ENCYCLOPÆDIA
BRITANNICA



MACROPÆDIA

The Encyclopædia Britannica
is published with the editorial advice
of the faculties of the University of Chicago;
a committee of persons holding
academic appointments at the universities
of Oxford, Cambridge, London, and Edinburgh;
a committee at the University of Toronto;
and committees drawn from members of the faculties
of the University of Tokyo
and the Australian National University.



THE UNIVERSITY OF CHICAGO

"Let knowledge grow from more to more
and thus be human life enriched."

The New Encyclopædia Britannica

in 30 Volumes

MACROPÆDIA
Volume 17

Knowledge in Depth

FOUNDED 1768
15TH EDITION



Encyclopædia Britannica, Inc.
William Benton, Publisher, 1943–1973
Helen Hemingway Benton, Publisher, 1973–1974
Chicago/London/Toronto/Geneva/Sydney/Tokyo/Manila/Seoul

©1979
by Encyclopædia Britannica, Inc.
Copyright under International Copyright Union
All rights reserved under Pan American and
Universal Copyright Conventions
by Encyclopædia Britannica, Inc.

Printed in U.S.A.

Library of Congress Catalog Card Number: 77-94292
International Standard Book Number: 0-85229-339-9

No part of this work may be reproduced or utilized
in any form or by any means, electronic or mechanical,
including photocopying, recording,
or by any information storage and retrieval system,
without permission in writing from the publisher.

Sonar

Sonar is the name of a technique for detecting the presence of objects underwater by acoustical echo. Having been developed during World War I, it antedates the better known radar, which uses electromagnetic echo. "Active sonar" employs an apparatus for radiating acoustical energy to bounce off underwater objects; "passive sonar" consists merely of the passive reception of acoustical energy generated by another source.

DEVELOPMENT OF THE TECHNIQUE

Interest in a means of detecting underwater objects was originally aroused by the problem of icebergs, dramatized by the sinking of the "Titanic" in 1912. The first proposal was put forward by a British meteorologist, L.F. Richardson, and the first successful application used in iceberg detection was made by the American radio pioneer, R.A. Fessenden. Development was stimulated by the outbreak of World War I and the impact of submarine warfare. A French physicist, Paul Langevin, played the leading role in research in which first British and then American scientists joined. A passive system of submarine detection, operational by 1916, employed a hydrophone (underwater microphone) and amplifier to pick up the noise emitted by submarine engines. By 1918 scientists had developed an active system in which a pulse of sound was transmitted and its rebounding echo used to detect a submarine even when its engines were shut down. The original term, "asdics," is said to have been derived from 'Anti-Submarine Division-ics,' although other explanations have been given. The name was long retained in the United Kingdom; the term sonar, from sound navigation and ranging, a United States acronym from World War II, is now used widely.

In the years between the wars, British and United States researchers refined techniques to such a point that the Allies enjoyed a substantial advantage over Germany in World War II underwater detection. They developed two types of beams, one vertical to bounce off the sea floor for depth determination, the other horizontal and capable of detecting underwater objects and obstacles near the surface. The depth detector, or echo sounder, was in widespread navigational use in the 1930s. A major American wartime contribution was the development of a system that could rapidly scan with a narrow beam either a sector or all around, without mechanical motion of the acoustic transmitters or receivers, making possible swift and methodical search for submarines.

In the years since 1945 a great deal of work has been done in naval acoustics by the United States, the United Kingdom, and the Soviet Union, and other maritime nations, but virtually all of it is secret and so has had little influence on peaceful applications. The most important of these is the use of echo sounders to detect shoals of fish, a potential discovered in the 1930s. By 1950 specialized equipment was being installed on fishing vessels, with notable success.

The echo sounder

In the early 1970s the simple echo sounder remained the basic sonar device used by fishing fleets. A type of sonar has been developed with a horizontal scan that locates fish at a distance of one kilometre (about 1,100 yards), greatly facilitating purse-seining (large net) operations by trawlers. The beam's direction can be changed by mechanical rotation of the underwater transmitting and receiving equipment, permitting a thorough search of a fair-sized area. The technique is slow, owing to the slow rate at which sound travels through water (1.5 kilometres per second).

Another problem in fish sonar is that there is as yet no technique for detecting fish close to the seabed in front of the ship, as the bottom-trawling method of fishing used by the British and other fishermen requires. There are, however, systems that accurately indicate the depth of a net towed behind the trawler, so that adjustments can be made to take advantage of any fish shoals detected on the main sonar.

Known military applications in the 1970s include the detection and location of submarines, control of antisubmarine weapons, sonar-equipped homing torpedoes, and mine hunting.

Detection ranges for civil and military sonar systems vary from 100 metres to 10 kilometres. Wave lengths for the acoustic signal range between 0.5 centimetres and 30 centimetres, corresponding to frequencies of approximately 300 kilohertz and 5 kilohertz.

BASIC PHYSICAL PRINCIPLES

Principles of a simple sonar system. As with most radar systems, sonar systems generally use the transmission and reflection of a pulse of energy as their basis of operation. The arrangement in Figure 1 is typical. Individual systems vary, with some, for example, using a single transducer both for generating the acoustic waves in the water and for detecting the reflected waves. The pulses they send out are always bursts of the transmitted, or carrier, frequency, although the term sonar may also include marine seismic work, using the sound pulses generated by a small explosion.

From B.K. Gazey and D.G. Tucker, *Applied Underwater Acoustics* (1966); Pergamon Press; reprinted by permission

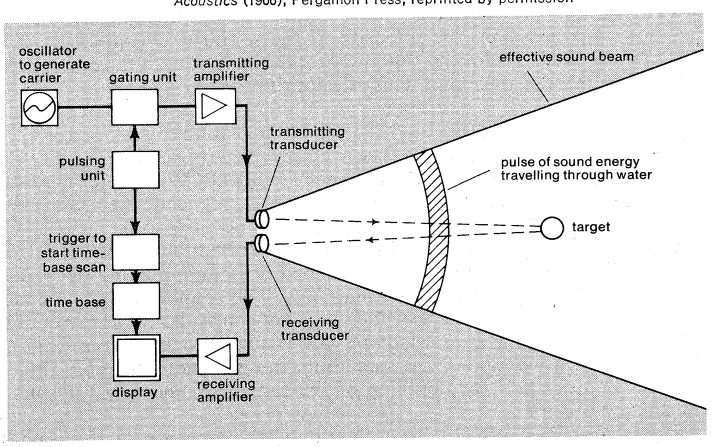


Figure 1: Schematic arrangement of a typical pulsed sonar system.

Display. The distance from the transducers to a reflecting target is indicated by the time elapsing between transmission of the pulse and the reception of an echo. The display, usually a pattern on a moving strip of paper, always includes a time base, the traverse of which is initiated by the transmitted pulse. For simple systems, the commonest display is a chemical recorder which records the received information on sensitized paper by means of a recording stylus drawn across paper. Since the echo actuates the marking mechanism, the position of the mark across the traverse indicates the range of the target. The

paper is moved slowly in a direction perpendicular to the traverse of the marking stylus to make successive traverses side by side. If the range of the target in relation to the transducers does not vary, the line produced by the echo is parallel to the direction of motion of the paper. If the range changes (due to the motion of the target or ship), the line slopes with respect to the paper motion.

Resolution. The resolution of the system in range, that is, the fineness of detail it can show in a direction outward from the transmitter, improves as the transmitted pulse is made shorter. The carrier frequency may have to be increased to permit this shortening. Greater bandwidth is then required in the receiving equipment. Improved angular resolution, that is, the ability to distinguish two targets at slightly different angles, is generally obtained by making the beams narrow (more directional).

Doppler sonar. Another sonar method, "Doppler detection," relies upon the relative speed of the target and the observing station to provide an indication of target speed. It employs the Doppler effect, in which an apparent change in frequency occurs when the observed and observer are in motion relative to one another. The classic example is the apparent change in the pitch of a train whistle as the locomotive approaches and passes an observer. Detection by this means requires a long pulse to give an aural recognition of tone. The received signal is mixed with a local signal of slightly different frequency, the frequencies being so arranged that mixing the two produces an audible signal. A shift of pitch between the transmitted and received signals can easily be detected by a trained ear; this method gives not only very sensitive detection but also valuable information regarding the speed of the target.

Doppler detection can be used on fast-moving ocean fish, but most fish move too slowly for the method to be of value for ordinary fishing.

The propagation of acoustic waves in water. If water did not introduce any losses, if it extended indefinitely in all directions, and if it were uniform in all respects, then the spreading of sound would be spherical. Under these conditions the intensity of the sound would decrease inversely to the square of the increasing distance from the transmitter. This relationship is usually expressed as a loss of six decibels per doubling of range for one-way transmission, or 12 decibels per doubling of range for the echo signal. The decibel is a logarithmic unit. If the intensity at one point is I_1 and at a farther point is I_2, then the loss in decibels between the points is ten times the logarithm of the ratio of I_1 to I_2.

Sound intensity in the sea

But the sea is not infinite in all directions. It is bounded by the surface and the seabed, so that spreading does not follow the inverse-square law. The sound intensity, channelled between surface and bottom, falls off less rapidly than the inverse-square law implies. In the extreme case, called cylindrical spreading, the loss increases only three decibels for doubling of the range for one-way transmission.

There is also some loss due to absorption of the sound energy by the water, which converts it into heat. This loss is variable; it is small at low frequencies but rises very rapidly with an increase in frequency. In freshwater, the loss expressed in decibels per kilometre is proportional to the square of the frequency and is about three at 100 kilohertz. In seawater, there is an additional loss at frequencies below one megahertz due to the dissolved salts. Below 100 kilohertz, this additional loss is approximately constant at around 15 decibels per kilometre.

Further propagation effects of importance result from variations of the velocity of sound in seawater. Under normal conditions the velocity is about 1,500 metres per second. The main causes of varying velocity are, in order of their importance: temperature, pressure due to depth, and salinity. The magnitudes of the effects cannot be expressed by any fundamental equation, though several empirical, or rule-of-thumb, equations have been proposed. Increased salinity, temperature, and depth all increase velocity over all normal ranges.

Perhaps the most serious effect of varying velocity is the refraction (bending) of the sound beam, which may cause a beam that is normally horizontal to be deflected to the sea bottom, where it will be reflected upward, then down again, and so on. Along any straight line through the transducer, therefore, there are intervals of range where detection is impossible and others where it is possible. This and similar effects are most serious in low-frequency systems since they have the greatest nominal range.

Effects of varying velocity

Another effect of varying velocity is the general scattering of the sound beam as it passes through turbulent water, leading to rapid fluctuations of signal strength.

INHERENT LIMITATIONS OF SONAR

Noise in water. Acoustic noise in water sets the ultimate limit to the range of detection because it eventually obscures the wanted signal. Noise is defined as power received that is not part of the desired signal and is not produced by the transmitter. Sources of noise can be inherent, can be caused by natural phenomena, or can be produced by man or animals.

Inherent noise. Inherent noise results from motion of molecules, caused by heat. The colder the water the less inherent noise is received. Basically, noise intensity from this source is independent of frequency although, like all noise, it is dependent on the bandwidth of the system.

Sea-state noise. The most important noise produced by natural phenomena is that resulting from wave action. This is called "sea-state noise." Its magnitude is dependent on the height of the waves. Essentially a low-frequency noise, its power diminishes rapidly as the frequency rises, becoming negligible in comparison with thermal noise around 50 to 150 kilohertz.

Animal and man-made noises. Noise-producing aquatic animals include several species of fish and shrimp, encountered frequently in warm waters. Man-made noise, particularly that generated by the ship carrying the sonar equipment, can be more serious. Although difficult to calculate in advance, it poses the main limitation to sonar performance.

Reverberation. When the sonar system must detect small targets, a random background due to "reverberation" can make detection difficult. Reverberation is the sum of all the numerous small echoes produced by reflections from sand and stone particles on the sea bottom, minute air bubbles, and other irregularities in the water. While background noise in the sea limits the maximum range of detection, reverberation may limit performance in all ranges. Since reverberation originates from the signal transmission, it has a power level at any time interval closely related to that of the signal. Consequently, to detect small objects in areas where reverberation level is high (due, for example, to a shallow and rough sea bottom), the beams must be as narrow as possible in the relevant dimension. When reverberation is the limiting factor in detection, increasing the transmitted power does not improve detection over most of the range. Consequently, many sonar systems operate at acoustic powers of fairly low peak level.

Directivity. Sonar is most efficient if the sound energy is confined to a narrow beam on transmission and the receiving transducer responds only to sound coming from a limited angle, or cone. If the transmitter is highly directional, then the sound intensity at any distant point, for a given power source, will be much higher than if the sound had been broadcast uniformly in all directions. Fewer unwanted objects will be illuminated by the transmitted beam. On reception, a high directivity means that the noise will be reduced. If the noise is isotropic (coming almost equally from all directions) the reduction is given by the Directivity Index. Directivity Index is defined for transmission as the ratio in decibels of the intensity at a particular point in the direction of maximum transmission to that which would result at the same point if all the transmitted power were spread uniformly in all directions. On reception, it is the ratio in decibels of the output power developed by a signal in the direction of maximum response to that developed by the same signal power if it were uniformly distributed over all directions. For ordinary transducers, Directivity Index is the same whether the transducer is used for transmission or reception.

The Directivity Index

NEW APPLICATIONS AND TECHNIQUES

Improved sonar techniques are leading to (and often originating from) new applications in various fields of underwater operations. Since naval requirements obviously call for longer ranges of detection, higher powers and larger transducers are being provided, augmented by the sophisticated computer technique of "signal processing," whereby low-level signals can be extracted from a noise background.

Side-scan sonar. In nonmilitary applications, one technique that is increasing rapidly in importance is side-scan sonar. This sonar has a narrow (about 1°) beam in the horizontal plane, but relatively wide (10° or 20°) in the vertical plane, looking sideways from the ship or towed body on which it is mounted. As the ship proceeds along its route, a map of the acoustic scattering from the seabed is built up on recording paper. With experience, these records can be interpreted in geological terms, giving in effect a map of the hardness and roughness of the seabed rocks; it reveals geological faults, sand waves, and ridges. The general arrangement is shown in Figure 2.

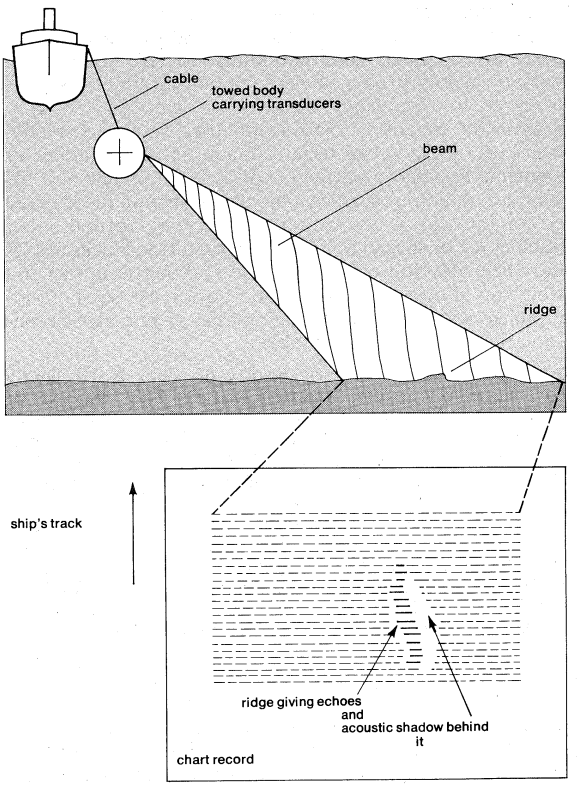


Figure 2: Principle of side-scan sonar with simplified record showing single small ridge with acoustic shadow behind.

Existing equipment for geological applications has a range of about 1,000 metres, but a recent model constructed by the British National Institute of Oceanography has a range of approximately 15 kilometres working at a relatively low frequency, of 6.6 kilohertz. The speed of search is high, but the results are seriously affected by refraction. Smaller, high-frequency, side-scan sonars, operating at 300 kilohertz (or even higher) with a range of about 200 metres, survey an area in finer detail. They can delineate such objects as underwater pipelines, oyster beds, and wrecks.

WPESS sonar. The growing trend towards higher resolution in sonars has led to the development of what is called within-pulse electronic-sector-scanning sonar (WPESS sonar). When very narrow sonar beams are used and it is desired to search a sector rapidly by swinging the beams around, mechanical beam swinging slows the rate of search, making the method unsatisfactory for fish finding and mine hunting when the ship is moving forward. In the WPESS method, a wide sector is covered by each transmitted pulse, and a narrow receiving beam is rapidly swung electronically across this sector. The beam is swung so rapidly that the whole sector width is examined for echoes before the acoustic pulse has travelled its own length through the water. In this way, the whole sector is examined on every transmitted pulse. Thus the search rate is very high, even if very narrow receiving beams are used to provide high resolution. The display is usually a cathode-ray tube giving some kind of plan view of the sector, as illustrated in Figure 3.

From *Underwater Journal* (September 1970)

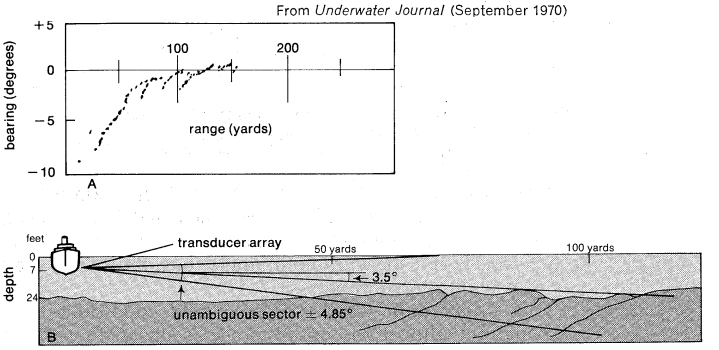


Figure 3: *Scanning sonar used with vertical 'scan' in a shallow reservoir.*
(A) Bearing-versus-range display.
(B) Geometry of bottom of reservoir deduced from A (vertical section).

Applications of WPESS sonar

The use of WPESS sonar is beneficial in the study of fish behaviour, in observing the behaviour of underwater gear, and in monitoring and directing the movements of divers. For some of these applications the acoustic frequency is made high (*e.g.*, 500 kilohertz) so that high resolution (*e.g.*, 0.5° in angle, 7.5 centimetres [three inches] in range) can be obtained using only small transducers. The maximum range is then relatively small, perhaps only 60 metres on 500 kilohertz equipment, or 200 metres in 300 kilohertz equipment. On the other hand, the system can be used equally well at lower frequencies. In early trials, 37 kilohertz was used with a range of detection on small fish shoals of nearly 800 metres.

Because of the trend to higher resolution, sonar can compete with optical methods for underwater viewing. Sonar makes it possible to view objects in muddy and turbid water where optical methods are ineffective. It is thus of great importance in police searches, in civil engineering (*e.g.*, river and harbour works), and in studying fish behaviour. In these applications it is essential to use a sonar that provides high angular (or lateral) resolution as well as high resolution in range. Range resolution can usually be increased by using a short pulse. Lateral resolution, however, raises difficulties.

Beam width considerations. Obtaining a beam of narrow angular width requires a transducer with dimensions equivalent to many wavelengths of the sonar carrier frequency. For example, a beamwidth of half a degree requires transducer dimensions of approximately 120 wavelengths. To keep the transducer small, a high frequency is necessary. Attenuation in water, however, increases rapidly with frequency. For example, at 500 kilohertz, it is in the region of 0.1 decibel per metre, while at one megahertz it is in the region of 0.3 decibels per metre. Thus the frequency cannot be increased indefinitely, and there is a limit to the reduction in size of the transducer. If, for example, the maximum range is to be about 50 metres, then 500 kilohertz would be a reasonable choice of frequency, the wavelength would be 0.3 centimetre, and the transducer length (l) for a 0.5° beam would be 36 centimetres (14 inches).

Far-field beamwidth

The concept of the 0.5° beam is based on the idea that the measuring point is far removed from the transducer from which the beam boundaries are drawn (far-field condition). But at close ranges the concept of angular beamwidth is invalid; if the transducer is straight (*i.e.*, unfocussed) the beam can hardly be narrower than the transducer length l, and the geometry becomes as shown in Figure 4. The far-field beamwidth θ is realized only after a range of about l^2/λ; *i.e.* 40 metres from the trans-

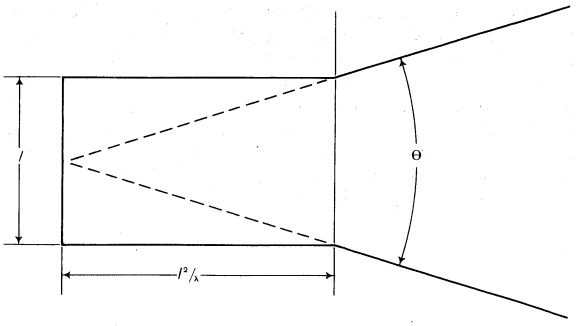


Figure 4: Geometry of beam near transducer (see text).
From B.K. Gazey and D.G. Tucker, *Applied Underwater Acoustics* (1966); Pergamon Press; reprinted with permission

ducer (λ is the wavelength in metres). This is almost the maximum range specified, so that "near-field" conditions exist almost throughout. The lateral resolution is therefore clearly not measurable in angle, but in distance, and is thus about 36 centimetres.

This situation in which the near-field conditions exist throughout almost the entire range is unsatisfactory. It can be overcome by using a system of signal processing, in which the receiving transducer is divided in half, and the signals from the two half-transducers are multiplied together. Thus a target can give a signal output from the receiver only if it lies in the beams of both half-transducers. Moreover, the far-field beamwidth is one-half of that corresponding to the length of the transducer used normally, and is realized at a relatively short distance from the transducer. Thus the near-field beam is extremely narrow, and lateral resolution of one-sixth of the transducer length (or less) is feasible. In the example taken above, the overall length of the transducer would be 18 centimetres, and the lateral resolution is quoted as three centimetres, or 0.5°, whichever is the poorer. This provides an enormous advantage for the multiplicative system.

Influence of modern electronics. Older forms of electronics with thermionic tubes or discrete transistors have been expensive to produce, somewhat unreliable, and costly to maintain. Because of the revolution in electronics that brought forth microelectronics, or integrated-circuit technology, however, it is now possible for a system to be sophisticated and complex and, at the same time, relatively inexpensive and reliable. This change in electronics has influenced sonar design, and already more refined and useful sonars are going into commercial production.

The microelectronic units that first became available were of the digital type; that is, they represented circuits that had switching, or on-off, functions of the type used in digital computers.

Several types of analogue circuits later became available in microelectronic form, and it is possible to make even a complex analogue system such as the WPESS sonar on this basis.

PRINCIPLES OF TRANSDUCERS

The purpose of transducers (strictly, electroacoustic transducers) is to convert an electrical signal into an acoustic signal, or vice versa. Three types of transducer are employed in underwater applications: magnetostrictive, piezoelectric, and electrostrictive.

Magnetostrictive. In a magnetostrictive type, a magnetic field is applied to a piece of suitable magnetic material, causing the dimension of the piece to decrease along the axis parallel to the field. When the field is alternating, application of a steady magnetic field (polarization) is necessary to give the acoustic wave the same frequency as that of the electrical signal. The magnetostrictive effect also operates in reverse: received acoustic signals cause compression of the material, altering the magnetic field, which in turn produces an electromotive force in the electrical winding.

Piezoelectric. Piezoelectric transducers use crystals that change in dimension according to the applied electric field. If the field is alternating, the crystals vibrate and radiate an acoustic wave. Conversely, if the crystals are acted on by acoustic waves, they generate an electric field. Piezoelectric materials used include quartz, ammonium dihydrogen phosphate, tourmaline, and lithium sulfate (see also PIEZOELECTRIC DEVICES).

Electrostrictive. Electrostrictive transducers are becoming the most widely used of the three. Materials used include barium titanate and lead zirconate. The change of dimensions depends on the magnitude but not the polarity of the applied electric field, so polarization is needed as with magnetostrictive transducers. Electrostrictive transducers generally have impedances of a few hundred ohms.

Transducers usually are operated at resonance; that is, the frequency at which they vibrate naturally, to get reasonable efficiency, making their frequency bandwidth rather narrow. When wide-band operation is required, the transducer is operated in such a way that its natural resonances are outside the operating frequency band. Various methods are used to reduce the sharpness of resonance as well as the cost of transducers. In one method, slices of crystal or ceramic material are alternated with plates of metal; such "sandwich" transducers are common.

It is difficult to generalize about transducer design since there are so many different kinds. In size transducers vary from tiny probes of perhaps one millimetre square to large low-frequency elements of perhaps ten kilograms in weight. Transducers (or, more correctly, transducer arrays) are usually made up of a number of elements to achieve particular directional effects and for a larger radiating surface that can transmit the desired power into the water without cavitation (formation of bubbles). For special applications, rigid transducer arrays have been developed measuring several metres in size and weighing almost a ton. For some stationary, sea-bottom-mounted sonars the array may consist of numerous separate transducers linked by cable.

BIBLIOGRAPHY. The early history of sonar is excellently covered in F.V. HUNT, *Electroacoustics: The Analysis of Transduction, and its Historical Background* (1954); sonar occurring in nature is treated in W.N. KELLOGG, *Porpoises and Sonar* (1961); and D.R. GRIFFIN, *Listening in the Dark: The Acoustic Orientation of Bats and Men* (1958). An elementary textbook is D.G. TUCKER, *Underwater Observation Using Sonar* (1966); more advanced textbooks include: D.G. TUCKER and B.K. GAZEY, *Applied Underwater Acoustics* (1966); J.W. HORTON, *Fundamentals of Sonar* (1957); and R.J. URICK, *Principles of Underwater Sound for Engineers* (1967). Sophisticated sonar techniques are described in a simple way in D.G. TUCKER, *Sonar in Fisheries: A Forward Look* (1967).

(D.G.T.)

Sonata

Deriving from the past participle of the Italian verb *sonare*, "to sound," the term sonata originally denoted a composition played on instruments, as opposed to one that was *cantata*, or "sung," by voices. Its first such use was in 1561, when it was applied to a suite of dances for lute. The term has since acquired other meanings that can easily cause confusion. It can mean a composition in two or more movements, or separate sections, played by a small group of instruments, having no more than three independent parts. Most frequently it refers to such a piece for one or two instruments. By extension, sonata can also refer to a composition for a larger instrumental group having more than two or three parts, such as a string quartet or an orchestra, provided that the composition is based on principles of musical form that from the mid-18th century were used in sonatas for small instrumental groups. The term has been more loosely applied to 20th-century works, whether or not they rely on 18th-century principles.

Sonata and sonata form

Quite distinct from all of the preceding, however, is the use of the term in "sonata form." This denotes a particular form or method of musical organization normally used within instrumental sonatas, string quartets, and other chamber music, and symphonies written since the beginning of the Classical period (the period of Mozart, Haydn, and Beethoven) in the mid-18th century.

The first concern of this article will be to establish the

principles of musical form often associated with sonatas for small and large groups of instruments. They will be approached through an examination of the principle of musical structure called "sonata form." A historical account of the origins and development both of the instrumental sonata and of sonata form will then endeavour to throw light on other meanings of the term. In conclusion, some estimate will be offered of the present and possible future roles of the sonata in musical life.

STRUCTURE OF THE CLASSICAL SONATA

Sonata form denotes a particularly fertile manner of organizing the musical structure of a single movement. It commonly occurs within the larger context of a multimovement scheme. Maturing in the second half of the 18th century, it provided the instrumental vehicle for much of the most profound musical thought until about the middle of the 19th century, and has continued to figure largely in the methods of composers down to the present day.

The basic elements of sonata form are three: exposition, development, and recapitulation, in which the musical subject matter is stated, explored or expanded, and restated. There may also be an introduction, usually in slow tempo, and a coda, or tailpiece, but these optional sections do not affect the basic structure. Although sonata form is sometimes called first-movement form, the first movements of multimovement works are not always in sonata form, nor does the form occur only in first movements. Likewise, another name for it, sonata-allegro form, is misleading, for it need not be in a quick tempo such as allegro.

At first glance sonata form may appear to be a species of three-part, or ternary, form. The three parts of ternary form are a first section (A), followed by a contrasting section (B), followed by a repetition of the first section (that is, ABA). The parts are interrelated not in terms of basic structure but by purely lyrical or character contrast. Actually, the three parts of sonata form developed out of the binary, or two-part, form prominent in the music of the 17th and early 18th centuries. In binary form the structure depends on the interrelationship not only of themes but also of tonalities, or keys, the particular sets of notes and chords used in each part. Thus, the initial part, which is repeated, leads directly into the second part by ending in the new key in which the second part begins. The second, also repeated, moves from the new key back to the original key, in which it ends. The second part thus completes the first.

Binary form

In sonata form the exposition corresponds to the first part of binary form, the development and recapitulation to the second. The exposition moves from the original key to a new key; the development passes through several keys and the recapitulation returns to the original key. This echoes the motion, in binary form, away from and back to the original key. In relation to binary form, sonata form is complex. It offers, in the exposition, contrasting musical statements. In the development these are treated dialectically; that is, they are combined, broken up, recombined, and otherwise brought into change and conflict. In the recapitulation they are restated in a new light. This organic relationship between parts marks the sonata form as a higher, more complex, type than the ternary form. The occasional designation of sonata form as compound binary form is useful in that it stresses its origins in the earlier form, but notes its added complexity.

The exposition. The emphasis on contrast, even conflict, is the element that distinguishes the exposition of a sonata-form movement from the first section of an earlier binary form. The first section of a binary movement in a Baroque suite or instrumental sonata, for example, might contain two clearly differentiated themes, but the stress is on continuity and on uniformity of musical texture rather than on contrast. In sonata form the emphasis is more dynamic; there is a stronger sense of contrast within the movement. The terms usually given the contrasting areas are "first subject/second subject" or "principal group/subsidiary group." These are misleading terms, for they imply a simple contrast of themes.

In reality it is contrast of key, or tonal contrast, that characterizes the sonata exposition. Usually the opening of the exposition is firmly rooted in the tonic, or "home," key of the work. The later segments of the exposition move decisively to a closely related but distinct key. The second key chosen is almost invariably one of the two keys most closely related to the home key. If the home key was a major key, the dominant key is chosen; if the home key was minor, the relative major is chosen. (The dominant key is the one whose keynote is five tones above that of the tonic, as C–G; the relative major has a keynote three tones above the relative minor, as A minor–C major.) The exposition thus creates an opposition of tonalities or key areas that the rest of the movement—the development and recapitulation—will strive to reconcile. Compared with the contrast of keys, the question of how many themes the movement possesses is of minor structural significance. Very often, a movement in sonata form has two clearly defined main themes, for example the first movement of Mozart's *Symphony No. 41 in C Major*. It may also have only one, like the first movement of Haydn's *Symphony No. 85 in B Flat Major*. Or it may have more than a half dozen strongly characterized themes, as does the first movement of Brahms's *Fourth Symphony*.

Thematic and tonal contrast

The thematic organization of a movement in sonata form may affect the character of the exposition, and thus of the whole movement, in two specific respects. When two themes or groups of themes are clearly differentiated, their distribution may help the listener to assimilate the cardinal points of the tonal design (that is, the arrangement of keys) of the movement. When, on the other hand, such differentiation between themes is obscured or set at variance with the organization of tonalities, the very tension between thematic design and tonal scheme may greatly enhance the subtlety and interest of the form. Such tension may produce not merely an interplay of melody and key within the movement but an interplay between two interplays. One fairly simple way of achieving this is shown in the first movement of Haydn's *Symphony No. 99 in E Flat Major*. Here, as in *No. 85*, the first theme is restated in the dominant key. This restatement could appear at first to be the second subject. But later it is followed by another distinct motif that, in terms of themes, is the real second subject. At the same time the neat, almost epigrammatic character of the second subject makes it similar to a codetta theme, which is often used to round off the exposition after both main subjects have been stated.

In the first movement of Mozart's *Symphony No. 35 in D Major (Haffner)* this interplay of interplays reaches a higher level of subtlety. The second theme, against which the first persists as a counterpoint, is stated "on" rather than "in" the dominant; that is, its harmonies suggest the dominant key, but remain part of the home, or tonic, key. The second theme is thus heard as a new perspective on the tonic. Later, when the dominant key is firmly established in its own right, Mozart introduces a new subject whose tune is closely related to the first theme. In this richly ambiguous structure, the newly introduced motif would be regarded by the criterion of key, as the second subject; in purely thematic terms, it might almost be said to constitute the beginning of the codetta, or concluding section.

The development and the recapitulation. The functions of the other two main sections follow naturally from what has been established in the exposition. Their purpose is to discuss and resolve the conflicts of tonality and theme that the exposition has raised. The development is an area of tonal flux—it usually modulates, or changes key, frequently, and any keys it settles in are likely to be only distantly related to the keys found in the exposition. It frequently proceeds by breaking the principal themes down into smaller elements and bringing these elements into new tonal or contrapuntal relations with each other. That is, themes or fragments of themes may appear in new keys; they may be combined to form apparently new melodies; they may be played against each other as counterpoint, or countermelody. One of the finest illustrations

Modulation and thematic alteration

of the methods of development used in the Classical period occurs in the first movement of Mozart's *Symphony No. 38 in D Major (Prague)*. Another resource of development is to seize on an apparently minor feature of the exposition and, by developing it extensively, to demonstrate its hidden importance. Another is to introduce entirely new material. This may provide a moment of relief in the course of a rigorous argument (as in the last movement of Mozart's *Piano Sonata in C Major*, K. 330); or it may allow the composer to expand the scope of a large-scale movement (as in the first movement of Beethoven's *Symphony No. 3 in E Flat Major*, the *Eroica*).

Sometimes such a theme may only *seem* to be new. In the first movement of Beethoven's *Symphony No. 4 in B Flat Major*, for instance, the theme in the development that is usually described as "new" is really a decorated version of a motif already heard in the exposition.

One common tactic in the Classical development is to begin the section with the codetta theme that ended the exposition. The first movement of Beethoven's *Cello Sonata No. 2 in G Minor*, Opus 5, No. 2, is an example.

The impact of this device, and of the development section as a whole, is often obscured by the common tendency among modern performers to ignore the composer's instruction, present in almost all sonata form movements of the Classical period, to repeat the entire exposition. When this repetition is omitted, the thematic balance of the movement is upset and the dramatic effect of the development's sudden departure from an established regularity can be ruined. Music is an art to which the controlled use of time is basic. The temporal structure of a movement cannot be altered without seriously changing the proportions of the whole.

Tension and recapitulation

Like the beginning of the development section, the point at which development passes into recapitulation is one of the most important psychological moments in the entire sonata structure. It marks the end of the main argument and the beginning of the final synthesis for which that argument has prepared the listener's mind. The Classical masters differ interestingly in their handling of this juncture. All of them usually prepare for it with a long passage of gathering tension. In Mozart the return of the tonic key and subject is managed with understated punctuality, the actual moment of recapitulation gliding in almost unnoticed. Haydn and Beethoven tend to celebrate its advent with panoply.

The recapitulation presents the principal subject matter of the movement in a new state of equilibrium. The main subjects of the exposition are heard almost always in the same order as before, but now both subjects are typically in the tonic key, whereas in the exposition the first was in the tonic, the second in the dominant key. As a result of the musical events in the development, the listener perceives the subjects in a new relationship—rather like a traveller who glimpses the constituent parts of a valley separately as he climbs a hill and then, when he reaches the summit, sees the entire landscape for the first time as a whole. The recapitulation can vary greatly in the literalness with which it repeats the elements of the exposition. Sometimes, as in the first movement of Mozart's *Sonata in B Flat Major*, K. 570, a tiny modification in the transition that originally led from the tonic to the dominant key is enough to effect the necessary change of key perspective and keep the second subject in the tonic key. In other cases (the first movement of Beethoven's *Eroica Symphony*, for instance, and many of Haydn's symphonic movements) far-reaching modifications and reshufflings of the original material are made in the recapitulation. As in any living manifestation of a principle of musical form, the methods differ vastly from work to work; but the effect is always to bring about the reconciliation of opposites that is essential to sonata form.

A large-scale sonata movement often creates conflicts of key and theme that cannot be completely settled even by the full process of recapitulation. In this case, the movement may be rounded off with a coda, or concluding section. Beethoven often extends the coda so greatly that it becomes almost a second development section, as in his *Appassionata* piano sonata. But this is no more an essential element of sonata form than the introduction that may precede the main movement.

Interrelationship of movements. The form described above is that exemplified in most first movements of sonatas, and of sonata-style compositions, in the Classical period. There are usually two, three, or four movements in the entire work. Two-movement and, more particularly, three-movement schemes are most common in sonatas for one or two instruments. Symphonies and string quartets almost always have four movements, and Beethoven, particularly in his earlier period, sometimes expanded the scheme of the instrumental sonata to four movements too.

The first movement in all of these patterns is usually fast; the second commonly provides the contrast of a slower tempo; and the last in most cases is again fast. When there are four movements, a simpler, dance-style movement of the type also found in the suite is included. This is usually placed between the slow second movement and the finale; in some cases it stands second and the slow movement third.

The forms of these other movements vary much more than that of the first, which in Classical examples is almost invariably the weightiest. Since their function is to complement the experience of the first movement through a new but related range of contrasts, their scope and manner depend on the point to which the issues of the work have already been taken. Simple ternary (ABA) form and variation form (*i.e.*, theme and variations) are among the most common patterns for the slow movement, but rondo and sonata forms are also used. In rondo form a recurring theme is contrasted with a number of intervening themes, as ABACA. When sonata form is used in slow tempos, the demands of overall proportion frequently cause the omission of the development section. Sonata form, rondo, and, less often, variation form are also used for the final movement. In final movements, also, the simple rondo pattern (ABACA) is often expanded into ABA-development-BA, with B in the dominant key at its first appearance and in the tonic key at its second. The result is a hybrid form known as sonata-rondo.

The dance movement

In the first part of the Classical period, the dance movement, when it appeared, usually consisted of a minuet in fairly simple binary form (the two-part form from which sonata form evolved). This was followed by a second minuet known as the trio, which tended in orchestral works to be more lightly scored. The first minuet was then repeated, normally without its own internal repeats. The minuet-trio-minuet structure forms an overall ternary pattern. Haydn frequently, and Beethoven still more often, chose to speed the traditional minuet up to the point at which it lost its dance character and became a scherzo, a quick, light movement usually related to the minuet in form. In some extreme cases, such as the ninth symphonies of both Beethoven and Schubert, the binary structures of both scherzo and trio were expanded into small but complete sonata-form structures. In this way, as with the sonata-rondo, the principles of thematic development and key contrast spread during the Classical period as the sonata form began to influence other movements.

Such are the outlines of the most fertile form in Western instrumental music since 1750. Before discussing its origins, growth, and later modification, there should be a warning against phrases like "true sonata form." If sonata form as described is considered the sole "true sonata form," the implication is that instrumental sonatas written in forms other than sonata form are not genuine.

Actually, the principle exemplified in sonata form represents one way of organizing the passage of sound through time. Its scope is enormous; it was the basis for some of the greatest works of Western music; and it still contains the seeds of potential further development. But it is only one episode in a complex chronicle of styles and principles of musical organization. In contrast to earlier forms, it emphasizes conflict instead of continuity and derives its impact from the explosive power of tonal organization instead of the smoother influence of melody.

THE SONATA IN MUSICAL HISTORY

The sonata in all its manifestations has roots that go back long before the first uses of the actual name. Its ultimate

Earliest origins

sources are in the choral polyphony of the late Renaissance (music having several equal melodic lines, or voices). This in turn drew at times on both liturgical and secular sources—on the ancient system of tones or modes of Gregorian chant, and on medieval European folk music. These two lines were constantly interweaving. Popular tunes, for example, were used as the starting point for masses and other religious compositions from the 15th to the early 17th centuries. Sacred and secular elements influenced the development of both the sonata and the partita (or suite) of the Baroque period.

The specific musical procedures that were eventually to be characteristic of the sonata began to emerge clearly in works by the Venetian composers of the late 16th century, notably Andrea Gabrieli (*c.* 1520–86) and Giovanni Gabrieli (1556–1612). These composers built instrumental pieces in short sections of contrasted tempo, a scheme that represents in embryo the division into movements of the later sonata. This approach is found not only in works entitled "sonata," such as Giovanni Gabrieli's *Sonata pian' e forte* (*Soft and Loud Sonata*) of 1597, which was one of the first works to specify instrumentation in detail; the instrumental fantasia and the canzone, an instrumental form derived from the chanson or secular French part-song, display a similar sectional structure. Like early sonatas, they were often contrapuntal (*i.e.*, built by counterpoint, or the interweaving of melodic lines in the different voices, or parts). At this stage sonatas, fantasias, and canzoni were often indistinguishable from each other, and from the fuguelike ricercare, though this form is generally more serious in character and more strictly contrapuntal in technique.

In the 17th century stringed instruments eclipsed the winds, which had played an at least equally important role in the sonatas and canzoni composed by the Gabrielis for the spacious galleries of St. Mark's Cathedral, Venice. Claudio Monteverdi (1567–1643) devoted more of his energies to vocal than to instrumental composition. The development of instrumental writing—and of instrumental musical forms—was carried on more and more by virtuoso violinists. One of these was Carlo Farina (fl. *c.* 1630), who spent part of his life in the service of the court of Dresden, and there published a set of sonatas in 1626. But the crowning figure in this early school of violinist-composers was Arcangelo Corelli (1653–1713), whose published sonatas, beginning in 1681, sum up Italian work in the field to this date.

Apart from their influence on the development of violin technique, reflected in the works of such later violinist-composers as Giuseppe Torelli (1658–1709), Antonio Vivaldi (1678–1741), Francesco Maria Veracini (1690–*c.* 1750), Giuseppe Tartini (1692–1770), and Pietro Locatelli (1695–1764), Corelli's sonatas are important for the way they clarify and help to define the two directions the sonata was to take. At this point the *sonata da chiesa,* or church sonata, and the *sonata da camera,* or chamber sonata, emerged as complementary but distinct lines of development.

Church and chamber sonatas

The *sonata da chiesa* usually consists of four movements, in the order slow–fast–slow–fast. The first fast movement tends to be loosely fugal in style (*i.e.*, using contrapuntal melodic imitation), and thus reflects, most clearly of the four, the sonata's roots in the fantasia and canzone. The last movement, by contrast, is simpler and lighter, often differing from the dance style typical of the *sonata da camera* only in that its sections are not repeated. The *sonata da camera* is altogether less serious and less contrapuntal than the *sonata da chiesa,* and it tends to consist of a larger number of shorter movements in dance style. If the *sonata da chiesa* was the source from which the Classical sonata was to develop, its courtly cousin was the direct ancestor of the suite, or partita, a succession of short dance pieces; and in the 18th century, the terms suite and partita were practically synonymous with *sonata da camera.* The two streams represented by church and chamber sonatas are the manifestation, in early Baroque terms, of the liturgical and secular sources found in Renaissance music. The Baroque style flourished in music from about 1600 to about 1750. Down to the middle of the 18th century the two influences maintained a high degree of independence; yet the injection of dance movements into the lighter examples of the *sonata da chiesa* and the penetration of counterpoint into the more serious suites and *sonate da camera* show that there was always some cross-fertilization.

Another characteristic of the Baroque sonata that Corelli's work helped to stabilize was its instrumentation. Around 1600 the musical revolution that began in Italy had shifted emphasis from the equal-voiced polyphony of the Renaissance and placed it instead on the concept of monody, or solo lines with subordinate accompaniments. The comparatively static influence of the old church modes was superseded by the more dramatic organizing principle of the major–minor key system with its use of contrast of keys. Although counterpoint continued to play a central role in musical structure for another hundred years and more, it became a counterpoint that took careful account of the implications of harmony and of chords within the framework of the major and minor keys.

The continuo part

In this context the continuo, or thorough bass, assumed primary importance. The composer that used a continuo part wrote out in full only the parts of the upper melody instruments. The accompaniment, which was the continuo part, was given in the form of a bass line, sometimes supplemented with numbers, or figures, to indicate main details of harmony, whence the term "figured bass." The continuo was "realized," or given its performed form, by a low melody instrument (viola da gamba, violone, or later cello or bassoon) in collaboration with an organ, harpsichord, or lute. The collaborating instrument improvised the harmonies indicated by the figures or implied by the other parts and so filled the gap between the treble and bass lines.

In Corelli's work, "solo" sonatas, for one violin with continuo, are found alongside others for two violins and continuo described as sonatas *a tre* (*i.e.*, for three), early examples of the trio sonata that was the principal chamber-music form until about 1750. The use of "trio" for sonatas played by four instruments is only superficially paradoxical: although trio sonatas were played by four instruments, they were considered to be in three parts—two violins and continuo. Moreover, specific instrumentation at this period was largely a matter of choice and circumstance. Flutes or oboes might play the violin parts, and if either harpsichord or cello or their substitutes were unavailable, the piece could be played with only one of them representing the continuo. But a complete continuo was preferred.

Corelli's importance is as much historical as musical. Perhaps because a vigorous line of Italian composers of violin music followed him, he is commonly accorded the main credit for late 17th-century developments in sonata style. But his undeniably vital contribution should not distract attention from equally important work that was done around the same time outside Italy.

Early development outside Italy

In France Jean-Baptiste Lully's lucrative monopoly of music at the royal court and the immense popularity of spectacular ballets used as courtly entertainments naturally led, through François Couperin (1668–1733), to a concentration on the smaller dance forms found in the ballet and courtly social dance. This concentration gave the French school its pre-eminence as producer and influencer of the 18th-century dance suite. The French, thus occupied with dance music, had little effect on the growth of the *sonata da chiesa.* But in Germany, where in 1619 Michael Praetorius (1571–1621) published some of the earliest sonatas, the sonata developed from an originally close relation to the suite into a more ambitious blend. As it evolved it combined the suitelike multisectional structure of the *sonata da camera* with the contrapuntal workmanship and emotional intensity of the Italian *sonata da chiesa* form.

One of the first contributors to this development of the Italian influence was the Austrian composer Johann Heinrich Schmelzer (*c.* 1623–80). In Nürnberg in 1659 he published a set of trio sonatas for strings, following it in 1662 with a set for mixed strings and wind instruments, and in 1664 with what may have been the first set of so-

natas for unaccompanied violin. The German composer Johann Rosenmüller (*c.* 1620–84) spent several years in Italy; his *Sonate da camera cioè sinfonie* (*i.e.*, suites or symphonies), published in Venice in 1667, are essentially dance compositions. But 12 years later, in Nürnberg, he issued a set of sonatas in two, three, four, and five parts that vividly illustrate the German trend toward more abstract musical structure and expressive counterpoint. During this period even pieces with dance titles began to lose their danceable character and became compositions meant only for listening.

Meanwhile, the greatest member of this school, Heinrich Ignaz Franz von Biber (1644–1704), published several sets of sonatas—some for violin and continuo, others in three, four, and five parts. In these, from 1676 onward, he took a penchant for expressiveness to extremes of sometimes bizarre but often gripping profundity that contrast sharply with the bland, polished style of Corelli. The titles of some of Biber's sets of sonatas specifically indicate his aim of reconciling church and chamber styles. The 1676 publication, for instance, is entitled *Sonatae tam aris quam aulis servientes* (*Sonatas for the Altar as Well as the Hall*). And being himself, like Corelli, a violinist of extraordinary powers, Biber made a valuable contribution to the development of instrumental technique in a set of sonatas for unaccompanied violin in which the practice of *scordatura* (adjustment of tuning to secure special effects) is ingeniously exploited.

The English composers were achieving a comparable intensification of expression during the 17th century, though in their case the technical starting point was different. In accordance with the characteristic time-lag of the English in the adoption of new European musical methods, the English continued to work with polyphony in the Renaissance manner, while the Italians were perfecting monody and the Germans fruitfully uniting monody with their own contrapuntal tradition. English polyphony in the 17th century attained a remarkable level of technical finish and emotional grandeur. Thomas Tomkins (1572–1656), Orlando Gibbons (1583–1625), John Jenkins (1592–1678), and William Lawes (1602–45) were the chief agents of this refining process. They and their predecessors, notably John Coperario (*c.* 1575–1626), made a gradual transition from the string fantasia bequeathed by William Byrd and other composers during the reign of Elizabeth I (1558–1603) and approached the new kind of musical form associated with the Baroque sonata; but they always stayed closer than their continental colleagues to the spirit of polyphony.

Contribution of English polyphony

When Henry Purcell (*c.* 1659–95), in his three-part and four-part sonatas, submitted this rich English tradition to the belated impact of French and Italian influence, he produced a fusion of styles that was the highest point of musical inspiration yet reached by the emergent sonata form.

The Baroque sonata. The years from the end of the 17th century to the middle of the 18th represent a moment of equilibrium in the interaction of counterpoint and monody that had created the Baroque sonata. The continuo device, as long as it endured, was a sign that the balance still held—and it did endure as long as the trio sonata kept its central position as a chamber-music medium. During the first half of the 18th century the later Italian violinists, most notably Vivaldi, were prolific creators of trio sonatas. Sometimes they leaned to a three-movement pattern (fast–slow–fast), influenced by the direction the Italian operatic sinfonia, or overture, was taking. More often the old four-movement pattern was preserved. In this well-tested shape, too, Georg Philipp Telemann (1681–1767) produced hundreds of examples that maintained a remarkably consistent standard of musical interest, George Frideric Handel (1685–1759), working for most of his life in England, composed some trio sonatas, and also some valuable sonatas for solo instrument with continuo. In France, Joseph Bodin de Boismortier (1691–1755) and the violinist Jean-Marie Leclair the elder (1697–1764) cultivated both solo and trio genres with charm although with less profundity.

Yet even while the sonata with continuo flourished, the forces of tonality, or organization in terms of keys, developed intensely toward a use of key contrast that would eventually drive the trio sonata from the scene. The continuo itself was being undermined by the growth of interest in instrumental colour, and the figured bass could not long survive the tendency toward scoring for specific instruments and exhaustive detailed musical notation.

Rise of tonal consciousness

Beginning in 1695 Johann Kuhnau (1660–1722) had published the first sonatas for keyboard instrument alone, some of them programmatic pieces on biblical subjects. J.S. Bach (1685–1750), the greatest composer of Baroque sonatas, continued the move away from the treatment of the keyboard in the subordinate, "filling-in" capacity that was its role in the continuo. He wrote a small number of trio sonatas after the traditional scheme, and also a few violin and flute sonatas with continuo; but at the same time he produced the first violin sonatas with obbligato harpsichord parts (that is, obligatory and fully written out, rather than improvised), others for flute or viola da gamba with obbligato harpsichord, and three sonatas (along with three partitas) for unaccompanied violin.

In these works, as in some of Telemann's later sonatas, the power of key or tonality to articulate sections of musical structure, and its ability to provide a harmonically derived eventfulness—a sense of expectation succeeded by fulfillment—began to make itself felt. These powers of key are the seed from which the Classical sonata form originated. But at this point the dualism engendered by tonal and thematic contrast had not yet supplanted the more continuous, unitary processes at work in a composition based on counterpoint. Nor was the consciousness of tonality any more advanced in the otherwise forward-looking work of Domenico Scarlatti (1685–1757). His harpsichord sonatas—555 movements survive, many designed to be played in pairs or in groups of three—are often original to the point of idiosyncrasy in expression. They introduced a valuable new flexibility in the treatment of binary form, and they had a powerful effect on the development of keyboard writing. But in formal terms they still belong in the old world of unity—even their strongest contrasts have an air of being suspended in time, quite unlike the far-ranging effects of conflict through time that are the basis of the Classical sonata.

A later generation of composers completed the transition from Baroque to Classical sonata. One of J.S. Bach's own sons, Carl Philipp Emmanuel Bach (1714–88), plunged enthusiastically into the new resource of dramatic contrast. In about 70 harpsichord sonatas, and in other works for chamber ensembles and for orchestra, he placed a new stress on key contrast not only between but, more important, within movements. Correspondingly, he emphasized the art of transition.

Transition from Baroque to Classical

In the development of sonata form in orchestral music, particular value attaches to the work of the Austrians Georg Matthias Monn (1717–50) and Georg Christoph Wagenseil (1715–77) and of the Italian Giovanni Battista Sammartini (1701–75). All three played vital roles in shaping the symphony, which assumed an importance equal to that of the solo or small-ensemble sonata. Their symphonies further stressed the individual characterization of themes and, in particular, the use of the second subject to shape form. Another of Bach's sons, Wilhelm Friedemann Bach (1710–84), made sporadic but interesting contributions to this development, and a third, Johann Christian Bach (1732–82), who settled in London, exploited a vein of melodic charm that influenced Mozart.

The Classical sonata and after. By about 1770 most of the specific changes that dictated the shift from Baroque sonata to Classical sonata were firmly established. Through the work of the Neapolitan school of opera led by Domenico Scarlatti's father Alessandro (1660–1725), the operatic sinfonia, or overture, had streamlined the traditional *sonata da chiesa*. It omitted the opening slow movement and abandoned the fugal manner that was the first allegro's link with the past. In the new three-movement pattern, a minuet sometimes replaced the fast, abstract finale. In other cases, the inclusion of both minuet and finale brought the number of movements back to four. The south German Mannheim school of composers

—most notably Johann Wenzel Stamitz (1717–57) and his son Karl (1745–1801)—developed the technique of the orchestra, whose resources now provided an ideal laboratory for experimentation with the dramatic effects of tonal contrast.

By this time the Classical sonata proper (*i.e.*, with at least one movement in sonata form), whether in the medium of sonata, trio, quartet, quintet, or symphony, could provide a vehicle for consolidating the process begun nearly two centuries earlier by the revolution from equal-voiced polyphony to monody, with its emphasis on melody and harmony. The Rococo style of the mid-18th century, generally known as style galant, had attained a halfway stage in which counterpoint had been virtually dropped and tunes had occupied the forefront of interest. But now, in the mature Classical style of Haydn and Mozart, superficial melodic interest was in turn subordinated. In this style the value of tunes lay in their role as functions of tonality. Key by this time had assumed a central role as the fundamental articulator of form. As a corollary, musical themes were often, though not always, reduced to the status of mere motifs, or tags. The theme's harmonic implications, which contribute to the feeling of key, took precedence over its attractiveness as melody.

Subordination of melody to tonality

The new musical principle—that of contrast of key—reached full expression in Haydn and Mozart through their use of sonata form as a principle of musical organization. Haydn's most valuable work in the sonata form is found in his series of over 80 string quartets, over 100 symphonies, 52 keyboard sonatas, and 31 trios for piano, violin, and cello. Unlike Haydn, Mozart was at his greatest in the fields of opera and of the solo concerto. (The latter, though it shared with sonata form such elements as the central principle of key contrast, was a medium that evolved, through the Baroque concerto grosso, from the fundamentally different source of the solo vocal aria and the vocal-instrumental concerto; see CONCERTO.) But in the last six symphonies, the last ten string quartets, about a dozen keyboard sonatas, and several trios, quartets, quintets, and serenades, Mozart achieved outstanding examples of sonata structures. The formerly prominent sonata for violin plays a relatively minor part in both men's output: the violin had been eclipsed by the rise of interest in keyboard instruments. It was reintroduced almost surreptitiously as a distinctly subordinate partner and regained a leading role only toward the end of the 18th century in Mozart's later violin sonatas and then in Beethoven's.

The strikingly individual details in Haydn's and Mozart's handling of sonata form are all features consonant with the general outlines of the form. Examples of different approaches include Haydn's taste for combining dualistic key schemes with monistic thematic material (that is, the use of the same basic theme in both keys). He also frequently set slow movements in keys only distantly related to the key of the first movement. Mozart preferred strongly differentiated themes, and he often reshaped his second subjects drastically when they reappeared in the recapitulation. Beethoven, in his sonata compositions (preeminently, the 32 piano sonatas, the 16 string quartets, the trios, the 9 symphonies, and the sonatas for violin and for cello), retained the basic sonata form. But he vastly extended its scale; *e.g.*, by increasing the importance of the coda, or concluding section, and by using unusual keys in the exposition, which was greatly increased in length. He also introduced extramusical implications of a profound philosophical nature. In his later sonatas and quartets he began to move away from the dualistic sonata principle and back to the monistic approach exemplified in variation form and fugue.

The case of Franz Schubert (1797–1828) is quite different. The first movement of the *Symphony No. 5 in B Flat Major* (written when he was 19) is one among several places that illustrates a changing attitude to sonata principles. In the recapitulation of this movement, the first theme is given in the subdominant key (the key whose keynote is five tones below that of the tonic, or home key, as F–C, just as the tonic's keynote is five tones below that of the dominant key, as C–G). This device enables Schubert to place the second theme in the tonic key (the goal of the recapitulation) without altering the transition between the two themes; for the same passage that, in the exposition, took the music from the tonic to the dominant serves, in the recapitulation, to take it from the subdominant to the tonic. This essentially labour-saving procedure is evidence of a certain lack of patience with the workings of sonata form as hitherto practiced. Up to this time, sonata form, first treated as a textbook study after Schubert's death, was not a set of rules codified by theorists and followed by composers. Rather, it was a principle of composition that grew out of earlier forms and that can be generalized from an examination of the actual work of Haydn, Mozart, Beethoven, and their contemporaries.

19th-century trends

Schubert's interests lay in new directions, and the first steps in two such directions are to be found in the greatest of his instrumental works. His later sonata-form compositions in all media—when they follow the rough traditional scheme of exposition, development, and recapitulation—modify it substantially. He frequently expanded the number of tonal centres (central keys) in the exposition, and sometimes also the number of basic themes, from two to three. This tendency to expansion affects the whole subsequent course of the Austro-German symphonic tradition. It is the direct ancestor of the expositions of Anton Bruckner (1824–96), with their three distinct thematic groups, and of the vastly extended sonata structures of Gustav Mahler (1860–1911). At the same time, Schubert's *Fantasy in C Major* (*Wanderer*) for piano (1822) exemplifies an opposite 19th-century trend toward contraction, through the fusion of the sonata's formerly separate movements in one closely integrated whole: the four movements of the fantasy are based on transformed versions of a single theme. Similarly, in France, Hector Berlioz (1803–69) in his *Symphonie fantastique* transformed the theme representing the artist's beloved (the *idée fixe*, or fixed idea) so that it took different forms in each movement. In this case the transformation was affected by the program or "plot" of the symphony. This was a departure from the abstract, or plotless, character of the Classical sonata. The tendency to fusion—that is, to thematic unity between movements—was the source of the thematic transformations used in symphonic poems, such as those of Franz Liszt (1811–86), as a basic principle of musical structure. But in these works the program rather than any abstract musical form suggests the particular course of the transformation of the themes. For this reason their specific form does not depend, as did that of the Classical sonata, on the exposition-development-recapitulation principle of contrast, conflict, and reconciliation of keys. A corresponding evolution away from the Classical form of the sonata for one instrument occurs in Liszt's one-movement *Piano Sonata in B Minor* (1853). In this work he used a single extended movement with subdivisions analogous to the sections of sonata form. But the specific use of his four themes, which are transformed and combined in free fashion, departs from the usual order of the classical Sonata.

The influence of program music

Robert Schumann (1810–56) likewise experimented fruitfully, especially in his *Symphony No. 4 in D Minor*, with the Schubertian idea of fusing movements together. The tendency to use thematic transformation in a manner that moved away from the Classical sonata form was complemented by César Franck, who adhered to the basic form but from 1841 utilized a "cyclic" approach; that is, one of fusion, or thematic relationships between movements. Brahms, on the other hand, carried the more familiar Classical sonata form to its highest point of complexity. In addition to making valuable innovations in rhythmic structure, he gave the role of counterpoint a new lease of vigour and interest and used the concept of thematic relationships between movements in a particularly subtle way. Chopin's three piano sonatas are concerned more with lyrical expression than with innovative formal methods. Similarly, the sonata compositions of Carl Maria von Weber (1786–1826) and Felix Mendelssohn (1809–47), which generally followed the patterns of their Classical predecessors and were highly regarded

in their day, contributed little to the evolution of sonata form away from its Classical state. This evolution, illustrated by the works of Berlioz and Liszt, was carried forward by Mahler. In his symphonies expansion, through inclusion of more than two tonal centres and groups of themes, and fusion, or the creation of unity between movements, are combined. This gives rise to expansive compositions held together by complex interrelationships between themes. Arnold Schoenberg (1874–1951), in such works as his *First String Quartet* (1904), carried the idea of fusion of movements to its logical conclusion: this is a one-movement work with contrasting sections in which all the themes used are derived from a few basic motifs.

Modern directions. Two important 19th-century developments tended to weaken the effectiveness of the Classical sonata form as an organizing principle. One, exemplified by Richard Wagner (1813–83), was an increasing use of chromaticism; that is, of notes and chords foreign to the key in which a passage of music is written. Chromaticism, when used extensively, broke down key feeling. Instead of being heard as a contrast to, or special modification of, the key, it became so prominent that the key itself was not heard strongly enough to establish itself in the listener's mind. Secondly, Liszt and his followers weakened the sonata form by using in their symphonic poems musical organizations based on program rather than on contrast of keys. But although the effectiveness of key as a basis for musical organization was weakened by the late 19th century, Mahler and Carl Nielsen (1865–1931) provided a modification of the sonata form that made use of tonality in a new way. This innovation, progressive tonality, used the home key as a goal to be worked toward from more or less distant key regions, so that a work ends in a different key from the one in which it began. Mahler and Nielsen arrived at the same notion independently at the same time. Nielsen's *First Symphony* and Mahler's *Second Symphony* (both 1894) are the first to use progressive tonality, and both composers forged highly individual new forms from it.

Weakening of the Classical sonata form

Most compositions written in sonata form after Wagner's era, however, lack a certain sense of vitality. Frequently, because the effectiveness of key or tonality has been weakened, such compositions centre on melody without the strong contrast of tonality that underlay the Classical sonata. Some composers made stylistic compromises. Schoenberg's *Fourth String Quartet* used his twelve-tone (dodecaphonic) approach to composition, an approach that began with a "row," or series, of the 12 tones of the chromatic scale, chosen by the composer to serve as the melodic and harmonic basis for the composition. In this work he fits the twelve-tone style into the outlines of sonata form. The result is based on contrasting themes, rather than on the Classical sonata principle of key contrast, because twelve-tone music, being atonal, deliberately avoids the creation of a sense of key. In a comparable way, though in the context of a different style, some of the sonatas of Sergey Prokofiev (1891–1953) use the outward formal divisions of the classical sonata form but stress the interest of melody as such, leaving tonality —still present in this case—to play a decorative rather than a structural role.

New principles

Other modern composers developed new principles of musical form. Although these principles appear in genres traditionally associated with the sonata, such as instrumental sonatas, string quartets, and orchestral works, they vary in the degree to which they are or are not related to the Classical sonata form.

One of the more useful of such principles has been the technique of constructing large-scale compositions from transformations and developments of a single germinal motif, often merely two or three notes. Like Schoenberg's approach, in which a twelve-tone row is transformed, this is actually the application at a more radical and consistent level of the 19th-century principle of thematic transformation. The symphonies of Jean Sibelius (1865–1957) are based on this method. So are those of Ralph Vaughan Williams (1872–1958), who also used some of the features of sonata form but imaginatively reshaped them and transformed their proportions to suit his purpose. In the nonsonata works of Schoenberg and his pupils Alban Berg (1885–1935) and Anton Webern (1883–1945), the twelve-tone method produced legitimate new forms of the highest historical importance; but when forced into an uncomfortable liaison with earlier schemes of organization such as the sonata, its effectiveness diminished. In the works of Béla Bartók (1881–1945), passages built on folk music scales, rather than on the major and minor scales of 18th- and 19th-century keys or tonalities, are used alongside atonal passages. His musical structure frequently takes the form of a combination of elements of sonata form with a simple "archlike" structure such as ABCBA. Paul Hindemith (1895–1963) contributed copiously to the sonata medium with works for almost every known instrument, but as far as the form was concerned his innovations were of minor significance. Michael Tippett (born 1905) in his *Second Symphony* and sonatas (*e.g.*, for piano; for four horns) uses tonality in a fresh and valid way, and he has effected a stimulating rapprochement of the sonata form with the equal-voice polyphony characteristic of the English fantasia and madrigal (a genre of part-song) of Elizabethan and Jacobean times. The *Second Symphony* of Wilfred Josephs (born 1927) shows yet another potentially valuable reinterpretation of the fused-movement approach to the sonata: its long first movement serves the function of exposition, three intermediate movements act on one level as development and on another level as a combination of slow movement and scherzo, and a brief finale serves as a kind of recapitulation.

Other musical approaches use metre and instrumental tone colour to mark important musical points much as traditional sonata form used contrast of keys. Elliott Carter (born 1908) has combined a use of germinal motifs with a new rhythmic technique known as "metric modulation," a controlled change of metre foreshadowed in Brahms by such passages as the end of the *Second Piano Concerto*. Carter's *Sonata for Cello and Piano* is an example of this use of metre. Carter also uses the idea of sharply differentiating the musical subject matter given to the individual instruments of an ensemble—a resource found earlier in the *Second String Quartet* of Charles Ives (1874–1954). Some of the many styles of Igor Stravinsky (1882–1971), particularly after his late adoption of the twelve-tone approach, make ingenious use of germinal motifs; but his music really bases its structure on the juxtaposition of large blocks of distinct musical character, rather than on "development" in the sense traditionally associated with the sonata.

Music in the 1970s is too various in form, medium, esthetic attitude, and social function to allow any confident predictions. But all of these examples suggest that the sonata, and its special manifestation, the sonata form, can still provide composers with fertile areas of activity. As in the time of Haydn, Mozart, and Beethoven, success will continue to reward those who develop musical forms that grow naturally from the specific principles of composition used in their works, much as the sonata form grew out of the principle of contrast, conflict, and resolution of tonalities that characterized the sonatas, symphonies, and chamber music of the 18th and 19th centuries.

BIBLIOGRAPHY. E. BORREL, *La Sonate* (1951), a general modern study; W. MELLERS, "The Sonata Principle from c. 1750," in *Man and His Music* (1957), a description of stylistic trends that led from the Baroque to the Classical attitude; M. BUKOFZER, *Music in the Baroque Era, from Monteverdi to Bach* (1947), a general survey covering the period of the pre-Classical sonata; D.F. TOVEY, "Sonata Forms," in *The Forms of Music* (1956), a brilliant study, *Essays in Musical Analysis*, vol. 1–2 (1935), perceptive program notes on specific symphonic works, *Essays and Lectures on Music* (1949), important studies of Haydn, Beethoven, Schubert, and Brahms, and *A Companion to Beethoven's Pianoforte Sonatas* (1931); F.H. MARKS, *The Sonata: Its Form and Meaning in Piano Sonatas by Mozart* (n.d.); H. KELLER,"Wolfgang Amadeus Mozart," in R. SIMPSON (ed.), *The Symphony*, vol. 1 (1966), an imaginative and learned application of modern analytical principles to Mozart's symphonies; H.C. ROBBINS LANDON, *The Symphonies of Joseph Haydn* (1955), a monumental, detailed study of one of the central areas of sonata-form history.

(B.Ja.)

Sophists

The Sophists were certain Greek lecturers, writers, and teachers in the 5th and 4th centuries BC, most of whom travelled about the Greek-speaking world giving instruction in a wide range of subjects in return for fees.

HISTORY OF THE NAME

The term sophist (Greek *sophistes*) had earlier applications. It is sometimes said to have meant originally simply "clever" or "skilled man," but the list of those to whom Greek authors applied the term in its earlier sense makes it probable that it was rather more restricted in meaning. Seers, diviners, and poets predominate, and the earliest Sophists probably were the "sages" in early Greek societies. This would explain the subsequent application of the term to the Seven Wise Men (7th–6th century BC) who typified the highest early practical wisdom, and to Pre-Socratic philosophers generally. When Protagoras, in one of Plato's dialogues (*Protagoras,* 317 a–b) is made to say that, unlike others, he is willing to call himself a Sophist, he is using the term in its new sense of "professional teacher," but he wishes also to claim continuity with earlier sages as a teacher of wisdom. Plato and Aristotle altered the meaning again, however, when they claimed that professional teachers such as Protagoras were not seeking the truth but only victory in debate and were prepared to use dishonest means to achieve it. This produced the sense "captious or fallacious reasoner or quibbler," which has remained dominant to the present day. Finally, under the Roman Empire the term was applied to professors of rhetoric, to orators, and to prose writers generally, all of whom are sometimes regarded as constituting what is now called the Second Sophistic movement (see below).

THE 5TH-CENTURY SOPHISTS

The Sophistic movement

The names survive of nearly 30 Sophists properly so-called, of whom the most important were Protagoras, Gorgias, Antiphon, Prodicus, and Thrasymachus. Plato protested strongly that Socrates was in no sense a Sophist—he took no fees, and his devotion to the truth was beyond question. But from many points of view he is rightly regarded as a rather special member of the movement. The actual number of Sophists was clearly much larger than 30, and for about 70 years, until *c.* 380 BC, they were the sole source of higher education in the more advanced Greek cities. Thereafter, at least at Athens, they were largely replaced by the new philosophic schools, such as those of Plato and Isocrates. Plato's dialogue *Protagoras* describes something like a conference of Sophists at the house of Callias in Athens just before the Peloponnesian War. Antimoerus of Mende, described as one of the most distinguished of Protagoras' pupils, is there receiving professional instruction in order to become a Sophist (*Protagoras,* 315 a), and it is clear that this was already a normal way of entering the profession.

Most of the major Sophists were not Athenians, but they made Athens the centre for their activities, although travelling continuously. The importance of Athens was doubtless due in part to the greater freedom of speech prevailing there, in part to the patronage of wealthy men like Callias, and even to the positive encouragement of Pericles, who was said to have held long discussions with Sophists in his house. But primarily the Sophists congregated at Athens because they found there the greatest demand for what they had to offer, namely, instruction to young men, and the extent of this demand followed from the nature of the city's political life. Athens was a democracy, and although its limits were such that Thucydides could say it was governed by one man, Pericles, it nonetheless gave opportunities for a successful political career to citizens of the most diverse backgrounds, provided they could impress their audiences sufficiently in the council and the assembly. After Pericles' death this avenue became the highroad to political success.

The Sophistic teaching

The Sophists taught men how to speak and what arguments to use in public debate. A Sophistic education was increasingly sought after both by members of the oldest families and by aspiring newcomers without family backing. The changing pattern of Athenian society made merely traditional attitudes in many cases no longer adequate. Criticizing such attitudes and replacing them by rational arguments held special attraction for the young, and it explains the violent distaste which they aroused in traditionalists. Plato thought that much of the Sophistic attack upon traditional values was unfair and unjustified. But even he learned at least one thing from the Sophists—if the older values were to be defended, it must be by reasoned argument, not by appeals to tradition and unreflecting faith.

Seen from this point of view, the Sophistic movement was a valuable function of Athenian democracy in the 5th century BC. It offered an education designed to facilitate and promote success in public life. All of the Sophists appear to have provided a training in rhetoric and in the art of speaking, and the Sophistic movement, responsible for large advances in rhetorical theory, contributed greatly to the development of style in oratory. In modern times the view occasionally has been advanced that this was the Sophists' only concern. But the range of topics dealt with by the major Sophists makes this unlikely, and even if success in this direction was their ultimate aim, the means they used were surely as much indirect as direct, for the pupils were instructed not merely in the art of speaking, but in grammar; in the nature of virtue (*aretē*) and the bases of morality; in the history of society and the arts; in poetry, music, and mathematics; and also in astronomy and the physical sciences. Naturally the balance and emphasis differed from Sophist to Sophist, and some offered wider curricula than others. But this was an individual matter, and attempts by earlier historians of philosophy to divide the Sophistic movement into periods in which the nature of the instruction was altered are now seen to fail for lack of evidence. The 5th-century Sophists inaugurated a method of higher education that in range and method anticipated the modern humanistic approach inaugurated or revived during the Renaissance.

NATURE OF SOPHISTIC THOUGHT

A question still discussed is whether the Sophists in general had any real regard for truth or whether they taught their pupils that truth was unimportant compared with success in argument. Plato's hostile judgment on both counts is still frequently repeated without question. The Platonic writings make frequent reference to what Plato calls "eristic" (Greek *eristikos,* "fond of wrangling") and "antilogic"; the two often have been incorrectly treated as identical. Eristic, for Plato, consists in arguments aimed at victory rather than at truth. Antilogic involves the assignment to any argument of a counterargument that negates it, with the implication that both argument and counterargument are equally true. Antilogic in this sense was especially associated with Protagoras; but Plato, no doubt correctly, attributes its use to other Sophists as well. He regards the use of antilogic as essentially eristic, whether it be used to silence an opponent by making his position seem self-contradictory, or whether it be used mechanically to negate any proposition put forward in debate. He concludes that the widespread use of antilogic is evidence that Sophists had no real regard for the truth, which must itself be free from antilogic.

But Plato himself believed, for much or possibly all of his life, that the phenomenal world was essentially antilogical inasmuch as no statement about it could be made possessing a greater degree of truth than the contradictory of that statement. For example, if a man is tall in relation to one object, he will be short in relation to another object. In so characterizing the phenomenal world, Plato certainly did not wish to be called eristic—he regarded the application of antilogic to the description of the phenomenal world as an essential preliminary to the search for the truth residing in the Platonic Forms, which are themselves free from antilogic.

Seen in this perspective, the Sophistic use of antilogic must be judged less harshly. To the extent that it was used irresponsibly to secure success in debate it was eristic, and the temptation so to use it must often have arisen. But where it was invoked in the sincere belief that antilogic

elements were indeed involved, or where it was used for analyzing a complex situation in order to reveal its complexity, then antilogic was in no way inconsistent with devotion to truth. This raises the question to what extent the Sophists possessed any general view of the world or gave expression to any genuine philosophical views, whether original or derived. Ancient writers, influenced by Plato and Aristotle, seem to have excluded the Sophists, apart from Protagoras, for their schematized accounts of early Greek thinkers. Modern writers have frequently maintained that, whatever else they were, the Sophists were in no sense philosophers. Even those who acknowledge the philosophical interest of certain particular doctrines attributed to individual Sophists often tend to regard these as exceptions and claim that, inasmuch as the Sophists were not a school but only independent teachers and writers, as a class they were not philosophers. Two questions are involved: whether the Sophists held common intellectual doctrines and whether some or all of these could actually be termed philosophical.

Philosophy of the Sophists

Among moderns, Hegel was one of the first to reinsert the Sophists into the history of Greek philosophy. He did so within the framework of his own dialectic, in which every thesis invokes its own opposite, or antithesis; thus he treated the Sophists as representing the antithesis to the thesis of the group of philosophers known collectively as the Pre-Socratics. Pre-Socratics such as Thales, Heracleitus, and Parmenides sought the truth about the external world with a bold enthusiasm that produced a series of explanations, each claiming to be correct. None of these explanations of the physical world paid attention to the observer and each was driven to reject more and more of the phenomenal world itself as unreal. Finally, with the Eleatics, a 5th-century school at Elea in Italy that held that reality is a static one, of which Parmenides and Zeno are representatives, little or nothing of the phenomenal world was left as real. This trend in turn produced a growing distrust of the power of human beings to attain knowledge of the ultimate basis of natural phenomena. Philosophy had reached an impasse, and there was a danger of complete skepticism. Such an extreme position, according to Hegel's view, provoked the "antithesis" of the Sophistic movement, which rejected the "thesis" of the objectivists and concentrated attention upon man rather than upon nature. To Hegel, the Sophists were subjective Idealists, holding that reality is only minds and their contents, and so philosophy could move forward by turning its attention to the subjective element in knowing. Reflection upon the contrast between the thought of the Sophists and that of their predecessors produced the "syntheses" of Plato and Aristotle.

Whether any of the Sophists actually were subjective Idealists may be doubted. The conclusion depends in part on whether Protagoras held that phenomena had subjective existence only, or whether he thought that all things perceived had objective existence but were perceived differently according to the nature of the percipient and their relation to him—*i.e.*, whether he interpreted phenomena subjectively or relativistically. It is fairly clear, however, that the Sophists did concentrate very largely upon man and human society, upon questions of words in their relations to things, upon issues in the theory of knowledge, and upon the importance of the observer and the subjective element in reality and in the correct understanding of reality.

This emphasis helps to explain the philosophical hostility of Plato and Aristotle. Particularly in the eyes of Plato, anyone who looks for the truth in phenomena alone, whether he interprets it subjectively or relativistically, cannot hope to find it there; and his persistence in turning away from the right direction virtually amounts to a rejection of philosophy and of the search for truth. Many a subsequent thinker for whom metaphysics, or the investigation of the deepest nature of reality, was the crowning achievement of philosophy has felt with Plato that the Sophists were so antimetaphysical that they have no claim to rank as philosophers. But in a period when, for many philosophers, metaphysics is no longer the most important part of philosophy and is even for some no part at all, there is growing appreciation of a number of problems and doctrines recurring in the discussions of the Sophists in the 5th and 4th centuries BC. In the 18th and early 19th centuries the Sophists were considered charlatans. Their intellectual honesty was impugned, and their doctrines were blamed for weakening the moral fibre of Greece. The charge was based on two contentions, both correct: first, that many of the Sophists attacked the traditionally accepted moral code; and second, that they explored and even commended alternative approaches to morality that would condone or allow behaviour of a kind inadmissible under the stricter traditional code.

Moral views of the Sophists

Much less weight has been attached to these charges since about the mid-19th century. First, many of the attacks on the traditional morality were in the name of a new morality that claimed to be of greater validity. Attacks upon particular doctrines often claimed that accepted views should be abandoned as morally defective. Furthermore, even when socially disfavoured action seemed to be commended, this was frequently done to introduce a principle necessary in any satisfactory moral theory. Thus when Thrasymachus in the first book of Plato's *Republic* argues that justice is unwarranted when it merely contributes to another's good and not to the good of the doer, Plato agrees. Finally, there is no evidence that any of the Sophists were personally immoral or that any of their pupils were induced to immoral actions by Sophistic teaching. The serious discussion of moral problems and the theory of morality tends to improve behaviour, not to corrupt it.

WRITINGS

In addition to their teaching, the Sophists wrote many books, the titles of which are preserved by writers such as Diogenes Laërtius, who probably derived them from library catalogues. It has usually been supposed that the writings themselves hardly survived beyond the period of Plato and Aristotle, but this view requires modification in the light of papyrus finds, admittedly few, that were copied from Sophistic writings in the Christian Era. It also has been possible to identify in the works of later writers certain imitations or summaries of 5th-century Sophistic writers, whose names are unknown. The most important of these are the discussion of law in the *Protrepticus*, or "Exhortation to Philosophy," by the 3rd-century AD Syrian Neoplatonist Iamblichus, and the so-called *Dissoi logoi* found in the manuscripts of Sextus Empiricus (3rd century AD). This evidence suggests that while most later writers took their accounts of the Sophists from earlier writers, especially from Plato, the original writings did in many cases survive and were consulted.

PARTICULAR DOCTRINES

As part of his defense of the Sophists against the charge of immoral teachings, the English historian George Grote (1794–1871) maintained that they had nothing in common with each other except their profession, as paid teachers qualifying young men to think, speak, and act with credit to themselves as citizens. This denial of common doctrines cannot be sustained—the evidence is against it. While the Sophists were not a sect, with a set of obligatory beliefs or doctrines, they had a common interest in a whole series of questions to which they sought to apply solutions along certain clearly defined lines.

There are great difficulties, however, in the precise reconstruction of individual Sophistic doctrines. No complete writings survive from any of the Sophists to check the accounts found in Plato, and later writers were often, but not always, dependent upon what they found in Plato. Plato doubtless knew well the doctrines of individual Sophists; but he was writing for those to whom these doctrines were already well known, and he was always more interested in following the argument where it led than in providing precise statements of other people's views for the sake of posterity. Consequently, almost everything that is said about particular Sophistic doctrines is subject to controversy.

Theoretical issues. Relativism and skepticism have often been regarded as common features of the Sophistic

movement as a whole. But it was early pointed out that only in Protagoras and Gorgias is there any suggestion of a radical skepticism about the possibility of knowledge; and even in their case Sextus Empiricus, in his discussion of skepticism, is probably right when he declares that neither was really a skeptic. Protagoras does seem to have restricted knowledge to sense experience, but he believed emphatically that whatever was perceived by the senses was certainly true. This led him to assert that the tangent does not touch the circle at a point only, but along a definite length of the circumference; clearly he was referring to human perception of drawn tangents and circles. Gorgias, who claimed that nothing exists, or if it does exist it cannot be known, or if it exists and is knowable it cannot be communicated to another, has often been accused of denying all reality and all knowledge. Yet he also seems to have appealed in his very discussion of these themes to the certainty of perceived facts about the physical world; *e.g.*, that chariots do not race across the sea. Others dismiss his whole thesis as a satire or joke against philosophers.

Relativism and skepticism

Probably neither view is correct. What Gorgias seems to have been attacking was not perceived reality nor one's power to perceive it but the attempt to assign existence or nonexistence (with the metaphysical implications of such an operation) to what we perceive around us. There is evidence that other Sophists (*e.g.*, Hippias) were interested in questions of this kind, and it is likely that they were all concerned to some degree with rejecting claims of any nonsensible existence, such as those of the Eleatics. The Sophists, in fact, were attempting to explain the phenomenal world without appealing to any principles outside of phenomena. They believed that this could be done by including the observer within the phenomenal world. Their refusal to go beyond phenomena was, for Plato, the great weakness in their thinking.

A second common generalization about the Sophists has been that they represent a revolt against science and the study of the physical world. The evidence is against this, inasmuch as for Hippias, Prodicus, Gorgias, and Protagoras there are records of a definite interest in questions of this kind. The truth is rather that they were in revolt against attempts to explain the physical world by appeals to principles that could not be perceived by the senses; and instead of framing new "objective" explanations, they attempted to explain things, where explanation was required, by introducing the perceiver as one element in the perceptual situation.

Nature and law

One of the most famous doctrines associated with the Sophistic movement was the opposition between nature and custom or convention in morals. It is probable that the antithesis did not originate in Sophistic circles but was rather earlier; but it was clearly very popular and figured largely in Sophistic discussions. The commonest form of the doctrine involved an appeal from conventional laws to supposedly higher laws based on nature. Sometimes these higher laws were invoked to remedy defects in actual laws and to impose more stringent obligations; but usually it was in order to free men from restrictions unjustifiably imposed by human laws that the appeal to nature was made. In its extreme form the appeal involved the throwing off of all restraints upon self-interest and the desires of the individual (*e.g.*, the doctrine of Callicles in Plato's *Gorgias* that might, if one possesses it, is actually right), and it was this, more than anything else, that gave support to charges against the Sophists of immoral teaching. On other occasions the terms of the antithesis were reversed and human laws were explicitly acclaimed as superior to the laws of nature and as representing progress achieved by human endeavour. In all cases the laws of nature were regarded not as generalized descriptions of what actually happens in the natural world (and so not like the laws of physics to which no exceptions are possible) but rather as norms that people ought to follow but are free to ignore. Thus the appeal to nature tended to mean an appeal to the nature of man treated as a source for norms of conduct.

To Greeks this appeal was not very novel. It represented a conscious probing and exploration into an area wherein, according to their whole tradition of thought, lay the true source for norms of conduct. If Callicles in Plato's *Gorgias* represents a position actually held by a living Sophist when he advocates free rein for the passions, then it was easy for Plato to argue in reply that the nature of man, if it is to be fulfilled, requires organization and restraint in the license given to the desires of particular aspects of it; otherwise the interests of the whole will be frustrated. Both Plato and Aristotle, in basing so much of their ethics on the nature of man, are only following up the approach begun by the Sophists.

Religion and society

Humanistic issues. The Sophists have sometimes been characterized by their attacks on the traditional religious beliefs of the Greeks. It is true that more than one Sophist seems to have faced prosecution for impiety, as did Socrates also. Protagoras wrote "concerning the gods, I cannot know either that they exist or that they do not exist nor what they are like in form," and Prodicus offered a sociological account of the development of religion. Critias went further when he supposed that the gods were deliberately invented to inspire fear in the evildoer. It is thus probably correct to say that the tendency of much Sophistic thought was to reject the traditional doctrines about the gods. Indeed this follows almost inevitably if the supposition is correct that all the Sophists were attempting to explain the phenomenal world from within itself, while excluding all principles or entities not discernible in phenomena. But in their agnostic attitudes toward the Olympian deities the Sophists were probably at one with most of the Pre-Socratic philosophers of the 6th and 5th centuries and also with most thinking people living toward the end of the 5th century. It is thus probably misleading to regard them as revolutionary in their religious beliefs.

The importance the Sophists attached to man meant that they were extremely interested in the history and organization of human societies. Here again most is known about Protagoras, and there is a danger of treating his particular doctrines as typical of the Sophistic movement as a whole. In the 5th century, human history was very commonly seen in terms of a decline from an earlier golden age. Another view supposed that there were recurring cycles in human affairs according to which a progression from good to bad would give way to one from bad to good. The typical Sophistic attitude toward society rejected both of these views in favour of one that saw human history in terms of progress from savagery to civilization. In a famous myth Protagoras explained how man achieved civilized society first with the aid of arts and crafts and then by gaining a sense of respect and justice in the ordering of his affairs. The general thinking of most of the Sophists seems to have been along similar lines.

One of the most distinctive Sophistic tenets was that virtue can be taught, a position springing naturally from the Sophists' professional claim to be the teachers of young men. But the word virtue (*aretē*) implied both success in living and the qualities necessary for achieving such success, and the claim that *aretē* could be taught by the kind of teaching that the Sophists offered had far-ranging implications. It involved the rejection of the view that *aretē* came only by birth—for example, by being born a member of a noble family—and it involved also the rejection of the doctrine that *aretē* was a matter of the chance occurrence of specified qualities in particular individuals. *Aretē*, in the Sophists' view, was the result of known and controllable procedures, a contention of profound importance for the organization of society. Moreover, what can be taught has some relation to what can be known and understood. The belief that teaching of a high intellectual calibre could produce success both for the individual and for governments has had a profound influence upon the subsequent history of education. Once again, it is through the acceptance of this doctrine by Plato and Aristotle that the Sophistic position came to be part of subsequent humanist tradition.

THE SECOND SOPHISTIC MOVEMENT

It is a historical accident that the name "Sophist" came to be applied to the Second Sophistic movement. Greek liter-

ature underwent a period of eclipse during the 1st century BC and under the early Roman Empire. But Roman dominance did not prevent a growing interest in sophistic oratory in the Greek-speaking world during the 1st century AD. This oratory aimed merely at instructing or interesting an audience and had of necessity no political function. But it was based on elaborate rules and required a thorough knowledge of the poets and prose writers of antiquity. Training was provided by professional teachers of rhetoric who claimed the title of Sophists, just as the 5th-century Sophists had adopted a name already used by others.

The revival of the Greek spirit under Hadrian and other emperors in the 2nd century AD who were also admirers of Greek culture found expression in a fresh flowering of Greek prose following principles developed and applied by the professors of rhetoric in the 1st century AD. Hence a group of Greek prose writers in the 2nd century AD were regarded as constituting the Second Sophistic movement. This was a backward-looking movement that took as its models Athenian writers of the 5th and 4th centuries BC; hence the label "Atticists" (Greek *Attikos*, "Athenian") applied to some of its leading members. The limits of the movement were never clear. It is usually taken to include Polemon of Athens, Herodes Atticus, Aelius Aristides, Maximus of Tyre, and the group of Philostrati. Dio Chrysostom of Prusa is often included, although others would regard him as preparing the way for the main period. Other writers, like Lucian, Aelian, and Alciphron, were influenced by the movement even if not properly members of it; and the writers of prose romances, such as Longus and Heliodorus, and the historians Dio Cassius and Herodian are also associated with the general trend. By the 3rd century AD, however, its impulse was weakening, and was shortly no longer distinguishable within the general stream of Greek literature.

BIBLIOGRAPHY

Fifth-century Sophists (Ancient sources and fragments): R.K. SPRAGUE, *The Older Sophists, a Complete Translation* (1972), and in K. FREEMAN, *Ancilla to the Pre-Socratic Philosophers* (1948), to be used together with Freeman's *The Pre-Socratic Philosophers*, 2nd ed. (1949). The original Greek texts are in H. DIELS and W. KRANZ (eds.), *Die Fragmente der Vorsokratiker*, 6th ed., vol. 2 (1952); and are edited with an Italian translation and commentary in M. UNTERSTEINER, *Sofisti*, 4 vol. (1949–62). (*General discussions*): T. GOMPERZ, *Greek Thinkers*, Eng. trans. by L. MAGNUS, vol. 1, bk. 3, ch. 5–7 (1901); M. UNTERSTEINER, *The Sophists*, Eng. trans. by K. FREEMAN (1954); W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 3, *The Fifth-Century Enlightenment* (1969).

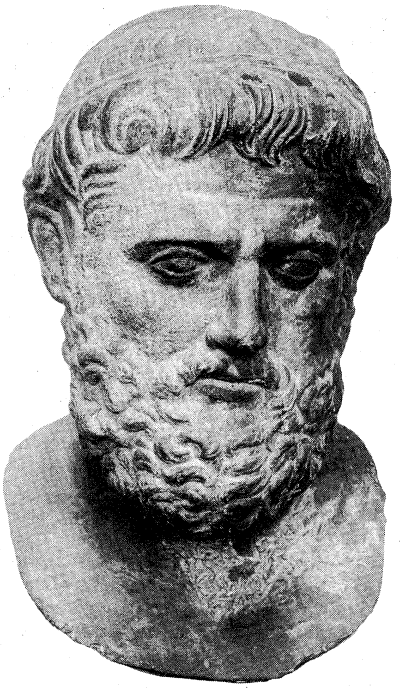
Second Sophistic period: G.W. BOWERSOCK, *Greek Sophists in the Roman Empire* (1969).

(G.B.K.)

Sophocles

More than any other dramatist of classical Greece—including his great fellow tragedians Aeschylus and Euripides—the poet Sophocles gave profound form to complex dilemmas of experience in the 5th century BC, while his native Athens grew to its fullest cultural, political, and economic development. His plays not only earned him the highest respect throughout his long life but they also translate the essential elements of classical civilization into permanently gripping theatre.

Public life. Sophocles was born *c.* 496 BC, when the Athenians had been experimenting for only a dozen years with a new, limited democratic rule. By the time he died, in 406, Athens was about to surrender to a force of Greek states (led by Sparta) that had banded together to overthrow the city's commercial and imperial "tyranny"; its democracy, too, seemed in danger of collapsing. Yet in his last play, *Oedipus at Colonus,* Sophocles sang the praises both of his own birthplace, Colonus, a village outside the walls of Athens, and of the great city itself—as though it were still in its heyday of 431, before the long war against Sparta had begun. Sophocles' faith in the city and its ideals—even during its troubles and defeats—bears out an ancient report that he refused invitations from kings to settle abroad at their courts.



Sophocles, bronze bust copied from a Greek original, 340–330 BC. In the Museo Archeologico, Florence.
Alinari

An event of Sophocles' adolescence could be said to have set the pattern for his future. Because of his beauty of physique, his athletic prowess, and his skill in music, he was chosen in 480, when he was 16, to lead the paean (formal choral chant dedicated to a god) celebrating a decisive Greek sea victory over the Persians at Salamis. At this turning point of Athenian history, which inaugurated 50 years of security, immense enterprise, maritime expansion, and cultural achievement, Sophocles was already a popular favourite, participating actively in the community and in religious ceremony and exercising outstanding artistic talents.

Public service

All the relatively meagre information about Sophocles' civic life is consistent with this. In 442 he served as one of the treasurers responsible for receiving and managing tribute money from some 300 subject states of the Athenian Empire. In 440 he was elected one of ten generals (high executive officials, as well as commanders of the armed forces); and, as a colleague of Pericles (died 429), the greatest leader of the period, he took part in an expedition to bring a wavering "ally" back into line. He served again as general perhaps two other times later. Around 421, while acting as priest for a religious society honouring a minor god of healing, he kept a snake sacred to the divine physician Asclepius in his own house until a temple under construction was ready to receive it. In 413, then aged 83, Sophocles was one of ten advisory commissioners granted special powers and entrusted with organizing Athens' financial and domestic recovery after a terrible defeat at Syracuse in Sicily.

These few facts are about all that is known of Sophocles' life. They imply distinguished, steady, earnest attachment to Athens, its government, religion, and social forms. Some scholars have inferred that Sophocles remained untroubled by the issues inherent in the growth of Athenian power, the party conflicts, and the cultural upheaval of the 5th century; and, drawing on the remarks of ancient writers, they have called him "genial" and "serene" in temperament and "orthodox" (if not reactionary) in viewpoint. If the tragedies had not survived, this conclusion might be acceptable. For Sophocles was a gentleman of the leisure class, wealthy by birth (his father probably a manufacturer of armour), highly educated, socially cultivated, noted for his grace and charm, on easy terms with aristocratic families, a personal friend of men such as Cimon, the conservative statesman, and He-

rodotus, the patriotic historian, happy in his sons (one of whom was also an esteemed tragedian), and, as antiquity saw it, "fortunate" to have died before the final surrender of Athens in 404. The plays, however, force one to realize that each of these and the other known facts is ambiguous: none proves how Sophocles felt about what he did nor what he felt about politics or people. And it is unlikely that a man blandly delighted by the glories of Athens would have staged in *Antigone* the confrontation between Antigone and Creon or in *Ajax* the hatred of Ajax for the Greek leaders, or that an uncomplicated "genial" man could have created characters such as the polluted Oedipus, the diseased pariah Philoctetes, or the bitter, wailing Electra.

The tragedies. The Sophoclean tragedies, after all, are the most decisive facts in Sophocles' life. He spent his last 65 years writing plays to be performed at the Great Dionysia festivals held annually in Athens, including their music and choreography, directing and probably training all his actors and choruses and sometimes even taking a part himself. As early as 468, when he was 28, he defeated the great Aeschylus in the festival playwriting competition, and he went on to write two dozen more plays before 450. By that date he had already modified the traditional form of tragedy: by replacing the Aeschylean trilogies with three plays each complete in itself, by extending the scope for dramatic conflict with the addition of a third speaking actor to the two that Aeschylus had adopted, and by reducing the importance of the chorus. These innovations retained the fundamental techniques and tone of tragedy but allowed Sophocles to perfect his distinctive vehicle: a complex play lasting little more than an hour that compresses varied characterization and meaning into a coherent, suspenseful situation. It was for this perfection of form that the philosopher Aristotle (in his *Poetics*, *c.* 335 BC) praised Sophocles above other tragedians, singling out *Oedipus Rex* as his masterpiece.

Participation in the Dionysia festivals

In all, Sophocles wrote 123 dramas for the festivals. This means that, since each author chosen to enter the competition usually presented four plays at a festival, he must have competed some 30 times. By any measure he was the most reliably creative of the three most famous Greek tragedians: he had a longer career than either his older contemporary Aeschylus or his younger contemporary Euripides; he wrote more, and he won more victories than either (perhaps as many as 24, compared to Aeschylus' 13 or so and Euripides' five). And this was during a period of richness in literature and the other arts probably unequalled before or since.

Only seven of Sophocles' tragedies survive (plus half a light satyr play, some fragments, and 90 titles). They provide the only access to Sophocles' inner experience and so are vital to any appreciation of his attitudes and personality. If it were known when each had been written, a chronological sketch of his development might perhaps be feasible. As it is, only a couple have fairly certain dates, and these suggest consistency more than evolution or development. Though some scholars would disagree, it seems best to assume unanimity of viewpoint among the seven.

Main subject of the tragedies

Sophocles' main subject is crisis—above all, crises of suffering or their climax, death. In Sophoclean tragedy, calamity, pain, and death are never accidental, "sad," or meaningless: in various ways they indicate or produce transformation, a transition from a spurious to a more authentic life. Deathlike agony (mental, physical, or both) leads to "rebirth" with increased understanding. This process is analogous to the transforming power of ecstasy, fertility, and growth that was embodied in the god Dionysus—a power that the Great Dionysia dramatic festival itself celebrated. It is probably also analogous to the basic spiritual content of the Mysteries (secret rites and visions) held at nearby Eleusis, into which Sophocles had almost certainly been initiated.

The cosmos. Sophoclean drama involves both characters and audience in the process of initiation and transformation, by staging the interaction of representative human figures (rather than the action of gods or natural powers). But the overall movement or pattern is more fundamental than the human figures themselves, since out of it grows the shape of each play, as well as the experience of many characters (and of the audience). This dynamic pattern is grounded in the universe as a whole, as Sophocles saw it. His conception of the universe, or cosmos, is unlike Christian, modern scientific, and other conceptions and cannot be grasped in terms of "God," "history," "matter," or "fate," though it includes analogues of all these. It is pluralistic—an interplay, or tension, of diverse forces: "divinity," "time," "nature," "necessity," and so forth. The cosmos is an ordering, a rhythmic patterning, of these forces, out of which comes the alternation of joy and despair, life and death, underlying the action of the tragedies. This universe is, in modern terms, as "natural" as it is "divine," as "personal" as it is "impersonal." Whether viewed as a coherent unity or as a driving power, it has permanency. The prime symbol for it in the plays is "the gods." "Gods" can "know everything" simply because they personify the permanent forces and structures of reality. Men, by comparison, live in dark ignorance because they are cut off from these forces and structures and are subjected to time, change, suffering, and death. Yet it is precisely time, pain, and death that can bring men into valid contact with the universal order of things. This is the main theme of Sophoclean tragedy.

Dramatic pattern of the plays

Each of the seven tragedies focusses on one phase in this process, while giving a glimpse of the whole. The main character may die early in the action (as Ajax and Antigone do); a large portion of the play afterward shows how death has justified the character and how it affects others and brings new awareness to those who survive (including the audience). Or the focus may be on the march of events and mixture of motives before the catastrophic outbreak of suffering (*Oedipus Rex, Women of Trachis*, Creon in *Antigone*). Or, again, the drama may display endurance of misery finally resulting in an act necessary and just, though grim (Oedipus at the end of *Oedipus Rex* and throughout *Oedipus at Colonus*; *Electra*; *Philoctetes*).

Death. Cosmic process generates life and fulfills humanity, but, since it always manifests itself as death and agony, it is fearful. Some Sophoclean characters naturally try to avoid the terrible trial; it overwhelms some; others meet it at least halfway (Oedipus, Antigone, Philoctetes). But, however they encounter it, all the main characters (and several secondary ones) at last "die." Many die physically, as Deianeira, Jocasta, Eurydice, Haemon, and Oedipus in the sacred grove at Colonus. Others die symbolically, as Electra in her despair at hearing Orestes is dead, Philoctetes when he fears he will be left on his island, Creon after losing his son and wife, or Oedipus in the blinding realization of his crimes. And at least two die both symbolically and physically: Ajax, whose humiliating loss of face makes life not worth living but who is ennobled by suicide; and Antigone, whose decision to oppose Creon's unjust edict foreshadows her eventual suicide in prison.

Symbolic and physical death

Whatever its form, "death" produces, beyond the expectation of even the most resolute character, a leap into a new realm of selfhood that is more in touch with truth. In this way it becomes evident that to accept reality beyond oneself, by giving up "life," can mean becoming more genuinely human, more essentially oneself. Oedipus, Ajax, Heracles, and others represent mankind passing through an ordeal by fire. Out of the fire, though, comes a personage enlarged in every respect: angrier, more tender, more compelling, more complex, grimmer, nobler. But these personages are now in fact larger than life: they have become "mythic" selves.

Mythic truth. In Sophoclean tragedy, myth symbolizes truth. The main character may learn and eventually accept the myth, or pattern, of his whole life; and this knowledge and acceptance illuminates his harsh, previously baffling experience (as with Philoctetes, Oedipus in both plays, Heracles, and, to an extent, Deianeira in *Women of Trachis*). Or the mythic pattern may emerge only in the total action of the drama, so that the audience and not any particular character grasps its significance

Dramatic function of the mythic truth

(*Ajax, Electra, Antigone*). Prophets and oracles announce these mythic truths in all the plays. Moreover, the cosmic process working through each Sophoclean tragedy is equivalent to its mythic pattern. Yet it must be realized how freely Sophocles adapted traditional stories in shaping his plays, and how myth, as he conceived and used it, is interwoven with his own specific themes, yielding multilayered dramas that no bare "plot" outline can even begin to suggest. Similarly, his characters, in becoming their essential, mythic selves, at the same time become the unique, individualized figures of Sophocles' imagination. By the end of *Oedipus Rex*, for example, Oedipus has become the paradoxical "Oedipus" of legend —"the most fortunate and most wretched of men," riddle solver, blindman, tyrant, king, criminal, and exile; but he has also become the "Oedipus" of Sophocles—aggressive, generous, domineering, preoccupied with truth, and passionate.

Freedom and responsibility. Sophocles' human agents face rending cosmic process, and so it is almost inevitable there should be controversy over his view of the relationship between universe and humanity and about the degree of freedom, autonomy, and responsibility of the individuals who so vividly confront one another in every scene. It seems reasonable to assume that Sophocles wanted to dramatize these issues rather than to eliminate them. His men and women are often sharply delineated and are free agents. But they are also "types," exemplifying attitudes, universal principles, moral qualities, social roles, and so on. No contradiction existed for Sophocles between free decision and cosmic necessity. He brought to the stage the new sense of individual autonomy that presumably developed along with wider democracy in Athens (after the 460s); but, rather than seeing divine, mythic truths blurred by novelty, diversity, and individuality, he rediscovered their permanently fresh meaning in the midst of the busy city. His cosmos, in the last analysis, includes rather than excludes man; however alien or brutal the gods or the facts may seem to men, their truth points toward the mysterious centre of man's being. Oedipus fled necessity as though it were external and imposed; he learns that it is his own actuality. Man diverges from truth by placing it "outside" himself and fighting the cosmos. This struggle would destroy man if he were merely an "individual" or an "ego." The Sophoclean tragedies show this struggle destroying the ego but thereby giving birth to indestructible spirit. *Oedipus at Colonus*, produced after Sophocles' death, expresses this most completely. The blind Oedipus is "reborn" in the course of the action, out of the extremes of calamity, enfeeblement, and pain. The sign of his regeneration is his increasing power of utterance (persuasion, command, prophecy). At the end, overtowering everyone else through his newly won awareness, he "sees" the way and leads others to the secret spot where he vanishes in a miraculous, gentle death that will transform him into a permanently protecting power.

Attitudes and personality. On the basis of this analysis of Sophoclean tragedy, it becomes possible to draw some conclusions about Sophocles the man.

Politics and community. Scholars agree that none of Sophocles' dramas is narrowly partisan or programmatic on the issues of his day. Sophoclean tragedy is, however, "political" in several other senses. The plays affirm what might be called a purified form of the Athenian community and of democracy that is grounded in the organic unity of the individual, domestic relationships, government and society, and the whole cosmic order. Usurpation of authority, abuse of power, or lack of respect for others violates this intimate unity. But, instead of "democratically" parading a number of equally valid opinions, Sophocles emphasizes that most people lack wisdom, and he always presents truth in collision with ignorance, delusion, and folly. Many scenes in each play dramatize flaws or failures in thinking, planning, decision, and communication (deceptive reports and rumours, false optimism, madness, hasty judgment), probably alluding both to the limitations of human knowledge and to weak spots in the Athenian system of popular policy making and jurisdiction. In the same way, though aware of the misery of the powerless and of the horrors of war, enslavement, tyranny, and treachery, Sophocles was no pacifist, no abolitionist, no radical democrat; the "social" virtues of compassion, pity, and sympathy are never the most highly valued in the plays. Though it is not known what he felt about specific events—not even those he took part in—it is quite certain that he located the source of the evils of life beyond the political system; he never suggests that political innovation could overcome them. His universe and his tragedies, like the Greek world of the 5th century, presuppose inescapable, unrelenting power struggles, the unpredictable rise and fall of men and states, and the impossibility of securing happiness by rational calculation.

Political themes in the plays

Vocation as poet. Though uncommercial and religious, the festivals of Dionysus at Athens were highly competitive. Some scholars consider Sophocles to have been primarily a "man of the theatre" who wrote the plays most likely to win. Yet his tragedies appear to make a less obvious appeal to popular favour than some by Aeschylus and Euripides and were surely as much "education" as "entertainment." Sophocles must have believed tragedy offered his audience wisdom that was vital to their lives. If so, he saw his work not as a replacement of nature by art nor as the imposition of order on chaos but as a product of the cosmic order imposing itself on him. Like prophet or oracle, he was spokesman for a power beyond himself—but a spokesman free and autonomous, who even wrote a theoretical treatise on dramaturgy— and the power expressed itself in an intricate, rhythmic, architectural structure of language perfectly articulating his world view. When, at an unknown date, Sophocles founded a literary society in honour of the Muses, goddesses of art, he must have been asserting his devotion both to the powers inspiring poetry and to the power of the human mind that voices it. In Sophocles' conception of poetry, man and cosmos are interdependent: human language itself can dwell in the realm of truth and can grasp the world order, giving it utterance and thereby pointing to man's essential self. Appropriately, Sophocles' last recorded act was to lead a chorus in public mourning for his rival Euripides before the festival of 406: to the end, Sophocles affirmed devotion to poetic drama itself, not merely to his own achievement in it.

Sophocles' conception of poetry

Religion. Scholars differ widely as to Sophocles' religious attitudes, some calling him a humanist, others seeing him as orthodox and pious. In the light of his sense of reality and of his own vocation, however, neither of these interpretations quite applies. His plays revere the divinity of the universe. As a playwright he was serving the religion of his community (and specifically the cult of Dionysus). He doubtless practiced ritual devotions throughout his life (as almost everyone did) and reportedly founded a shrine honouring Heracles. After his death he himself was honoured as a hero, apparently in recognition of his service to the god of healing, Asclepius. But, while the plays assume the truthfulness of the oracle of Apollo at Delphi, they cannot be said to venerate any special cult, god, or gods; in fact, they dramatize all the ambiguities of the gods' relation to human justice and virtue. One must suppose that Sophocles had an all-inclusive religious commitment of a peculiarly classical texture, hard to convey in modern terms. He lived in awe of eternal powers but did not prostrate himself before them. No ascetic, no mystic, he acknowledged divinity by his full local life. He accepted dark, uncanny forces in his ordered universe. He reverenced the wrenching alternation of life and death that yields, at times, a stark, sublime harmony. And, without being fatalistic, stoical, or resigned, he found in cruel necessity the fulfillment of man.

Personality. The man revealed by Sophocles' tragedies was as impressively and richly integrated as the common ancient and modern idealizations could wish but even more complex. Just as the surface artistry and control of the plays exist only in tension with their underlying dynamics and darkness, so there is a virtually paradoxical relationship between Sophocles' outer calm and poise and his inner grimness and passion. The plays justify admira-

tion for his stability because of their consistency over the years; but their brilliant, perplexing diversity implies his willingness to view experience from every angle. The plays support the traditional opinion that he was essentially affirmative; but the sheer pain and despair displayed in their course indicate that, if he was "orthodox," it was with an intense personal faith, in spite of the pervasive evils of his century; and that if he was "serene," it was a serenity achieved only after having faced realistically the worst that death and suffering could do.

MAJOR WORKS

Aias (probably before 441 BC; *The Ajax*); *Antigone* (441 BC); *Oidipous Tyrannos* (soon after 430 BC; *Oedipus Tyrannus*; *Oedipus the King*; *Oedipus, King of Thebes*); *Trachiniai* (possibly after 430 BC; *The Trachiniae*); *Electra* (between 418 and 410 BC); *Philoctetes* (408 BC); *Oidipous Epi Kolono* (produced posthumously in 401 BC; *Oedipus Coloneus*; *Oedipus at Colonus*).

Translations of all Sophocles' extant plays are available in the "Loeb Series" (2 vol., 1912–13); and in the "Penguin Classics Series" (1947–). Translations by R.C. Jebb (1912) and Gilbert Murray (4 plays, 1941–48) have been influential; the best current versions, however, are in the *Complete Greek Tragedies*, ed. by David Grene and Richard Lattimore, vol. 3–4 (1954–57).

BIBLIOGRAPHY. The standard Greek text is A.C. PEARSON, *Sophoclis Fabulae* (1924, reprinted 1957). Influential interpretations of Sophocles' life and works are C.M. BOWRA, *Sophoclean Tragedy* (1944); T.B.L. WEBSTER, *An Introduction to Sophocles* (1936, reprinted 1969); C.H. WHITMAN, *Sophocles: A Study of Heroic Humanism* (1951); and ALBIN LESKY, *Die griechische Tragödie*, 3rd ed. (1964; Eng. trans., *Greek Tragedy*, 1965). Discussions of individual plays include B.M.W. KNOX, *Oedipus at Thebes* (1957) and *The Heroic Temper: Studies in Sophoclean Tragedy* (1964); H.D.F. KITTO, *Greek Tragedy*, 3rd ed. (1961), *Form and Meaning in Drama*, 2nd ed. (1964), and *Poiesis: Structure and Thought* (1966). See also the film *Oedipus Rex*, in W.B. Yeats's translation, directed by TYRONE GUTHRIE (1957); the soundtrack is on Caedmon Records, as is *Antigone* in the Fitts and Fitzgerald translation.

(T.W.)

Sōtatsu

Combining the heritage of the native school of Japanese narrative scroll painting, known as *yamato-e*, with the bold, decorative designs of the great screen painters of the Azuchi-Momoyama period (1574–1600), Tawaraya Sōtatsu created an art that in both style and subject matter is traditionally Japanese and at the same time corresponds to the aesthetic sensibility of the modern age. The extraordinary popularity of this 17th-century painter developed principally in the mid-20th century, when he was recognized by Japanese and Western art critics as one of the most original and truly Japanese of all painters. He had formerly been regarded, along with the painter Kōrin, the calligrapher Kōetsu, and the potter Kenzan, as merely one of a group of decorative artists of the Edo, or Tokugawa, period (1603–1867).

Known biographical information

The dates of Sōtatsu's birth and death are unknown, and the only established facts about his life are that he came from a well-to-do merchant's family whose trade name was Tawaraya and that he was active in Kyōto from approximately 1600 to the 1630s. It is further known that he was given the rank of *hokkyō*, an honorary title conferred on distinguished artists. Only a few documents and references to Sōtatsu or his work have been found. These consist largely of three letters and a colophon, or inscription, on one of his scrolls. The reason for this paucity of documented information is that Sōtatsu did not belong to officially recognized Kanō, Hasegawa, or Tosa schools of painting, nor did he come from the nobility or warrior class (samurai), which traditionally had supplied most of the outstanding Japanese painters.

Judging from the evidence of Sōtatsu's few dated works, it generally is believed he was born around 1570 and emerged as a prominent artist by the beginning of the 17th century. Since he sometimes called himself Tawaraya Sōtatsu, it is conjectured that he was associated in Kyōto with the Tawaraya weaving factory, which is famed for its Chinese-style textiles. Probably he worked in the Tawaraya factory during the early years of his career, and only after receiving the rank of *hokkyō* did he sign his works simply Sōtatsu. A diary and a popular novel of the time refer to "Tawaraya pictures" and "Tawaraya fans," suggesting that the pictorial compositions produced by this factory enjoyed great popularity. This would explain why so many of his works are fan pictures.

Just who his teachers were and what instruction he received is not known, but his style is clearly derived from the native Japanese tradition of colourful, decorative painting rather than ink painting (*sumi-e*) in the Chinese style, although he did work in monochrome, as well as colour. His name has often been linked with that of Kōetsu, and it has been recorded that these two masters were married to sisters. There are, in fact, several scrolls dating from the beginning of the 17th century that combine the calligraphy of Kōetsu with decorative designs executed in gold and silver by Sōtatsu.

Since none of Sōtatsu's art was dated, it is impossible to establish a chronology. A group of paintings, however, that he executed for the sliding doors of the Yōgen-in in

By courtesy of the Seikado Collection, Tokyo



The "Miotsukushi" screen by Sōtatsu, illustrating a portion of the *Genji monogatari*, 6-panel *byobu* screen, colour on paper. In the Seikado Foundation, Tokyo.

Kyōto (rebuilt in 1621 for the wife of the Tokugawa shogun, the military ruler of Japan) indicates that he was already famous at that time, and it has been suggested that he may have been given the title of *hokkyō* in connection with this commission. With the Yōgen-in doors Sōtatsu must certainly have attained his artistic maturity and independence, and it is likely that he produced most of the works that he is celebrated for today during the succeeding decade.

Works attributed to Sōtatsu

Among the works attributed to him, the most outstanding are a pair of sixfold screens depicting episodes from *The Tale of Genji*, the great 11th-century novel by Murasaki Shikibu. Now in the Iwasaki Collection in Tokyo, they originally came, along with several other of his most famous works, from the Sambō-in in Kyōto. Other works of note are the twofold screen showing *bugaku* dancers in the Sambō-in and the twofold screen with the god of thunder and the god of wind in the Kennin-ji, also in Kyōto. Among works by Sōtatsu in the United States, the most outstanding is a pair of sixfold screens in the Freer Gallery of Art (Washington, D.C.) representing the pine-clad isles of Matsushima. Two other fine examples in American collections are the deer scroll in the Seattle Art Museum and the Cleveland Museum of Art's album leaf (originally one of a set of 40) illustrating an episode from the *Ise monogatari*, a 10th-century tale.

MAJOR WORKS

Very few works are definitely attributable other than these: "Bamboo, Plum, Azalea and Ivy: A Poem Scroll" (*c.* 1610; Issei Hatakeyama Collection, Tokyo); "Saigyō monogatari-e" (the only firmly dated work, *c.* 1630; Akira Watanake Collection, Tokyo); "Matsushima" (*c.* 1630; Freer Gallery of Art, Washington, D.C.); "Genji monogatari: 'Sekiya'" (*c.* 1630; Seikado Foundation, Tokyo); "Genji monogatari: 'Miotsukushi'" (*c.* 1630; Seikado Foundation, Tokyo); "Bugaku Dancers" (*c.* 1630; Sambō-in, Kyōto); "Wind God and Thunder God" (*c.* 1630; Kennin-ji, Kyōto); "Cottages in Early Spring" (*c.* 1630; Sambō-in, Kyōto); "Ise monogatari: 'The Akutagawa Story'" (*c.* 1630: Yamato Bunkakan, Ōsaka).

BIBLIOGRAPHY. YAMANE YUZO, *Sōtatsu* (1962), the best illustrated and most scholarly book on the artist (in Japanese, with English captions and summaries); WILLIAM WATSON, *Sotatsu* (1959); ELISE GRILLI, *Tawaraya Sotatsu* (1956); and JUDITH and ARTHUR HART BURLING, *Sotatsu* (1961), three brief popular introductions to the work of Sōtatsu by Western scholars.

(H.Mu.)

Soto, Hernando de

Hernando de Soto, the Spanish explorer of what was to become the southeastern United States, opened the North American continent to European colonization.

Early years

Neither the date nor the place of de Soto's birth is known with any certainty. He may have been born in 1496, 1499, or 1500, at either Villanueva de Barcarotta or Jerez de los Caballeros. Both towns lie in Extremadura, the "cradle of conquistadors," which was also the birthplace of Hernán Cortés, Francisco Pizarro, and Vasco Núñez de Balboa. De Soto spent his youth in the family manor house at Jerez de los Caballeros. His parents intended him to be a lawyer, but in 1514, while still in his teens, he told his father of his desire to go to the Indies. So, instead of taking the road to Salamanca and its famous university, he galloped toward Seville, armed like a knight and bearing the coat of arms of his family.

The age of discovery was one of excitement, especially in Spain, which was humming with marvellous tales of foreign adventure. Since Christopher Columbus' discovery of the New World 20 years before, the Spanish had been advancing on the American continent. They were established in Santo Domingo and in Haiti, from which they spread out to the Caribbean coasts, where two provinces had just been founded: New Andalusia, from Darién to Cabo de la Vela (modern Colombia), and Castilla del Oro, from Darién to Honduras (modern Panama and Costa Rica). Pedrarias Dávila had just been named their governor, and it was in his service that de Soto enlisted.

While de Soto was accompanying Pedrarias in his explorations along the coast and into the interior from 1516 to 1520, the conquistadors, greedy for treasure and jealous of each other, were falling out among themselves. Balboa, who had just crossed the Isthmus of Panama and discovered the South Sea—the Pacific Ocean—was beheaded by Pedrarias, his father-in-law. De Soto took no part in these quarrels and in 1523 participated in the conquest of Nicaragua, where for several years he was military commander. Upon the death of Pedrarias in 1531, he joined the expedition commanded by Francisco Pizarro that left Panama in search of Peru, then still unknown. In the course of Pizarro's victorious campaign, de Soto was sent as ambassador to the Inca king Atahuallpa and, in order to impress him, put on a dazzling equestrian show. He became Atahuallpa's friend, but Pizarro's brutal execution of the King and the circumstances of the Spaniards' seizure of power in Peru disgusted him with colonial life. In 1536 he returned to Spain.



De Soto, engraving from *Historia general de las Indias Occidentales* by Antonio de Herrera y Tordesillas (1559–1625).
By courtesy of the Library of Congress, Washington, D.C.

While de Soto was in Spain, his comrades were leading extended and dangerous expeditions into North America, but with neither plan nor organization, and they produced no practical result. A leader was needed, and Hernando de Soto assumed the role.

Abandoning his magnificent estate in Extremadura, he decided to head a large-scale expedition for the exploration of North America. In April 1538 he embarked from the port of Sanlúcar de Barrameda in command of ten ships, 1,000 men, and 350 horses. After a brief stop in Cuba, the expedition landed on May 30, 1539, in Florida, south of Tampa Bay.

Exploration of southern North America

De Soto left about 100 men to guard the ships and immediately headed north. At the small Indian village of Apalache, not far from present-day Tallahassee, the expedition spent the winter and explored the surrounding area. In the spring of 1540, de Soto resumed his journey. His next stop was at what is today Mobile, on the border of Alabama and Mississippi. Along the way, de Soto made contact with many Indian chiefs, some of whom welcomed him into their territory. At Mobile, on the other hand, he engaged in hard fighting with the local chief. Then the expedition resumed, heading northwest through the upper part of Alabama and Tennessee. The fort of Alibamo, near modern New Albany, Mississippi, fell into de Soto's hands after a severe engagement. On May 21, 1540, the Spaniards saw for the first time the Mississippi, the "Father of the Waters," south of Memphis. They crossed the river, made their way through Arkansas, and penetrated into the plains of Oklahoma. Along the way, they fought bloody battles against Indian tribes who used their traditional weapons, the tomahawk and scalping knife. When possible, de Soto negotiated treaties. Then, in May 1540, he turned back to the Mississippi, certain that at the end of the journey he would discover fabulously rich cities. He was never disillusioned, for, overcome by fever, he died on May 21,

1542, near modern Ferriday, in Louisiana. In order to prevent his body from falling into the hands of the Indians, his comrades buried it in the Mississippi.

Although Hernán Cortés and Francisco Pizarro, the discoverers and great captains of Mexico and Peru, have been regarded as the principal luminaries of the Spanish conquest of the New World, de Soto must be accorded a place as an excellent lieutenant, a redoubtable commander, and a leader of dangerous missions. (J.De.)

BIBLIOGRAPHY. The most scholarly tracing of de Soto's routes is the *Final Report of the United States De Soto Expedition Commission*, 76th Congress, 1st session, House Document no. 71 (10328) (1939). Directed by JOHN R. SWANTON, Smithsonian Institution, it includes an extensive bibliography and rests upon an examination of the terrain and on the de Soto narratives: Elvas, Garcilaso, Biedma, and Rangel. For Elvas, see JAMES ALEXANDER ROBERTSON (ed. and trans.), *True Relation of the Hardships Suffered by Governor Fernando de Soto . . . by a Gentleman of Elvas*, 2 vol. (1932–33). For Garcilaso, see JOHN G. and JEANNETTE J. VARNER (eds. and trans.), *The Florida of the Inca* (1951). For Rangel and Biedma, see EDWARD GAYLORD BOURNE (ed.), *Narratives of the Career of Hernando de Soto in the Conquest of Florida . . .*, 2 vol. (1904). The best biography of de Soto is THEODORE MAYNARD, *De Soto and the Conquistadores* (1930, reprinted 1969).

Sound

When a human being speaks he is said to produce sound, and anyone nearby who has adequate hearing ability is said to hear the sound. The experience is a manifestation of a phenomenon of great significance to man, who is immersed in sound stimuli that influence all of his activities.

One suggestion of the importance of sound is the large number of words in every language that are descriptive of the different kinds of sound within human awareness. In nature, the storm thunders, the gale howls, surf roars and pounds, wind whistles and moans, trees whisper, rain patters, and running water ripples and gurgles. The vocabulary descriptive of the sounds of living things is even richer: the barking and snarling of dogs, the mewing of cats, crowing of fowl, roaring of lions, hissing of snakes, lowing and bellowing of cattle, blatting of goats, grunting of pigs, chirping of birds and insects, screaming of gulls, crying and yelling of infants, not to mention the sneezing, wheezing, and snoring of humans.

In addition to the racket all around him, over which he has little or no control, man is subjected to an almost limitless variety of sounds that he himself produces. There are the unpleasant boom of the cannon, crack of the pistol, rattle of machine guns, whine of a shell, blast of an explosion, roar of engines, rumble of traffic, whine of jet planes; more agreeable generally, although distracting to some people, the whir of machinery and ticking of clocks; providing aesthetic enjoyment, the melody and harmony of music.

The purpose of this article is to set forth the physicist's explanation of the phenomena of sound, though reference will be made, when necessary, to other disciplines that play a considerable role in human production of and reaction to sound. The branch of physics most concerned with understanding sound is called acoustics, from the Greek word meaning hearing. Inevitably man's major association with sound is through the human ear. Though the anatomical and physiological details of this interesting and important sensory organ are dealt with in the text article on EAR AND HEARING, HUMAN, a few remarks here will set the stage for the further discussion of sound.

Threshold of hearing

The normal human ear is a remarkably sensitive organ, and it has been estimated that as little as 10^{-11} erg (an erg is a unit of work equal to that done by the dyne, a unit of force, moving an object through one centimetre) or about 10^{-25} kilowatt-hour of sound energy is enough to excite it and produce recognition of the sound.

The ear is built ruggedly enough to tolerate a wide range of sound energy, but hearing becomes painful, and irreversible ear damage may ensue if the amount of sound energy reaches 10^{-11} kilowatt-hour, which is, however, 10^{14} times that of the minimum audible energy level. These figures must be taken as order of magnitude values and not as specific results of precise measurement. In any case, human ears differ from individual to individual and vary even in the same individual with age and environmental conditions. The noise with which people are surrounded in large cities takes a heavy toll in physiological damage and psychological impact.

The fact that the human being has two ears enables a hearer to judge the direction from which sound in the open reaches him (the binaural effect) or to separate a signal from ambient noises. A fundamental characteristic of what may be called regular sounds, like musical notes, is their frequency—*i.e.*, the number of cycles per second expressed in hertz (one cycle per second), abbreviated Hz, of the causative vibration. Low-frequency sound is heard as low pitch, and higher frequencies as correspondingly higher pitch (see VIBRATIONS). The normal human ear is sensitive to sound within the range of about 20 Hz to 20,000 Hz, with some variation from individual to individual. Advancing age usually leads to a decrease in the upper limit. Sound of less than 20 Hz is called infrasonic; that of more than 20,000 Hz, ultrasonic. These inaudible sounds are actually of greater interest in modern physics and in their technical application than those in the sonic or audible range (see ULTRASONICS AND INFRASONICS).

This article is divided into the following sections:

I. Historical survey
II. The nature and properties of sound
    The general nature of sound waves
    The properties of sound waves
III. Sound production and perception
    Sources of sound
    The reception of sound
IV. Applications of acoustics

I. Historical survey

The fact that a solid object struck in air produces sound must have been observed from the earliest times. The additional observation that under certain conditions the sounds so produced are particularly agreeable to the ear furnished the basis for the creation of music, which must have taken place long before the beginning of recorded history. Thus the germ of the idea that the pitch of a musical sound depends somehow on the frequency of vibration of the sound-producing object is usually attributed to the 6th-century-BC Greek philosopher Pythagoras.

Further, that sound transmission through the air has something to do with the motion of the air was recognized at least by Aristotle, though it is not clear that he grasped the idea that the motion of air involved in sound is a wave motion rather than a streaming of air from the sound source to the listener. The Roman architectural engineer Marcus Vitruvius Pollio of the 1st century BC understood the wave concept and actually drew analogies between sound travelling in air and surface waves in water, but his ideas were more or less forgotten during the Middle Ages, when a corpuscular, or atomic, theory of the transmission of sound was widely held. In the 17th century, a French philosopher and mathematician, Pierre Gassendi, in reviving the atomic theory, attributed the propagation of sound to the emission and transfer of a stream of very small, invisible particles, which, after moving through the air, are able somehow to affect the ear.

Irrespective of the theory of sound transmission, the question of the speed of propagation began to engage attention in the early 17th century. Gassendi in 1635 made one of the first recorded measurements of the velocity of sound in air. His experiments involved firearms and were based on the assumption that the flash of light is seen by the observer instantaneously. Timing the delay between powder flash and the noise of the discharge, he determined a speed of sound the equivalent of 478.4 metres per second (1,569 feet per second, or 0.3 mile per second). This was shown to be too high by other observers during the 17th and 18th centuries. It was about 1750 that the first measurements made in windless, open air that can be considered at all precise were carried out under the direction of the Academy of Sciences of Paris, with a

cannon as the source of sound. The result, calibrated for 0° C (32° F), was 332 metres or 1,089 feet per second, or 0.206 mile per second. Careful repetitions during the succeeding two centuries resulted in figures differing from this last result by less than 1 percent.

Velocity of sound in liquids

The first serious attempt to measure the velocity of sound in a liquid, such as water, was apparently made in the 19th century by Daniel Colladon, a Swiss physicist who, with the assistance of a French mathematician, Charles-François Sturm, conducted such studies in Lake Geneva. The average velocity as found by their measurements was 1,435 metres (4,707 feet) per second at 8° C (46° F). It had already been established around 1808 that the velocity of sound in an iron pipe greatly exceeds that in either air or water and in fact is of the order of 5,000 metres (16,000 feet) per second.

Though the wave theory of sound had been in existence in a qualitative way for many centuries, the first attempt to apply it to a theoretical derivation of the velocity of sound was probably made by Sir Isaac Newton, who, on the basis of arbitrary assumptions, reported that the velocity of sound in air (or any gaseous fluid) should be equal to the square root of the ratio of the atmospheric pressure to the density.

Newton's derivation was criticized for giving too low a speed when appropriate pressure and density values were substituted in his equation. This difficulty was resolved by a French mathematician, Pierre-Simon Laplace, who pointed out that the Newtonian formula was based on the assumption that the compressions and rarefactions associated with the propagation of sound in air take place at constant temperature. Laplace felt it to be more reasonable to assume that the compressions and rarefactions in sound propagation do not take place at constant temperature. His assumption led to the correction of Newton's formula by multiplying the ratio of the pressure to the density by that of the specific heats (heat required to raise the temperature of one gram of gas one degree Celsius) of the gas at constant pressure and constant volume, respectively. Thus corrected, the result agreed with that from experiments at ordinary frequencies.

In the latter half of the 18th century and the first quarter of the 19th century, many scientists tried to theorize about mechanical waves in continuous media, mainly on the basis of a mathematical expression for wave motion proposed by d'Alembert in 1747. Much of this work was stimulated by the work of two mathematicians, the Swiss Leonhard Euler and Joseph-Louis Lagrange in France. These mathematical considerations made it possible to understand the vibrations of sound sources like strings, rods, membranes, plates, organ pipes, and musical instruments in general. Though earlier work—based largely on empirical or relatively simple mathematical analysis—had been carried out, it required the methods of the calculus to understand in any precise way the behaviour of even the simplest types of sound vibrators.

Mathematical analysis of wave motion

The ear early attracted scientific attention as the receiver of sound for human beings. A French physicist, Félix Savart, in 1830 attempted to establish the frequency limits of audibility and obtained results not too different from those of modern measurements made with more sophisticated equipment. The corresponding problem of the threshold of audibility was attacked around 1870 by the Viennese physicists, August Toepler and Ludwig Boltzmann, who arrived at the figure 10^{-11} watt (one watt is a unit of power equal to one joule of work per second) per square centimetre, considerably in excess of the modern figure of 10^{-16} watt per square centimetre but indicating, nevertheless, the remarkable sensitivity of the ear as a sound receiver. Further important studies on the ear were made in Germany by Georg Simon Ohm, a physicist, who in 1843 theorized that the ear is able to analyze any complex sound into a combination of simple tones in terms of which the sound may be expressed mathematically by the use of a theorem of a French mathematician, Jean Fourier (see ANALYSIS, FOURIER). Ohm's work was further extended by Hermann von Helmholtz, a German physicist and physiologist, whose book, *Sensations of Tone*, marked the beginning of physiological acoustics.

Though early scientific work in the production and reception of sound in closed spaces (*i.e.*, room or architectural acoustics) was reported around 1850, the modern quantitative foundation of the subject dates from the work of a United States physicist, Wallace Clement Sabine, who in 1900 found a law to connect reverberation (echo) time in a closed room with the volume of the room and the amount of absorbing material. His finding made applied architectural acoustics possible.

Rayleigh's contributions

In another sense, however, the modern period in acoustics may be said to date from the publication in 1877 by the English physicist Lord Rayleigh (John William Strutt) of *The Theory of Sound*, in which the whole basic theory of all aspects of the production, propagation, and reception of sound was presented in rigorous fashion, together with much of the relevant experimental material. Most of 20th-century acoustics is an exploitation and development of the fundamental ideas in Rayleigh's treatise with the use of more sophisticated equipment.

II. The nature and properties of sound

THE GENERAL NATURE OF SOUND WAVES

Wave motion, velocity, and intensity. As has been noted, it was early recognized that sound is propagated as wave motion from a sound-producing object. It is important to distinguish this motion from that of such things as automobiles, projectiles, and flowing water. When a person opens his mouth and utters a sound, he moves the air in front of his mouth, but he does not push away all of the air in his vicinity. Instead, the motion is a squeezing one that compresses a small mass of air near his mouth into less than its normal volume. Being elastic, air when compressed tends to expand again when the compressing influence is removed. In expanding, the first mass of air moves some of the air adjacent to it, which is thereupon compressed in turn and in turn expands and pushes against another adjacent bit of air. So the original compressing motion is passed on in temporal succession through the space surrounding the speaker until it reaches a receiver of sound like the human ear and finally yields up some of itself to the receiver as a stimulus. The response in the case of the human ear is audition or hearing.

A sound wave in air or any fluid medium is thus a compressional wave. The simplest example is the pressure pulse produced by the firing of a pistol or a short speech ejaculation: it takes time for such a pulse to travel through the medium, and hence such a wave has a finite velocity. The velocity of sound in air is about 344 metres (1,128 feet) per second at 20° C (68° F). It increases with temperature in a way that is treated below in the section on the properties of sound waves. It is independent of the pressure or density of the air over wide ranges. Because the velocity of sound in air is so much less than the velocity of light (3×10^8 metres per second), the sound of a distant cannon is heard sometime after the puff of smoke is seen.

The everyday distinction between loud and soft sounds rests on the differences in the magnitude of the pressure changes involved in sound-wave propagation. In normal conversation the pressure of the air in front of the mouth of the speaker is changed at most by only about one millionth of the normal atmospheric pressure (*i.e.*, by only about one dyne per square centimetre out of 10^6 dynes per square centimetre; the dyne, a unit of force, is that force required to change the velocity of a mass of one gram one centimetre per second each second). At the same time the accompanying motion that produces the compression corresponds to a fluctuating air-flow velocity with a maximum value of only about 0.02 centimetre per second (about 0.008 inch per second).

A sound wave represents the transmission of mechanical energy. An average rate of transfer of only 10^{-16} watt per square centimetre is enough to produce the hearing of an identifiable sound in a normal young person. This means that mechanical energy of only about 10^{-11} erg or about 10^{-25} kilowatt-hour can produce in the ear an identifiable sound stimulus. The average power transmission in a sound wave per unit area of the medium perpendicular to the direction of the wave is called the intensity. The unit

Definition of decibel

for this quantity is the decibel (abbreviated dB). If the intensity of a given sound in absolute units is I and the minimum audible intensity is I_0, the sound is said to be a number of decibels above the minimum audible equal to ten times the logarithm to the base ten of I divided by I_0. The decibel scale is thus relative to some arbitrary level, but the latter is commonly taken as the minimum audible. On this basis, ordinary conversational speech has an intensity of 60 decibels at a distance of about three feet from the mouth. Traffic at a busy street intersection can provide an average of 75 decibels, and a boiler factory can reach 130 decibels.

As a source of sound in the open, such as an airplane, moves away from the listener, it is heard less distinctly. This is basically a matter of geometry. The intensity of a sound wave from a point source varies inversely as the square of the distance from the source. A point source in the open spreads sound energy in all directions, and the farther from the source, the greater is the spherical area through which the same sound energy flows per unit time. Because the area of a sphere increases as the square of the distance from the centre, the intensity of the wave, which is the average flow of energy per unit area, must decrease inversely with the square of the distance. Actual observation indicates that the falling off of intensity with distance in the open is always greater than this purely geometrical decay. This effect is attributed to reflection, refraction, diffraction, and scattering, as well as to the absorption of the sound caused by the viscosity and other similar properties of the medium. The diminution of sound intensity from a localized source is further complicated if the source is on the ground or underwater near the surface, and even more so if it is inside a closed room or auditorium.

Frequency, pitch, and wavelength. An air siren produces sound waves by interrupting jets of air at different rates, by means of a rotating disk with orifices in its periphery. The greater the number of separate puffs of air per unit time, the higher is the perceived pitch. This leads to a consideration of periodic sound waves, which are much the easiest to study analytically (see the section on the properties of sound waves). Waves are called harmonic waves because they result from a source of sound in which the disturbance repeats itself regularly. The simplest periodic sound, a harmonic sound wave, is characterized by a definite frequency. The frequency is defined as the number of times per second the disturbance at any point in the medium transmitting the wave is repeated.

The relation between pitch, as perceived by the human ear, and physical frequency is more complex in the case of the human voice or ordinary musical instruments than it is for a pure harmonic wave as emitted by a carefully made tuning fork or loudspeaker, activated by an electronic oscillator under precise frequency control. The

Meaning of pure tone

fork or speaker is said to produce a pure tone of definite single frequency. In such a harmonic wave the time for one complete cycle of pressure change at any point in the medium is called the period of the wave, which is one divided by the frequency in cycles per second, or hertz.

The sounds of all frequencies are not heard equally well. An object vibrated in air at a frequency of less than about 20 hertz will emit a compressional wave but will not be heard as sound by most normal ears. Such waves are said to be infrasonic. An object vibrating at a higher frequency than this will tend to be audible if the intensity is sufficiently great. When the frequency of a compressional wave in air is increased beyond 20,000 hertz (20 kilohertz), the audibility ceases for most people, even if the intensity becomes great, though an uncomfortable nervous sensation is experienced by many. With advancing age of the listener, the frequency threshold decreases decidedly. These inaudible high-frequency sound waves, called ultrasonic, play an important part in modern acoustical science and engineering. By means of appropriate sources (see below *Sources of sound*) it has been possible to produce ultrasonic radiation of frequency of about 10^{10} hertz. Such waves, however, are rapidly absorbed in fluid media.

Every harmonic sound wave is characterized by a definite wavelength; that is, the distance between identical successive points in a spreading wave at which the disturbance is exactly the same and doing the same thing (*i.e.*, increasing or decreasing the compression of the fluid). At two such successive points the disturbances are said to be in phase with one another. Because the disturbance (*e.g.*, the onset of compression) in a wave travels a distance of one wavelength in one period, the product of wavelength and frequency equals the wave velocity. This important relationship between wavelength, frequency, and velocity, which holds for harmonic waves of all kinds (*i.e.*, sound waves, light waves, water waves, etc.), means that high-frequency acoustic waves in a given medium have smaller wavelengths than low-frequency sounds. Thus, in air at 20° C (68° F), a wave of frequency 1,000 hertz has a wavelength of 0.344 metre (1.128 feet), whereas a wave frequency 10^6 hertz or one megahertz has a wavelength of 0.344 millimetre (0.0135 inch). The corresponding values in water are, respectively, 1.5 metres (five feet) and 1.5 millimetres (0.06 inch).

Few sounds in daily experience are pure tones characterized by a single frequency; most can be considered as arising from more or less complicated combinations of harmonic waves, each with its own frequency. One or more frequencies will often predominate (*i.e.*, have greater intensity) and help give the sound its observed quality.

THE PROPERTIES OF SOUND WAVES

Wave functions. The simplest of all wave motions to imagine are waves that travel along a flexible rope when one end is flicked. This type of wave is called transverse because the disturbance of the medium (in this case the rope) takes place at right angles to the direction of travel. A transverse wave differs physically from a sound wave in which the disturbance (compression) takes place in the same direction as the wave itself; thus sound waves are called longitudinal. Each particle of the medium in a longitudinal wave is displaced from its equilibrium position along the direction of the wave propagation.

The longitudinal nature of sound waves

Though sound waves are not necessarily periodic, those of greatest interest are characterized by frequency and wavelength (*i.e.*, are periodic). Like transverse waves, these longitudinal sound waves can be expressed in a harmonic equation in which the displacement, or disturbance (ξ), is equal to the amplitude (A) times the sine of two pi (π), times a quantity equal to the frequency (f), times the elapsed time (t), minus the distance (x) from the origin divided by the wavelength (λ), or:

$$\xi = A \sin 2\pi (ft - x/\lambda). \qquad (1)$$

The amplitude of the wave represents the maximum value of the disturbance. This equation can be used to determine the size of the disturbance for any value of the time (t) or position (x). In the equation, the quantity inside the parentheses can take successive values of 0, ⅛, ⅜, ⅜, etc., so that the disturbance is equal to the amplitude times the sine of 0, $\pi/4$, $2\pi/4$, $3\pi/4$, etc.; that is, the disturbance is alternately equal to plus and minus the amplitude and passes through zero in between.

Concept of excess pressure

For a harmonic sound wave in a fluid, the displacement ξ is more conveniently replaced by the excess pressure (p_e) produced by the wave. Excess pressure may be defined as the difference between the actual pressure (p) in the medium at any place and time and the normal pressure (p_0) in the originally undisturbed medium. Thus, symbolically, $p_e = p - p_0$. This excess pressure for a harmonic wave varies in space and time like ξ as shown in equation (1).

Velocity of sound waves in fluids and solids. A mathematical study of the propagation of a compressional wave in a fluid (gas or liquid) also shows that the velocity (V) of the wave is equal to the square root of the ratio of excess pressure to excess density:

$$V = \sqrt{p_e/\rho_e}, \qquad (2)$$

in which p_e is the excess pressure in the medium produced by the wave and ρ_e is the associated excess density (difference between the actual and normal, $\rho_e = \rho - \rho_0$). Whenever a fluid is squeezed by the application of extra

pressure, the density goes up, and when the excess pressure becomes negative the density goes down. Equation (2) can be used to provide expressions from which the velocity V may be calculated in special cases. Hooke's law of elasticity (see ELASTICITY) states that elastic stress is always proportional to elastic strain; that is, stress divided by strain is equal to a quantity B, in which B represents a proportionality constant. In the case of a compressional wave in a fluid, provided these quantities are not too great, one can write:

$$\frac{p_e}{\rho_e/\rho_0} = B, \tag{3}$$

in which ρ_0 is the average density of the undisturbed medium. The quantity B, called the bulk modulus of the medium, is the coefficient of proportionality between the excess pressure p_e (the stress) and the fractional excess density ρ_e/ρ_0 (the strain). By substituting equation (3) in (2), one obtains the formula for the sound velocity: the velocity is equal to the square root of the ratio, bulk modulus divided by the average density of the undisturbed medium, or,

$$V = \sqrt{B/\rho_0} = \sqrt{1/\rho_0 K} \tag{4}$$

in which K, called the compressibility, is the reciprocal ($K = 1/B$) of the bulk modulus. For example, the bulk modulus of water has been measured to be (around room temperature) 2.1×10^{10} dynes per square centimetre. Because the density of water under the same condition is about one gram per cubic centimetre, one has a numerical value for the velocity of sound in water under these conditions (from equation [4]),

$$V = \sqrt{\frac{2.1 \times 10^{10} \text{ dynes/cm}^2}{1 \text{ gram/cm}^3}} = \sqrt{2.1} \times 10^5$$

centimetres per second, which comes out to be 1,450 metres (4,756 feet) per second, a figure in good agreement with the experimentally measured velocity. As a matter of fact, confidence in equation (4) is such that it is commonly used to determine the bulk modulus of liquids from the observed velocity of sound, in place of the static compression method that demands great precision, because liquids are hard to squeeze.

It can be noted from equation (4) that the velocity is expressed in the form of the square root of the ratio of a factor (B) that depends on the elasticity to a specific inertia factor (ρ_0). This type of dependence holds for elastic waves in general (*i.e.*, in solids as well as in gases and liquids). It means that for a given specific inertia factor (such as density), the more elastic a medium is the greater is the sound velocity. It may be that the two factors are interdependent, and then the sound velocity formula can take on a different form. This is true, for example, for gases, as is made clear in the next paragraph.

To use equation (2) for gases requires the so-called equation of state, which connects pressure, density, and temperature. For an ideal gas (see GASEOUS STATE) the equation of state says that the pressure (p) divided by the density (ρ) is proportional to the absolute temperature, or

$$p/\rho = RT, \tag{5}$$

in which T is the absolute temperature (Celsius temperature plus 273°) and the constant of proportionality, R, is the ideal gas constant per gram. It can be shown that if the temperature remains constant the ratio of the excess pressure to the excess density is also equal to RT—that is, $p_e/\rho_e = RT$—and therefore that the velocity of sound is equal to the square root of the gas constant times the absolute temperature and also equal to the square root of the ratio of the pressure to the density; *i.e.*, $V = \sqrt{p/\rho}$.

Newton's law of sound propagation

As noted above, this is the result obtained by Newton. It does not agree with experiment. If one follows Laplace's assumption that the pressure and density changes that take place in a gas in sound transmission are adiabatic instead of isothermal—that is, the temperature in a gas through which a sound wave passes does not remain constant—one must use the equation in which the pressure divided by the density raised to the γ power is a constant, or p/ρ^γ = constant, in which γ (gamma) is equal to c_p/c_v, the ratio of the specific heat of the gas at constant pressure (c_p) to the specific heat at constant volume (c_v). Then, the velocity of sound is equal to the square root of gamma times the ratio of the actual fluid pressure to its density, or

$$V = \sqrt{\gamma p/\rho}. \tag{6}$$

This formula agrees so well with experiment at sonic and low ultrasonic frequencies that it is often employed to determine the value of gamma from measurement of sound velocity. It might at first be supposed that if the sound-induced changes in pressure took place adiabatically at audible frequencies, they would remain adiabatic as the frequency increases. That would follow from the reasoning employed by Laplace to deduce equation (6). Laplace's reasoning was faulty, however, and at very high ultrasonic frequencies (perhaps of the order of 10^{12} hertz) the velocity of sound follows the Newtonian formula.

For non-ideal gases and vapours—*i.e.*, those that do not obey the equation of state expressed by equation (5)—the simple formula (6) must be corrected, by methods given in advanced treatises.

In any case, for an ideal gas at what may be called ordinary frequencies, equation (6) shows that the velocity of sound does not depend on pressure or density but only on the absolute temperature. Thus equation (6), with the use of the equation of state, states that velocity is equal to the square root of the product of gamma, the gas constant, and the absolute temperature, or $V = \sqrt{\gamma RT}$. This equation can be rewritten in more convenient form. If the velocity of sound in dry air at 0° C (32° F) is denoted by V_0 that at temperature $t°$ C becomes equal to the velocity at 0° C times a factor, the square root of the quantity 1 plus the Celsius temperature divided by 273, or

$$V_t = V_0\sqrt{1 + t/273}.$$

For dry air under standard conditions (*i.e.*, 0° C), $V_0 =$ 331.3 metres (1,086.7 feet) per second. For room temperature (20° C or 68° F) the equation then yields a velocity of 344 metres (1,128 feet) per second or 0.214 mile per second.

As the preceding equations indicate, under what may be termed ordinary conditions, the velocity of sound even in an ideal gas and indeed in all actual gases is independent of the frequency—*i.e.*, there is no so-called dispersion (see WAVE MOTION). If, however, the ratio of frequency to pressure (*i.e.*, f/p) increases sufficiently, the velocity of sound, even in a relatively permanent gas like hydrogen, shows an increase. In the case of hydrogen, for example, as the ratio of frequency to pressure goes from one megahertz per atmosphere to 30 megahertz per atmosphere, the sound velocity increases by about 8 percent from its standard low frequency value of 1,284 metres (4,212 feet) per second.

Absence of dispersion

In the case of liquids to which equation (4) applies, strictly speaking, the adiabatic bulk modulus or compressibility must be used. For liquids, however, these values as a rule differ little from their isothermal counterparts. Because the dependence of the compressibility of a liquid on the temperature is a complicated affair, it is not possible to derive simple equations to express the dependence of the sound velocity on temperature. Empirical formulas found in advanced treatises show that the sound velocity in water, for example, increases as the temperature rises to 74° C (165° F) and thereafter decreases. For most other liquids, decrease of velocity with increase in temperature is the rule.

Compressional waves in solids are usually considered to be sound waves. Their velocities follow the pattern described in the beginning of this subsection, viz., they are expressible as the square root of the ratio of an elasticity factor to a specific inertia factor. The velocity (V) of a compressional wave in a long solid rod, for example, is expressible in the following way: the velocity (V) is equal to the square root of Young's modulus (the "stretch" modulus of elasticity, Y) divided by the equilibrium density (ρ_0) of the rod material; *i.e.*, $V = \sqrt{Y/\rho_0}$.

For most hard solids the velocity here has a range of values from about 1,000 metres (3,300 feet) per second to 6,000 metres (19,800 feet) per second, depending on the elasticity and density. Thus the compressional wave velocity in a steel rod is many times that in a lead rod, as the density of steel is less than that for lead, and the Young's modulus for steel is about 13 times that of lead.

For a three-dimensional solid—*i.e.*, one not in the form of a long thin rod—the compressional wave velocity is similar to equation (4), except that the bulk modulus is increased by a quantity connected with the shear modulus (see ELASTICITY). It is $V = \sqrt{(B+4\mu/3)/\rho_0}$, in which B is as usual the bulk modulus, ρ_0 is the equilibrium density, and μ is the shear modulus.

Wave fronts and rays. It has already been pointed out that an audible sound wave tends to spread out in all directions from the source. It is helpful in describing this situation to introduce the concept of wave front: an imaginary surface of such a kind that at any instant the disturbance characterizing the sound wave is the same at every point of it.

An example is provided by a bird radiating a sound wave into the surrounding air. At any instant the state of compression of the air in this wave will be the same (to a good approximation) on the surface of a sphere with its centre at the bird. This sphere is a spherical wave front. The whole space surrounding the quasi-point source is filled with such wave fronts, each with its appropriate phase of the disturbances. The propagation of the wave can be thought of as equivalent to the spreading of the sphere throughout space as its radius grows larger at a rate equal to the speed of sound in the medium. The expanding ripples on the surface of water produced by a dropped stone provide a clear analogy. A flat plate, on the other hand, vibrating in a direction perpendicular to its surface alternately compresses and rarefies the air and so produces a wave front that, if confined in a tube of uniform cross section, is approximately a plane wave front. Plane and spherical wave fronts are the two most important varieties in the study of sound-wave propagation.

Spherical wave front

Though wave fronts are an essential element in the description of sound-wave propagation through three-dimensional space, it is often convenient to describe the propagation in terms of rays—*i.e.*, lines perpendicular to a wave front. The ray passing through a particular point at a given instant indicates the direction in which the wave transmission is taking place at that time and place. Sound rays can often be useful, though not so much so as light rays, as it is harder to produce a genuine beam of sound concentrated in one direction because the wavelength of sound is in general so much larger than that of light. At ultrasonic frequencies, however, it is possible to approximate sound beams, and then ray acoustics becomes of value.

Reflection and refraction. The existence of an echo when a sound is emitted in the neighbourhood of a large, hard surface is a manifestation of the fact that sound waves can be reflected when they strike a surface separating media of different properties. Moreover, sound-wave fronts are bent (refracted) when sound crosses such a boundary; *e.g.*, in going from air to water or vice versa.

Huygens' principle of wave propagation

The laws governing the reflection and refraction of sound waves are best understood in terms of a fundamental principle stated by a Dutch scientist, Christiaan Huygens. According to this principle each point on a wave front (F_1 in Figure 1, for example) may be assumed to be the source of a hemispherical wavelet that moves outward from F_1 in the direction of propagation. The new wave front after a short time is the mathematical envelope of all these wavelets—*i.e.*, the surface tangent to them all (see F_2 in Figure 1). A simple but important illustration of Huygens' principle is the establishment of the laws of reflection and refraction of a plane wave at a plane interface separating two fluid media with different densities and different sound velocities. In Figure 2, SS' represents the trace (intersection) in the plane of the page of the plane interface between a fluid medium (I) in which the velocity of sound is V_1 and the mean density is

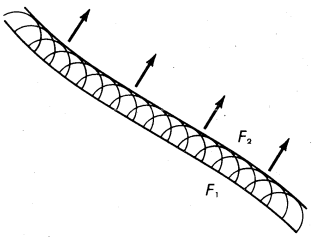


Figure 1: Huygens' construction showing points on wave front F_1 generating wavelets to produce new wave front F_2 (see text).

ρ_1, and a medium II in which the corresponding quantities are V_2 and ρ_2, respectively. AB denotes the trace of a plane wave front (perpendicular to the plane of the page) approaching the interface SS' from medium I. If the wave front AB is incident on SS' as CD with angle of incidence i, as in the figure, then AC and BD are incident rays that make an angle i with the normal CN to the interface. The reflected and refracted wave fronts can be constructed by means of Huygens' principle. At the instant the disturbance at one end of the incident wave front has reached C on the interface, the disturbance at the other end is still at D in medium I. In a time equal to the distance DE divided by the velocity (DE/V_1) that it takes the disturbance at D to reach E on the interface, the disturbance will travel out in all directions from C. In particular, in medium I it will traverse a distance equal to DE.

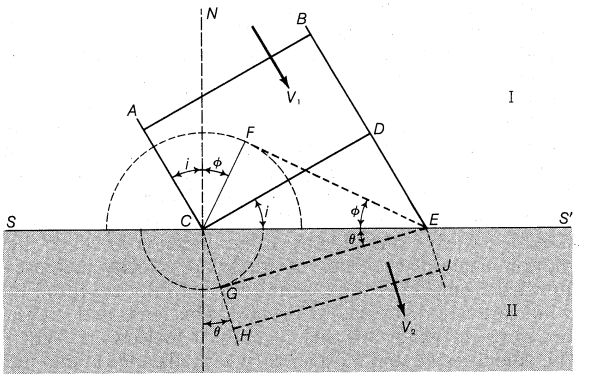


Figure 2: Huygens' construction showing wave fronts AB and CD that are reflected as FE at the boundary between media I and II and refracted as GE and HJ to give the law of reflection and Snell's law of refraction (see text).

If, with C as centre, a semicircle is drawn in medium I with radius DE and a straight line from E is drawn tangent to this, the result is EF, which makes the angle $\phi = i$ with SS'. That EF is the trace of the reflected wave front can be confirmed by making a Huygens' construction for other points on the incident wave front. It is seen that the reflected wave front makes the same angle with the interface as the incident wave front. This, combined with the fact that both wave fronts are perpendicular to the same plane (plane of the page), constitutes the law of reflection. Equally well expressed in terms of the incident and reflected rays, this law is the same law that holds for light, namely, that the angle of reflection is equal to the angle of incidence.

The same kind of Huygens' construction shows that, after transmission across the interface, the incident wave front CD is changed to the refracted wave front GE, making the angle θ with SS', such that the ratio of the sine of the angle of incidence to the sine of the angle of refraction is equal to the ratio of the velocity in the first medium to that in the second—*i.e.*,

$$\sin i/\sin\theta = V_1/V_2 = n_{12}. \qquad (7)$$

This is Snell's law for the refraction of plane waves at a plane interface, with n_{12} called the index of refraction of medium II with respect to medium I. A similar law, in somewhat modified form, holds for the transmission of wave fronts through a medium with continuously varying

properties. The law as expressed in equation (7) accounts well for the passage of sound waves from air to water, or vice versa, when the media are at rest.

In Snell's law, equation (7), if the velocity in medium I is greater than that in medium II ($V_1 > V_2$), the index of refraction is greater than one ($n_{12} > 1$) and refraction takes place for every angle of incidence. If the velocity in medium I is less than in medium II ($V_1 < V_2$), however, the index of refraction will be less than one ($n_{12} < 1$), and when the incident ray makes an angle i such that sin $i = n_{12}$, Snell's law gives sin $\theta = 1$; *e.g.*, the refracted ray just grazes the interface in the second medium—that is, the refracted wave front moves so as to remain perpendicular to the interface. For sin i greater than n_{12}, as may well be possible since the largest value of sin i is equal to one, sin θ becomes greater than one, which is mathematically impossible. The physical result under this condition, is that no refraction takes place and total reflection ensues. The angle (arc sin n_{12}) is called the critical angle of incidence and for angles greater than this, all the sound is reflected, analogously to the similar phenomenon in light.

Total reflection

When the two media—*i.e.*, air and water—are moving relative to each other through the action of wind and water currents, the refraction of sound becomes more complicated and so the result will be given only for a relatively simple case in which the velocity of flow of medium I is U_1 parallel to the interface and that of medium II is U_2, also parallel to the interface. Then the law of refraction becomes:

$$V_1/\sin i - V_2/\sin \theta = U_2 - U_1. \qquad (8)$$

If the two velocities are equal ($U_2 = U_1$)—*i.e.*, the two media are moving together with no relative velocity—equation (8) reduces to equation (7). Equation (8) also applies, as more elaborate analysis shows, when the sound velocity and flow velocity vary continuously in a stratified or layered medium in which the flow is parallel to the layers. Then all quantities in (8) vary from place to place in a direction perpendicular to the stratification. Examination discloses that even if velocities V_1 and V_2 are constant, the effect of the motion of the medium is to bend the wave front and ray in the direction in which the motion takes place. Thus, if the wind velocity increases upward from the ground, sound-wave fronts are lifted to windward and depressed to leeward. This tends to decrease the range to windward and increase it to leeward. When the velocities V_1 and V_2 also vary with distance above the ground, as is commonly the case in the atmosphere, the wave-front propagation becomes even more complicated. Nevertheless, many practical cases have been worked out successfully. These considerations are of great importance in sonic gun-ranging.

Energy propagation: intensity of sound. As has already been pointed out, the propagation of sound may be considered as a form of transmission of energy through the material medium. When, for example, a fluid is disturbed at any point, work is done, and energy is expended. The reappearance of the disturbance at a distant point by wave propagation corresponds to the transfer of the original energy. The average rate of energy transfer per unit time per unit area of the wave front is, by definition, the intensity of the wave. This may also be expressed as the average flow of power per unit area, the practical unit of which is the watt per square centimetre (watt/cm^2).

The intensity in a plane sound wave may be evaluated by finding the time average of the product of the excess pressure (p_e) and the flow velocity ($\partial\xi/\partial t$) in the medium. Inasmuch as pressure is force per unit area, the product just cited is valid as far as the consistency of dimensions is concerned. (The units of pressure times velocity are dyne centimetre^{-2} · centimetre second^{-1}, whereas the units of power per unit area are erg second^{-1} centimetre^{-2}. These are equivalent, as an erg is a dyne centimetre.) It is necessary to take the average of the product in question because both the excess pressure and the flow velocity vary in time and space as the wave progresses. Hence the flow of energy actually fluctuates, and to get a constant effective measure of the flow of energy per unit area, an average must be taken. The appropriate mathematical analysis shows that the intensity of a plane sound wave is given by:

$$I = p^2_{e,\,max}/2\rho_0 V. \qquad (9)$$

in which $p_{e,\,max}$ is the maximum excess pressure on the wave, ρ_0 is the average equilibrium density of the medium, and V is the sound velocity. This expression is independent of frequency. The use of the decibel notation in the representation of intensity has already been described in the section on the general nature of sound waves, and examples were given. It is sufficient to recall here that for ordinary conversational speech about one metre from the speaker's mouth, the maximum excess pressure is one dyne per square centimetre (dyne/cm^2). For dry air at 20° C (68° F) and normal atmospheric pressure, the equilibrium density (ρ_0) is 0.001205 gram per cubic centimetre, and the velocity (V) is 344 metres (1,128 feet) per second. Substitution of these values into equation (9) and subsequent division by 10^7 to convert to watts per square centimetre yields an intensity (I) of about 10^{-9} watt per square centimetre, or 70 decibels above the minimum audible intensity of 10^{-16} watt per square centimetre.

Intensity of speech

From equation (9) it follows that the intensity for given excess pressure depends on the medium and in particular on the product equilibrium density times sound velocity ($\rho_0 V$). The latter quantity is known as the specific acoustic resistance of the medium for a plane wave. For water, for which the specific acoustic resistance is about 3,800 times that for air under standard conditions, it takes an excess pressure some 60 times that for air to produce the same intensity. It will be seen below in the section on sound sources that for given frequency and excess pressure a solid sound source is a more efficient radiator in water than in air.

The intensity of a spherical wave diverging from an effective point source can still be written in the form given by equation (9), provided it is remembered that for this type of wave the maximum excess pressure is no longer constant but varies inversely as the square of the distance from the source. Hence the intensity as given in equation (9) still applies if the maximum excess pressure is the excess pressure amplitude at the place where the intensity is to be evaluated.

Flow of sound energy across a boundary. When a sound wave strikes the interface separating two media of different physical properties (*e.g.*, air and water), as was already seen, some of the wave disturbance is reflected and some transmitted (and in general refracted) through the second medium. The transmission of the sound energy is thus affected by the presence of the boundary.

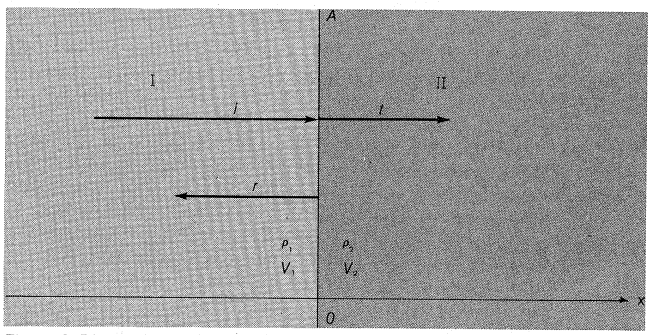


Figure 3: Direction of incident, transmitted, and reflected waves at boundary *A0* between media I and II (see text).

Figure 3 represents a plane boundary $A0$, separating two media I and II, that is perpendicular to the direction of plane wave propagation along the x axis. In medium I the mean equilibrium density is represented by ρ_1 and the sound velocity by V_1. The corresponding quantities in II are ρ_2 and V_2, respectively. The arrow i denotes a plane harmonic sound wave travelling from left to right in medium I in the x direction and striking the boundary at x equal to zero. Similarly, r denotes the wave in I reflected from the interface, whereas t denotes the wave transmitted across the boundary into medium II.

The amplitudes of the excess pressure in the incident, reflected, and transmitted waves may be represented by

p_i, p_r, and p_t, respectively. These quantities are not independent inasmuch as the sum of the incident and reflected excess pressure amplitudes must equal the transmitted excess pressure amplitude. There is a similar condition connecting the flow-velocity amplitudes in the incident, reflected, and transmitted waves. The result of applying these two conditions can be expressed by the following two equations:

$$\frac{p_r}{p_i} = \frac{1 - \rho_2 V_2/\rho_1 V_1}{1 + \rho_2 V_2/\rho_1 V_1}, \tag{10}$$

$$\frac{p_t}{p_i} = \frac{2}{1 + \rho_1 V_1/\rho_2 V_2}. \tag{11}$$

In the case in which the density times velocity (ρV value) is the same in both media ($\rho_2 V_2 = \rho_1 V_1$), equations (10) and (11) give p_r equal to zero and p_t equal to p_i; in words, all the sound energy is transmitted across the boundary, and none is reflected. This would be realized if the two media were precisely the same: the interface would then of course disappear. It may well be possible, however, to find two different media for which the ρV values are the same and hence secure complete transmission without reflection. Thus there is a certain kind of soft rubber whose ρV value is about the same as that of water. A water tank lined with this material would appear to have no reflecting walls, and this is helpful in the calibration of certain acoustical instruments.

Transmission of sound without reflection

The greater the difference between the ρV values of the two media, the greater is the percentage of reflection at the interface. The actual power-transmission ratio (P_r), which is the ratio of the intensity of the transmitted sound wave to that of the incident sound wave, also depends only on the factors mentioned above. Calculation shows that the power-transmission ratio is given by equation (12):

$$P_r = \frac{4\rho_2 V_2/\rho_1 V_1}{(1 + \rho_2 V_2/\rho_1 V_1)^2}. \tag{12}$$

Again from equation (12) it is seen that when the ρV value is the same in both media, the power-transmission ratio is equal to one, meaning that 100 percent of all the sound energy or power is transmitted. The specific acoustic resistance ρV is clearly an important criterion for the transmission of plane sound radiation across boundaries.

Formula (12) together with the corresponding formulas for the case of oblique incidence have had important application to acoustic transmission in the atmosphere between layers of air at different temperature and humidity and to transmission in seawater with temperature gradients.

Acoustic impedance. The significance of the product ρV in sound transmission prompts further attention to it. Referring to this quantity as the specific acoustic resistance of a plane harmonic wave inevitably suggests the analogy of electrical impedance in alternating electric current, defined as the ratio of the alternating electromotive force to the alternating current (see ELECTRICITY). It has therefore been thought helpful to define the acoustic impedance as the ratio of the excess pressure to the rate of volume flow. The volume flow is obtained from the flow velocity ($\partial\xi/\partial t$) by multiplying by the area (S) of the wave front. Hence the impedance (Z) has a value equal to the excess pressure divided by the product of the area and the flow velocity, or $Z = p_e/(S\partial\xi/\partial t)$. In words, the acoustic analogy of electromotive force is taken to be the excess pressure, and the acoustic analogy of the electric current is the volume flow rate.

In the mathematical theory of alternating currents the impedance is a complex quantity (so called because it involves the square root of minus one, an imaginary number) of which the real part is the resistance of the circuit and the imaginary part (having a factor $\sqrt{-1}$) is the reactance. In the acoustic wave case the quantity Z for a plane wave is real and equal to the density times velocity divided by wave front area ($\rho V/S$). Hence it is taken to correspond to an electrical resistance. The acoustic resistance for unit area of wave fronts (not per unit area) is the product ρV and this is justifiably called the specific acoustic resistance. Analysis shows that the acoustic impedance for a spherical wave has both real and imaginary components, corresponding to resistance and reactance, respectively. The physical meaning of this is connected with the spreading of the wave in all directions. The reactance of a diverging spherical wave turns out to be positive and behaves like an electrical inductance (*i.e.*, an inertia factor), whereas the reactance of a spherical wave converging to a point is negative and thus behaves like an electrical capacitance (*i.e.*, the reciprocal of a stiffness).

Specific acoustic resistance

The impedance notation has the advantage of permitting acoustic transmission calculations to be replaced by analogous electrical computations. Many acoustical effects can be readily estimated in terms of known values of acoustic impedance at strategic points. For example, the effectiveness of a horn (*i.e.*, a tube of variable cross section) as both a receiver and a transmitter of sound waves is directly proportional (for a given volume current) to the value of the acoustic resistance at the horn's throat. The effective acoustic transmission through a tube in which a branch line is inserted at some point can be shown to depend in a relatively simple fashion on the acoustic impedance at the branch point. Problems in room acoustics involving the reflection of sound from the walls of a room can often be simplified by the use of the impedance notation.

Selective transmission: filtration. An important case of the transmission of acoustic energy arises when a finite layer of different acoustic material (medium II) is inserted in a medium (medium I) in which the propagation takes place. A simple example is illustrated in Figure 4 (top). To determine the influence of the layer of medium II on the transmission, a mathematical analysis can be made of the transmission of a plane harmonic sound wave as it passes from medium I through medium II and back into medium I. This yields a power transmission ratio (P_r)—*i.e.*, the ratio of the average power emerging from medium II to that entering it—which is found to depend on the frequency of the sound and the thickness of medium II, as well as the ρV value for each layer. The plot of P_r as a function of frequency has the general form shown in Figure 4 (bottom). The transmission is said to be selective with respect to frequency and passes through a succession of maxima and minima as the frequency increases.

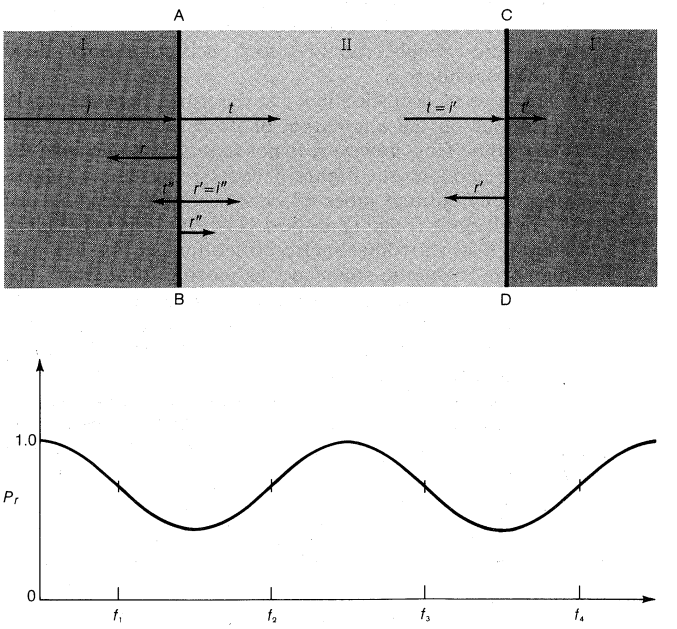


Figure 4: (Top) A layer of medium II situated between layers of medium I (see text). Waves (*i, i', i''*) incident on both faces AB and CD are reflected (*r, r', r''*) and transmitted (*t, t', t''*). (Bottom) Plot of power-transmission ratio (P_T) as a function of frequency (*f*) transmitted through medium I as above. (P_T=1 represents total transmission.)

This suggests the possibility of arranging a collection of

media that will act as an acoustic filter—*i.e.*, transmit waves of certain frequencies well, but fail entirely to pass waves of other frequencies. Further theoretical study indicates that this result can be realized approximately by a structure consisting of a sequence of alternating layers of media I and II. As the number of layers increases the transmission ratio approaches more nearly the form shown in Figure 5. Here the power-transmission ratio is

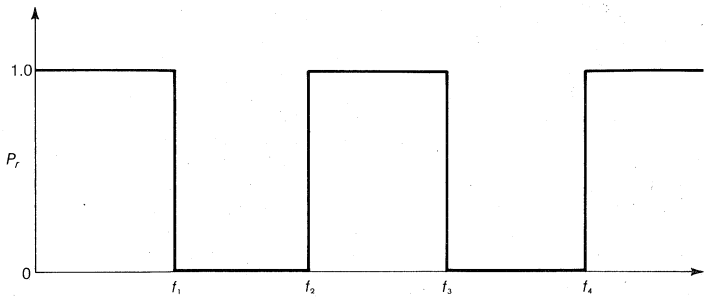


Figure 5: Power-transmission ratios (P_T) according to frequencies (f). Frequencies lying between 0 and f_1, f_2 and f_3, etc., called pass bands, are totally transmitted. Frequencies between f_1 and f_2, f_3 and f_4, etc., called attenuation bands, are totally absorbed.

equal to one for frequencies lying between zero and f_1, equals zero for frequencies between f_1 and f_2, is unity again for frequencies between f_2 and f_3, etc. A frequency band such as f_2f_3, in which the power-transmission ratio is one, is called a pass band, whereas a band such as f_1f_2, in which the power-transmission ratio is equal to zero, is called an attenuation band. The structure of alternating layers thus acts as an acoustic filter with a series of pass and attenuation bands. Because the first pass band extends from zero to f_1, it is commonly called a low-pass filter.

Attenuation band

The type of acoustic filter just described depends for its behaviour on the alternation of properties of two different media. This, however, is not necessary to achieve the same general result. Figure 6 (top) represents a succession of cylindrical tubes of cross-sectional areas S_1 and S_2 and lengths l_1 and l_2, respectively. Analysis and experimental tests show that such a structure acts as a low-pass filter for harmonic sound waves passing through it. The transmission pattern of pass and attenuation is indeed very similar to that shown in Figure 5. Here the ratio S_1/S_2 of the cross-sectional areas of the constricted and expanded tubes plays the same role as the ratio of the ρV values of the specific acoustic resistances in the iterated-media filter. The lengths l_1 and l_2 of the section also have a decided influence on the frequency range of the pass and attenuation bands.

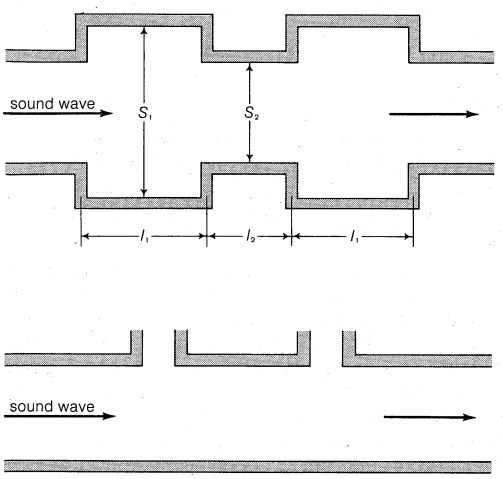


Figure 6: (Top) Low-pass acoustic filter, consisting of a succession of cylindrical tubes, cross-sectional areas S_1 and S_2 and lengths l_1 and l_2, for harmonic sound waves (see text). (Bottom) High-pass acoustic filter through which only high-frequency waves pass.

A tube with equally spaced side orifices, such as are shown schematically in Figure 6 (bottom), also acts as an acoustic filter, but in this case the lowest frequencies are attenuated, and hence it is called a high-pass filter. Many other types of sound filters have been constructed to suppress or silence certain frequency bands in blowers, mufflers, and other equipment producing noise by gas flow. Filters consisting of solid sections have also been constructed. Many musical instruments involve acoustical filtration in their behaviour.

Sound absorption and dispersion in fluids. In the previous discussion of sound filtration no account was taken of the dissipative absorption of sound radiation as it passes through any material medium and its transformation into heat. This leads to a loss in intensity with distance traversed that has nothing essentially to do with the selective transmission just treated. Also, as has already been mentioned, a decrease in intensity of a diverging spherical wave is caused by geometrical spreading. Superposed on this decrease is the absorption due to the interaction between the sound wave and the medium. The properties of the medium responsible for this absorption are shear viscosity, heat conduction and radiation, diffusion, and molecular energy exchange. Shear viscosity is a kind of internal fluid friction that manifests itself in the tendency of parallel moving layers of fluid to have their relative velocity reduced. The coefficient of shear viscosity (usually denoted by η) is a measure of this tendency (see DEFORMATION AND FLOW).

To remain correct at all frequencies, the formula for the intensity of a plane harmonic acoustic wave in terms of the distance traversed must take account of the fact that, in propagation through a viscous medium, the sound velocity changes with frequency; *i.e.*, there is dispersion. Unlike the situation in light in which, except for anomalous dispersion, the velocity decreases as the frequency increases, the effect of the viscosity in a medium is to increase the sound velocity with increase in frequency. The effect is a small one, however, even in viscous media like the liquid, glycerin.

Effect of viscosity on dispersion

The medium also produces sound absorption by conducting away heat produced in the compressional disturbance involved in the propagation of sound. For monatomic gases (*e.g.*, helium and argon) the contribution of the heat conduction to the absorption coefficient is about 70 percent of that due to viscosity. For polyatomic gases (*e.g.*, hydrogen sulfide and carbon dioxide) the influence of heat conduction is much less, and in liquids it is generally negligible. The effect of diffusion is also negligible save in mixtures of gases with considerably different molecular weights; *e.g.*, helium and krypton.

Experimentally determined values of the absorption of sound in fluids, especially polyatomic gases, are in general far in excess of those calculated above on the basis of viscosity and heat conduction. The explanation for this disparity has been sought successfully in a closer look at the energy exchanges among the molecules composing the fluid. After all, the propagation of sound is really a molecular phenomenon, which is somehow masked in the usual macroscopic theory of disturbances in elastic media. When a fluid is compressed during the passage of a sound wave, the average translational velocity of the molecules is increased, and it is the transfer of the associated momentum and kinetic energy by collision to the nearby molecules that provides for the transmission of the wave. But this is not all that happens. In a gas made up of polyatomic molecules, the molecules have internal energy of rotation and vibration as well as translational energy of their centres of mass. In the collision of the molecules, this internal energy is continually being exchanged with the external translational energy. When the translational energy is increased by the passage of a sound wave, some of the increase goes into the internal energy states. There is, however, a lag in its return to the translational form. Hence changes in the translational energy fail to keep pace with the propagation of the wave, and some of the original energy of the disturbance is rendered unavailable for transmission during each frequency cycle. It is therefore lost in dissipation and the

result is absorption. This type of mechanism, called relaxation absorption, has proved adequate to account for the experimentally determined values. It has led to the establishment of a 20th-century branch of the theory of sound called molecular acoustics, by which sound measurements are able to give insight into the interaction of molecules in fluids.

Molecular acoustics

Diffraction and scattering. When a sound wave strikes an obstacle of finite size, some of the sound energy is reflected, some is transmitted into the obstacle, and some gets around the obstacle and is said to be diffracted. It is customary to say that the original beam of sound is scattered by the obstacle. The scattering pattern depends on the relation between the wavelength of the primary incident wave and the dimensions of the obstacle. In the case, for example, of a rigid sphere of radius a as the obstacle, if the wavelength (λ) of the incident radiation is small compared with its circumference ($2\pi a$) the scattered wave on the side of the sphere opposite to the oncoming radiation combines with the latter to produce zero intensity by destructive interference (so called because, being out of phase with one another by 180°, their displacements are annulled). Hence the spherical obstacle forms a shadow that is sharper the smaller the ratio of wavelength to circumference ($\lambda/2\pi a$). This is analogous to an optical shadow in the case of light and justifies the use of sound rays (geometrical acoustics) in place of the propagation of wave fronts. For audible waves (λ in air greater than about two centimetres or 0.8 inch), however, this type of shadow-forming scattering is important only for relatively large obstacles.

At the opposite extreme, if the wavelength is much greater than the circumference ($\lambda >> 2\pi a$), the primary wave is scattered in all directions and there is no distinct shadow. The intensity of the scattered wave is directly proportional to the square of the volume of the scattering particle (in this case a sphere) and inversely proportional to the fourth power of the wavelength. This is the Rayleigh scattering law in acoustics; its optical analogue helps account for the blue colour of the sky (see LIGHT). In the scattered radiation the higher frequencies will produce the high intensities. Sound waves scattered by a grove of trees appear to be raised in pitch. A similar phenomenon occurs when underwater sound waves are scattered by air bubbles.

An interesting case of scattering is that in which light is scattered by sound. Optical diffraction by means of a diffraction grating is well-known. A sound wave can serve as a kind of diffraction grating in a material medium (a liquid as well as a gas) and hence diffract light. Study of the light-diffraction spectrum can actually provide a measure of the intensity of the sound.

The fluctuations in density in any material medium (at any temperature above absolute zero) due to molecular motion and ionic vibrations may be considered a form of sound radiation at extremely high frequencies—*i.e.*, about 10^{12} hertz and higher. This is sometimes called the pretersonic, or ultrahigh, ultrasonic range. Such sound waves also scatter light and the result is the so-called Brillouin scattering, which in the 1960s and 1970s showed great promise in the study of material structure, particularly of liquid structure. It has proved advantageous in the study of crystal lattices (a crystal is said to have a lattice structure because its atoms are arranged in an orderly array) to associate a sound "particle" with the ultrahigh-frequency waves travelling back and forth in the solid as a result of the lattice vibration. This hypothetical particle is called the phonon. It plays the same role with respect to ultrahigh-frequency acoustic radiation that the photon, or light particle, plays with respect to electromagnetic radiation.

The phonon

Another example of the interaction of light and sound is the use of laser light (a narrow beam of coherent light) to stimulate the production of high-frequency sound waves in a crystal. A similar technique can also serve to amplify existing ultrasonic waves.

Waves in bounded media: standing waves in a tube. When sound waves travel in a bounded medium with reflecting surfaces, the superposition of waves travelling in many directions produces important and interesting effects. The simplest case is that of a cylindrical tube of length l and cross-sectional area S, closed rigidly at both ends. A disturbance produced anywhere in the air in the tube will be reflected from both ends and produce in general a series of waves travelling in both directions along the tube. From the geometry of the situation and the constant value of sound velocity, these waves must be periodic with frequencies determined by the boundary conditions at the ends of the tube; *i.e.*, zero resultant air displacement. The allowed frequencies turn out to be a multiple of the sound velocity in the air in the tube divided by twice the tube length, or

$$f_n = nV/2l, \tag{13}$$

in which V is the sound velocity and n is any integer. These frequencies f_n in which $n = 1, 2, 3, \ldots$ are called the characteristic frequencies or normal modes of vibration of the air column. The fundamental frequency, or first harmonic, is described by an integer value of one and is $f_1 = V/2l$ with the associated wavelength $\lambda_1 = 2l$. The frequency f_2 is called the second harmonic and so on for $n = 3, 4, 5, \ldots$. The harmonics beyond the fundamental are also often referred to as overtones.

Approximately the same set of normal modes holds for a cylindrical tube open at both ends, though the boundary conditions are different. It is assumed that the excess pressure p_e is zero at an open end, since it is open to the surrounding air. The justification of this assumption cannot be easily explained, but it can be confirmed experimentally by placing a probe microphone at the open end. Actually, it turns out that the point of zero excess pressure is never precisely at the open end but somewhat beyond it, producing an effect called an end correction to the normal mode frequencies. Application of the ideal boundary conditions just mentioned to a tube closed at one end and open at the other (the ideal case of a so-called closed organ pipe) leads to an equation for the normal mode frequencies that is similar to equation (13), but the whole integer is replaced by a half integer value:

End correction for open pipes

$$f_n = (n - \tfrac{1}{2})\, V/2l, \tag{14}$$

in which n equal to one corresponds to the fundamental, and n equal to two to the second harmonic, etc. Because of the open-end effect, however, the length ($2l$) in equation (14) must be modified by adding a quantity that depends on the tube diameter. Thus, the actual frequencies are

$$f'_n = (n - \tfrac{1}{2})\, V/2\,(l + S/c),$$

in which S is the area of cross section of the tube and c is a quantity with the dimensions of length of about the magnitude of the diameter of the tube. It is called the acoustic conductivity of the open end of the tube. The acoustic conductivity is difficult to calculate theoretically and is often evaluated by the actual measurement of the frequency f'_n.

In an ideal sound-wave disturbance corresponding to a given normal mode in a tube open at both ends, there are

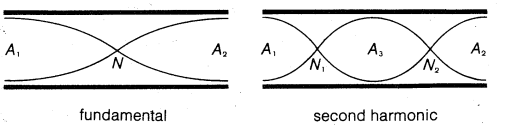


Figure 7: Nodes *(N)* and antinodes *(A)* in an open-ended tube (see text).

positions in the tube, namely, those at distances (x) from one open end that satisfy the equation, $\cos n\pi x/l = 0$, at which the displacement is zero at all times. Such a situation cannot exist in a progressive wave, as shown by equation (1). Hence the resultant wave disturbance corresponding to a normal mode is appropriately termed a standing or stationary wave. The positions of permanent zero displacement are known as nodes, whereas the intermediate positions for which $\cos n\pi x/l$ is equal to one instead of zero are known as antinodes or loops and correspond to maximum displacement. At a node the excess

pressure is a maximum, whereas at an antinode it is zero. The distance between successive nodes is equal to a half wavelength of the particular mode. As an example, the fundamental in an open-ended tube is shown schematically in Figure 7 (left) by means of portions of sine waves (the maxima of a wave indicate the positions along the tube where the displacement is greatest). Displacement antinodes are indicated by A_1 and A_2, whereas N (at centre) is a node. In Figure 7 (right) the second harmonic is indicated with antinodes at A_1, A_2, and A_3, and nodes at N_1 and N_2. The latter are, respectively, ¼ and ¾ of the length of the tube from the open end at the left. In Figure 8 the corresponding situation is shown for a tube closed at one end and open at the other. In Figure 8 (left) the closed end is node N and the open end is an antinode A. The length of the tube is $\lambda_1/4$. In Figure 8 (right) the closed end is still a node N_1 and the open end an antinode A_2, but there now is another antinode A_1 and another node N_2. For the nodes in the tube closed at one end (at $x = 0$), the analytical condition that exists is $\sin(n - \frac{1}{2})(\pi x/l) = 0$, in which n equal to one corresponds to the fundamental, etc.

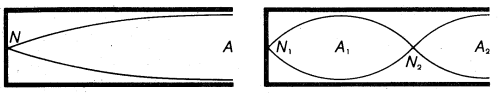


Figure 8: Nodes *(N)* and antinodes *(A)* in a tube closed at one end (see text).

Acoustical properties of a room

The preceding analysis can be applied to any confined space as, for example, a room with or without open windows. The reflections of a sound disturbance from the boundaries of the room produce a three-dimensional standing wave pattern that can be analyzed with a set of normal nodes, depending on the shape and dimensions of the room. Some of these nodes may be much more prominent and correspond to more energy flow than others and thus have a good deal to do with the acoustical properties of the room. Here absorption at the walls and in the air of the room plays an important role in diminishing the influence of these normal nodes.

Radiation pressure. For a plane harmonic progressive wave the excess pressure produced by the disturbance as it moves through a fluid is a sinusoidally varying function of time and space and hence vanishes if averaged over the time at any particular place. Associated with a progressive sound wave in any physically realizable fluid medium, however, there is also an excess pressure that does not become zero on the average, thus making a net addition to the static pressure prevailing in the fluid. This is known as the radiation pressure. It arises because the simple linear expressions for displacement and excess pressure, like that in equation (1), hold only for sound of relatively small intensity. For sounds of greater intensity, sine-squared terms and the like enter into the expression for the excess pressure, and it is these terms that, when averaged over the time, give a nonvanishing result, corresponding to the steady radiation pressure. The subject is admittedly a difficult one, and it turns out that there are at least two different kinds of radiation pressure, depending on definition, theoretical calculation, and experimental measurement, but the most practical from an experimental point of view is expressed simply as the ratio of the intensity (I) of the sound radiation to its velocity (V) and has the value $P_{rad} = I/V$. To get an idea of the order of magnitude of the radiation pressure, it is interesting to consider sound in air of intensity 10^{-2} watt per square centimetre, which is about 140 decibels above the minimum audible intensity. This intensity is at about the threshold of pain in the human ear and is equivalent to 10^5 ergs per square centimetre-second. In this case the radiation pressure comes out about three dynes per square centimetre.

Nonlinear acoustics

Radiation pressure is an illustration of what has come to be called nonlinear acoustics, resulting in effects from such large sound intensities that the ordinary linear wave equation no longer applies. Intensities of interest here are about 150 decibels above minimum audible or higher (greater than absolute intensity 1/10 watt per square centimetre). Such sound waves are now referred to as macrosonic. They have become of increasing importance in connection with practical applications of acoustics in industrial processing, as in metallurgy, in underwater sound for communication and detection, and in the biomedical field for the study of individual living cells.

Shock waves. In a macrosonic or high-intensity sound wave, (*e.g.*, that due to an explosion) a disturbance that starts out with a symmetrical profile—*i.e.*, a sinusoidal curve given by equation (1)—will soon develop a steepening of the wave front, which will then take the form of a sawtooth curve if the excess pressure at a given instant is plotted as a function of distance in the direction of the wave propagation. At successive points in this wave there are, ideally, discontinuities in pressure, density, and flow velocity. Empirically, these ideal discontinuities become large changes in these quantities over a very small space interval—*e.g.*, about 10^{-4} centimetre (4×10^{-5} inch). This transition interval travels through the medium faster than the normal acoustic velocity. Called a shock wave, it can result not only from an explosion but is also emitted from an object like a missile or jet airplane moving faster than the velocity of sound. In the case of an airplane breaking the sound barrier, as the phrase goes, it results in the sonic bang or boom on the ground. Such sonic disturbances have produced a serious problem in connection with the flying of supersonic airplanes.

III. Sound production and perception

SOURCES OF SOUND

Types of sources. Any local change in the density of an elastic medium can serve as a source of sound. This accounts for the great variety of acoustic sources because density changes may be produced in a great many ways, including mechanical, thermal, electrical, magnetic, and chemical actions. The most common sound waves are produced by the mechanical vibrations of solids, liquids, and gases. Solid vibrators include strings and rods, membranes and plates, shells (*e.g.*, bells), as well as three-dimensional extended objects like the Earth itself. Liquid sources are not as common, but the turbulent flow of water or air provides an example. Gaseous sources include organ pipes, whistles, singing flames, and explosions, as well as turbulent airflow.

Methods of producing local density changes in elastic media vary greatly. One of the simplest is the impact of one solid on another, such as the impact of a pile driver on a pile, the blow of a piano hammer on a piano string, a drumstick on a drum, a clapper on a bell, as well as pounding on a floor or wall. Still another common method is exhibited by the friction associated with the rubbing of two solids together, as in the bowing of a violin string and in the motion of air and other gases past the surfaces of solids, as in the sound of the wind and in aerodynamic noise.

More sophisticated sound sources became technically feasible when mechanical vibrations were made possible by means of magnetic and electrical effects. The attraction of a piece of iron or some magnetic alloy to the pole of an electromagnet is of course the basis for the electric bell and the telephone receiver. The motion of a wire carrying an alternating current in a magnetic field is the basis for the so-called electrodynamic sound source used in radio and television loudspeakers. They are appropriately called electroacoustic devices because they transform electrical oscillations into mechanical vibrations, usually of rods, membranes, or diaphragms.

Electroacoustic devices

Still more sophisticated in many respects are the magnetostrictive and piezoelectric sound sources. The former utilize the change in dimensions produced in a magnetic solid, such as nickel, by a changing magnetic field. The latter exploit the fact that many crystals, such as quartz, change their size and shape when placed in a changing electric field. These sources have received wide use in the production of underwater sound radiation and in medical diagnosis and therapy, particularly at ultrasonic frequencies. It is customary to call such sound sources trans-

ducers because they involve the transformation of one type of vibratory energy into another. These devices are usually suitable not only for the production of sound but also for its reception as well.

Explosions caused by the rapid combustion of organic materials produce another source of sound usually in the form of noise, as exemplified by the internal-combustion engine.

Mechanical vibration and radiation efficiency. The properties of a sound source depend essentially on the nature of its elastic vibrations. The analytical details will be found in the text article VIBRATIONS. Here the discussions will be confined to general characteristics having to do with the ability of an elastic vibrator to radiate sound energy into the surrounding medium.

When a mechanical system vibrates in a vacuum, it loses energy at a rate equal to a damping factor times the average of the square of the displacement velocity of the vibrator—*i.e.*, $R\overline{\dot{\xi}^2}$ in which R is the damping factor and $\dot{\xi}$ is the displacement velocity. The origin of the damping may be found in the internal frictional forces inherent in the vibrator and its mounting. If now the vibrator is allowed to vibrate in a material medium like a gas or a liquid, it will be found to lose energy at a greater rate because of the production of mechanical radiation, which is transmitted through the medium. The average rate of radiation of energy by the vibrator as a source of sound can be written in the form $R_r\overline{\dot{\xi}^2}$, in which R_r now represents the damping due to the radiation. In order to maintain this radiation energy at a constant average rate, it is necessary to continue to supply energy to the vibrator at the same average rate by whatever is driving it. The radiation damping factor R_r that controls the rate of sound radiation by the source is found to depend on the size and shape of the source, as well as the density of the medium into which the radiation takes place and the velocity of sound in this medium. It also depends on the frequency of the radiation.

Radiation damping factor

The analysis of a pulsating, spherical shell, for example, indicates that its radiation efficiency for a given displacement velocity is much greater at higher frequencies than at low and this is one reason for the technological interest in ultrasonics.

A pulsating sphere is not usually a practical sound source. Similar considerations, however, apply to the membrane or diaphragm source in loudspeakers or sonar sources for underwater sound radiation. The mathematical expressions differ in detail, but in every case, other things being equal, the radiation efficiency increases with the frequency. At all frequencies the radiation efficiency depends directly on the density of the surrounding fluid medium. This dependence makes it difficult to get radiation from any source into a medium having a density that is small compared with that of the solid vibrator itself. Such a low efficiency accounts for the often-cited experiment of the difficulty of hearing a bell in an evacuated jar, commonly explained by saying that "sound is not transmitted through a vacuum." Actually, sound waves will still travel short distances without abnormal dissipation through gases at the low pressure prevailing in such an experiment; the difficulty is to get the wave into the gas from the source in the first place.

Sound beams. In view of the relatively long wavelength of audible sound, it is obviously difficult to produce beams of such sound without the use of extremely large acoustical mirrors and lenses. Nevertheless, at ultrasonic frequencies the sound wavelength becomes small enough to make sound beam formation practical, although never with the sharpness of light beams, even at the highest ultrasonic frequencies (*e.g.*, 10^3 megahertz). Such beams are basic to the operation of sonar, the underwater echo ranging system (see SONAR).

Practical simple sources: voice and siren. At the beginning of the section on sound sources various types of sound sources were mentioned; here are presented some specific examples. First to be considered as a source is the human voice (see SPEECH, PHYSIOLOGY OF, and PHONETICS). It operates by forcing air from the trachea to vibrate the vocal cords. This in turn sets into vibration the air in the cavities of the throat and mouth; the resulting disturbance emerges from the lips. The part the vocal cords play in singing and speech has been established by motion pictures of their vibration.

Studies of the acoustical power of conversational speech have shown that in the case of an average speaker the overall level directly in front of the mouth can vary from 81 decibels at 15 centimetres (six inches) from the mouth to 65 decibels at 100 centimetres (40 inches). These decibel levels are relative to a standard reference level of 10^{-16} watt per square centimetre (the so-called minimum audible threshold). The mouth is only an approximate point source of sound emitting spherical waves, for at 100 centimetres directly back of the head, in the case just mentioned, the level is 62 decibels. This effect is due to diffraction by the head and is present to some extent in all practical sound sources of appreciable size.

The power in speech sound waves varies with the sound emitted, being much larger for vowels than for consonants. Ordinary speech involves the superposition of many waves of different frequencies. Moreover, the recognition of speech as mediated by the ear is not as simple as the detection of ordinary sound sources by appropriate electromechanical receivers. In the 1970s researches were directed toward improved methods of speech analysis and machine synthesis of speech. Much attention was also paid to measures of speech perception and intelligibility, as well as to better systems for the transmission of speech. Efforts were made to improve machine recognition of speech—*e.g.*, the phonetic typewriter, which types out spoken dictation. The coupling of sound recognition to computer technology was being developed.

The siren is one of the simplest sources of intense sound. In principle it consists of a disk containing a number of small holes placed at regular intervals around the same periphery. If this disk is rotated past an orifice lined up with the holes and compressed air is driven through the orifice, the result will be a succession of puffs of air. If the number of puffs per second (governed by the speed of rotation of the disk) is sufficiently large the result will be a periodic sound wave of frequency equal to the number of puffs per second.

Principle of the siren

Jet-edge tones and organ pipes. Another standard sound source operated by blown air is the jet-edge system coupled to an air column, as in the organ pipe. In the familiar type of pipe a jet of air impinges on the narrow edge of the bottom of the pipe. The resulting vibrations are coupled to those of the air in the pipe. The natural frequency of the edge tone coupled to the pipe increases approximately linearly with the velocity of airflow, but the coupling forces the fundamental tone of the pipe on the system for small velocities. When the flow velocity reaches the value for an isolated edge tone of frequency equal to the second harmonic of the pipe, the tone emitted rises to this harmonic, and the pipe is said to be overblown. The production of edge tones is associated with turbulence (vortex motion) in the airflow in the vicinity of the edge.

The large noise output of the jet-engine airplane in the 1960s and 1970s stimulated considerable interest in the theory of the generation of sound by airflow. The theory is intricate but leads to the result that the sound-power output of a jet of air is directly proportional to the fifth power of the mach number of the flow. The mach number is the ratio of the air flow velocity (relative to the stationary medium) to the velocity of sound in the stationary medium. Flow with mach number greater than one is known as supersonic.

Electroacoustic sources. The most important sound sources in technological use employ electrical energy. These may be divided into two main classes: irreversible and reversible. In an irreversible source the energy for sound radiation is supplied by a flow of fluid (air or water) under pressure and the electrical energy is used to modulate or control. A variable magnetic field, for example, can produce vibrations in a reed or diaphragm that will control the flow of compressed air through a variable orifice. Such a device can produce extremely intense sound waves with relatively high efficiency.

A reversible electroacoustic source transforms electrical energy directly into mechanical energy and thence into acoustical radiation. As the name implies, such a source can function in reverse to transform acoustical into electrical energy via some mechanical structure. It can therefore function equally well as a receiver of sound. As can be shown, under certain specific conditions, which hold approximately in practice, a good reversible source of sound will act also as a good receiver.

Transducers

Any system that transforms electrical to acoustical energy or vice versa is called an electroacoustical transducer. The important practical characteristics of a transducer used as a source are its acoustical output, its efficiency (ratio of the radiated power to the input power), its directivity, and its departure from linearity. The nonlinearity means the extent to which the mechanical force producing the vibrations leading to radiation is not strictly proportional to the electrical field or voltage applied to the input. Such nonlinearity leads to distortion in the radiated output. The directivity refers to the variation in intensity of the radiation with direction from the source.

To come to detailed examples, in the electromagnetic transducer, an alternating current of the desired frequency is passed through a coil that activates an electromagnet to produce an alternating force on a diaphragm of iron or some magnetic alloy. This is the type that has been employed in the conventional telephone receiver. The electrodynamic transducer is now more common. In this transducer the oscillating current passes through a coil suspended in a permanent magnetic field and undergoes the usual force. The coil, called the voice coil, is fastened mechanically to the sound-radiating surface. In the well-known loudspeaker, the latter is normally a cone attached to the voice coil at or near the vertex. The cone is the acoustical element, radiating directly into the surrounding medium. Various practical devices have to be incorporated in such a source to ensure large output and to diminish the distortion due to the resonant or normal mode frequencies of the cone.

The efficiency of an electrodynamic loudspeaker tends to vary directly with its linear dimensions. The speakers in common use in radio and television sets have efficiencies varying from 2 percent to 5 percent. By the use of an attached flaring horn this figure can be increased to about 50 percent. All these figures depend on the frequency range over which the speaker is driven. For a typical small-cone receiver ten centimetres in radius, the efficiency is 2 percent only over the frequency range from 75 hertz to 1,000 hertz; at 10,000 hertz the efficiency drops practically to zero. The total acoustical power output of an electrodynamic loudspeaker can have maximum values ranging from 0.5 watt to 150 watts, depending on size and the associated electric circuit. Such a speaker at maximum output risks nonlinear behaviour—*i.e.*, large diaphragm motion that is no longer directly proportional to the applied electrical voltage.

For ultrasonic radiation, the most important electroacoustic sources employ the piezoelectric and magnetostrictive effects already mentioned earlier in this section. A plate of quartz cut with its faces parallel to the optic axis of the crystal (a so-called *X*-cut), and used as the dielectric in a capacitor subjected to an alternating electric field, becomes an ultrasonic source with the frequency of the oscillatory field. It will, of course, vibrate with maximum intensity and radiate correspondingly if this frequency matches one of its own natural frequencies of vibration. The various resonance frequencies are inversely proportional to the thickness of the quartz crystal. Higher harmonics of the fundamental frequency of the quartz can be used to provide higher frequencies of radiation without using crystals too thin to be practical.

Quartz crystal as an ultrasonic source

In practical use, a number of *X*-cut crystals are usually arranged in the form of a mosaic to provide a larger radiating surface. Many types of piezoelectric materials are now available in addition to quartz; these include Rochelle salt, barium titanate, and lead zirconate titanate. With such crystals, radiation of ultrahigh frequency has been achieved. In the early 1970s it proved possible to reach frequencies of the order of 10^{10} hertz.

Magnetostrictive oscillators have also proved popular as ultrasonic radiators. If an alternating current passes through a coil surrounding a rod of magnetizable material already magnetized in the direction of its length, the rod will vibrate lengthwise with the frequency of the current. If this frequency coincides with one of the natural frequencies of the rod, large amplitudes can be obtained. A nickel rod ten centimetres (four inches) long, for example, will radiate strongly at its fundamental frequency of about 24 kilohertz. Simple magnetostrictive sources are less efficient than piezoelectric devices, largely because of the loss of energy through the production of eddy currents. Magnetic materials, the so-called ferrites, are also in use in cutting down heating loss in magnetostrictive transducers. These ferrites are ceramic-type materials, such as ferric oxide mixed with nickel oxide and compressed so as to form solid rods.

Magnetostrictive transducers are often used to provide underwater acoustic beams, usually beyond the audible frequency range.

Doppler effect. In the discussion of sound sources it has been tacitly assumed that the source is at rest relative to the medium. Moving sources—*e.g.*, airplanes, submarines, and missiles of various kinds—however, are important in applications. A significant effect of such motion on acoustic reception is the perceived change in pitch in accordance with the Doppler principle. If a harmonic source of frequency f moves with linear velocity v_s with respect to a stationary medium and a receiver is moving with velocity v_r along the same line, the receiver will estimate the emitted frequency as a frequency f' that is given by

$$f' = f(V - v_r)/(V - v_s),$$

in which V is the velocity of sound in the stationary medium. In the formula v_r and v_s are reckoned as positive in the direction from the source to the receiver and negative otherwise. If, for example, the receiver is at rest and the source in motion, the apparent frequency at the receiver is lower than the emitted frequency if the source is receding and is higher than the emitted frequency if the source is approaching. Knowledge of the emitted frequency and measurement of the apparent frequency thus permit the determination of the velocity of the source, a matter of significance in military applications such as sonar.

THE RECEPTION OF SOUND

To be studied and applied, sound energy must be detected —*i.e.*, transformed to make it perceptible to people or sensible to cybernetic (communication and control) devices. This is the function of a sound receiver. Any reversible sound transducer can be used as a sound receiver.

The ear and hearing. The ear is the most important acoustical receiver in man's environment. Normal people can detect sound intensity as low as 10^{-16} watt per square centimetre and can stand intensities up to about 10^{-4} watt per square centimetre before pain ensues. These are figures averaged over representative samples of the population.

For the description of the structural details of the ear as a receiver of sound see the text articles EAR AND HEARING, HUMAN and SOUND RECEPTION. During the early 1970s theories of hearing remained in a state of flux, with no one point of view commanding universal assent.

Thresholds of feeling and hearing

The threshold values of hearing and feeling as a function of frequency are based on the work of a physicist in the United States, Harvey Fletcher, and confirmed by other investigators (Figure 9). The lower two curves relate minimum audible intensity to frequency. These curves rise at both ends of the audible range and have a minimum at about three kilohertz. The upper curve is the corresponding plot of the threshold of feeling. Intensities greater than those along this curve produce pain. Even so-called normal people show variation in thresholds and the curves represent averages over suitably large samples. Advance in age results in general in hearing loss, appearing in Figure 9 as a rise in the lower curves. The tendency is for the upper-frequency limit to decrease with age.

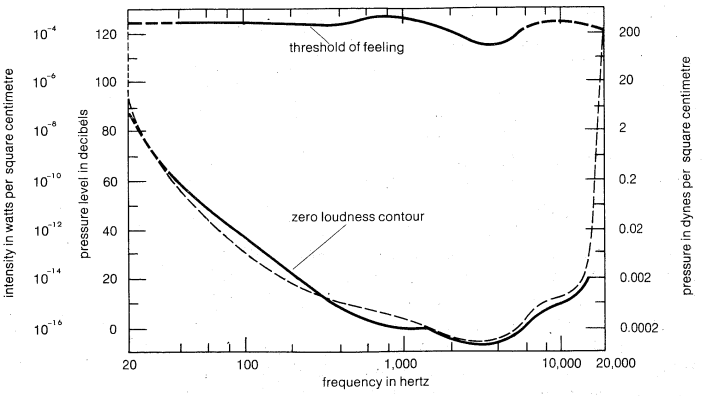


Figure 9: Auditory area between threshold values of feeling and hearing (see text).
From *Speech and Hearing in Communications* by Harvey Fletcher (©1953); D. Van Nostrand Company, Inc.

Loudness is a sensation related to the intensity of the received sound but not directly proportional to it. Intensity is an objective physical quantity, whereas loudness is subjective and defined in terms of the so-called average listener. It is customary to distinguish between loudness level and true loudness. The former is measured by varying the intensity of a reference tone until it appears to a listener to be as loud as the sound stimulus under test. The number of decibels by which the reference tone has been raised above the minimum audible threshold is said to be the loudness level of the tone under test in phon units. This applies, strictly speaking, only if the reference tone has a frequency of one kilohertz and is placed at a distance of one metre from the head of the listener.

Measurement of true loudness

The measurement of true loudness demands the introduction of a scale such that when loudness of a certain number of units on this scale is multiplied by any arbitrary factor the magnitude of the new auditory sensation then produced will be heard by a listener as equivalent to the corresponding number of units times that of the original sensation. The unit of loudness introduced for this purpose is called the sone. It is defined as the loudness produced by a tone of one kilohertz at 40 decibels above the minimum audible threshold. By a statistical study of experiments in hearing, a loudness scale in sones has been established in which there is a logarithmic relation between loudness (L) in sones and loudness level (P) in phons, $\log_{10} L = 0.03\ P - 1.2$, that holds for a reference tone of one kilohertz. This means that subjective loudness is equal to ten raised to the power $.03P - 1.2$. The loudness level is of course a measure of acoustic intensity on the decibel scale.

Hearing with two ears, the so-called binaural effect, leads to the ability to detect the direction of the source of sound radiation. A source of sound on the right of the plane bisecting the line joining the ears and perpendicular to it will be recognized as on the right. This effect is not due primarily to difference in intensity at the two ears but rather to difference in phase or time of arrival at the two ears. This effect has proved useful in acoustic detectors for gun ranging and underwater sound. It is also of great importance in stereophonic radio broadcasting and phonograph reproduction. Still more important, indeed, than the directionality effect is the ability conferred by the binaural effect to distinguish different sounds, that is, for example, meaningful speech from noise. It seems that the difference between meaningful sound and noise is enhanced, by some means as yet not too clear, by the difference in arrival time at the two ears, so that the listener and associated nervous system can make the necessary distinction.

Summation and difference tones

From the fact that the eardrum is an asymmetric vibrator (*i.e.*, its displacement for given applied pressure is not the same in the inward as in the outward direction) it follows theoretically that listeners should be able to hear summation and difference tones. What this means is that if two tones of widely separated frequencies f_1 and f_2 ($f_1 > f_2$) are presented to the human ear, in addition to these tones, the listener detects tones of frequency equal to their difference ($f_1 - f_2$) and equal to their sum ($f_1 + f_2$). As a matter of fact, tones of other combinations of frequency ($jf_1 + kf_2$), in which j and k are small arbitrary integers, have also been detected. If the original two tones, f_1 and f_2, are so close together that their difference is small compared with either, the usual phenomenon of beats is observed—*i.e.*, a periodic rise and fall in intensity with a frequency equal to their difference ($f_1 - f_2$).

Microphones. In modern terminology the word microphone is applied to any electroacoustic transducer used for the detection of sound. Actually, it might legitimately be used to denote a transducer capable of transforming the excess pressure in a sound wave into any other more readily observable physical effect (see SOUND RECORDING AND REPRODUCING). For example, a sensitive flame formed by burning gas issuing at high speed from a very small orifice and kept just below the point of flaring was an early form of microphone. In the presence of even a rather weak sound wave, the flame makes a perceptible jump. It can be made particularly sensitive to high-frequency, inaudible sound.

Such early microphones have now been entirely superseded by the electroacoustic variety. Such an electroacoustic microphone transforms the excess pressure changes due to an impinging sound wave into alternating electric currents that can be amplified and visualized on an oscillograph.

One of the most important properties of any microphone is its sensitivity or response, defined as the ratio of the electrical output (usually measured as the open-circuit voltage) to the root-mean-square (a statistical way of averaging) excess pressure at the microphone. Corrections must be made to take care of the fact that the presence of the microphone and its housing disturbs the acoustical field the instrument is inserted to measure. The response is commonly expressed in terms of decibels relative to some arbitrary reference level, often chosen as one volt per dyne per square centimetre.

Because the microphone diaphragm has mechanical elements, its vibration depends on the frequency of the incident radiation and the same holds true for the response. A typical response versus frequency curve for the simple carbon microphone used widely in telephony is shown in Figure 10. In this figure, zero dB corresponds to one volt

From *Applied Acoustics* by H. Olson and F. Massa (1939); by courtesy of Blakiston Division, McGraw-Hill Book Company, Inc.

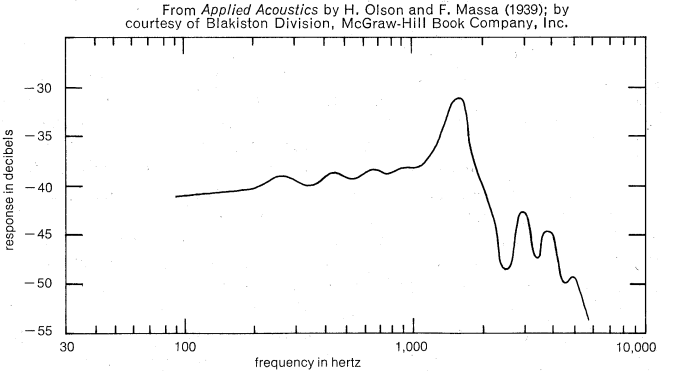


Figure 10: Frequency response of carbon microphone for acoustic pressure of one atmosphere (see text).

per dyne per square centimetre (dyne/cm^2). Hence −40 dB means 40 decibels in intensity level below one volt per dyne/cm^2. Power is proportional to the square of the voltage, and therefore the number of decibels corresponding to the voltage ratio (e_1/e_2) is $20 \log_{10} (e_1/e_2)$. Hence −40 dB relative to one volt per dyne/centimetre^2 is 10^{-2} volt per dyne/centimetre^2. It is seen from the figure that the response stays near this value from 100 hertz to one kilohertz. Between one kHz and two kHz it rises to between −30 dB and −35 dB and thereafter falls off sharply. The uneven response is due to the various modes of vibration of the microphone diaphragm.

The response referred to here corresponds to a plane acoustic wave incident normally on the diaphragm of the microphone; but another important characteristic of a microphone is its directivity pattern. This is a plot of the

response at a given frequency as a function of the angle of incidence of the radiation. In this sense some microphones—*e.g.*, the so-called ribbon-type—are more directional than others.

In the 1960s and 1970s great progress was made in the application of semiconductors (solids having special properties of electrical conductivity; see SEMICONDUCTORS AND INSULATORS, THEORY OF) to the construction of transducers for the detection of both static stresses and stresses at acoustic frequencies. One such type of energy conversion at high frequency is the piezoresistive variety in which stress applied to a semiconducting solid produces a change in its electrical resistance, which in turn can be translated into a voltage change and measured accordingly. Another type is based on the magnetoresistance effect in which the resistance of a semiconductor element varies with the change in the magnetic field intensity across it. If the magnetic field can be made to change with a displacement of the element produced by stress, the resulting device can be used as a sensitive microphone. The future of such devices seems bright.

Sound radiation detected by a microphone can be recorded or processed in a variety of ways. The voltage produced can be fed directly into a suitable voltmeter via an amplifier and measured directly in volts or translated into decibels. Special continuously recording meters have been devised for this purpose. When it is desirable to reproduce the signal from the microphone in direct visual form, a cathode-ray oscilloscope may be used. This is particularly appropriate when the radiation is in the form of a succession of pulses. Rather elaborate electronic circuitry is usually employed in modern acoustical reception techniques.

IV. Applications of acoustics

RECORDING AND REPRODUCTION

The visual recording of sound waves began with the use of devices such as the sensitive flame, mentioned in the last section. The development of sound recording for entertainment and other commercial purposes began with Thomas Edison's invention of the cylindrical phonograph record in 1877.

For audible sound recording and reproduction, three types of systems are in common use: mechanical, as in the disk phonograph; optical, as in the sound film; and magnetic, as in the tape or wire recorder (see SOUND RECORDING AND REPRODUCING).

The phonograph record

In the mechanical system the record is cut into a rotating disk of wax or foil with a cutting stylus, which is usually driven by electroacoustic coupling with the source being recorded. The record then consists of a spiral groove extending from the circumference of the disk to the centre. In conventional commercial records, the stylus vibrates transversely to the groove.

The standard, so-called long-playing 33⅓ rev/min (revolutions per minute) microgroove record has about eight grooves to the millimetre, with a maximum excursion of the needle or stylus of 0.035 millimetre (0.0014 inch). From a metal reproduction of the original cutting, copies can be made in large numbers on plastic disks.

For proper reproduction the disk must rotate at the same speed at which the recording was made, and a needle must follow the groove. This needle is part of a pickup microphone (usually of magnetic or piezoelectric-crystal type) that transforms mechanical vibrations into electrical oscillations and then back again into mechanical vibrations in a loudspeaker. An important requirement in cutting the original groove and in reproducing from the disk is uniformity of frequency response of the microphones and associated amplifying equipment in order to avoid distortion. The modern high-fidelity phonograph provides fairly uniform frequency response over the range from 25 hertz to 10 kilohertz.

Consideration of direction is important in achieving realistic reproduction from a spatially distributed source—*e.g.*, a large orchestra. When such an orchestra is listened to in a concert hall, one gets the sense of direction from the various parts. When the sound, however, is picked up by a single microphone for recording purposes, this sense of direction is lost. If two recordings are made, each from a microphone placed near or at each end of the ensemble, and these recordings are played back through a two-speaker system spaced so that their angular separation corresponds to that which the listener would experience in a live concert, the reproduction produces the impression of a stereophonic effect. Much progress was made in the 1960s and 1970s in improving stereophonic recording and reproduction so that in recent equipment the two signals are incorporated in a single groove on the record, one up and down (the so-called hill-and-dale type) and the other the conventional transverse type. A single transducer can pick up these two signals and channel them to separate loudspeaker systems.

In optical systems of recording, a beam of light is modulated by means of the sound waves. A photographic record of the modulated light then constitutes the sound track; this is the basic plan used in sound motion pictures. The sound may modulate the intensity of the light (now rarely used) or control the beam electromechanically by means of a chopper. The sound track may be either of variable density or variable area. In sound motion pictures it is usually placed directly on the film. In reproduction the sound track passes a beam of light that then falls on a photocell connected through an amplifier to a loudspeaker.

Magnetic-tape recording

In magnetic-tape recording, a thin plastic tape is coated with a film of magnetizable material. This tape is then moved at appropriate speed past an electromagnet (the so-called sound head) through which the output of a sound-recording microphone is passed. The current through the head magnetizes the tape and provides a more or less permanent record. For reproduction the tape is passed by another sound head (or in most modern tape recorders, the same sound head after the microphone is switched off) in which it induces a variable current that is amplified and fed into a loudspeaker. The record on the tape can be readily erased by using a larger demagnetizing field or more simply by the recording of another message on the same tape. Much used in commercial broadcasting and in many industrial operations (*e.g.*, computer storage), the magnetic-tape recorder has proved in the 1970s a serious competitor to the ordinary disk phonograph for home reproduction and general entertainment purposes.

ARCHITECTURAL ACOUSTICS

The production and reception of sound in closed spaces such as lecture and concert halls, theatres, church auditoriums, school classrooms, and home constitute the subject of architectural acoustics (see ACOUSTICAL ENGINEERING).

When a sound source is excited in a closed room, the resultant sound intensity at any point depends not solely on the sound waves reaching that point directly from the source but also on the arrival there of sound that has been reflected from the walls and other surfaces of the room. For rooms of the size encountered in ordinary experience the sound intensity builds up rather quickly, and the phenomenon known as reverberation ensues. If the room surfaces were perfect reflectors and the source were to continue to emit sound at a constant power output, the sound intensity would increase indefinitely, except for the attenuation of sound caused by the transmission through the air in the room. Actually, sound is always absorbed to a certain extent by any surface, and this keeps the intensity from building up indefinitely. But if the source stops it takes time for the intensity to decrease so as to render the sound inaudible. The time taken for the average sound intensity in a room to diminish to one-millionth of its original value at the instant the source was turned off is called the reverberation time. It is found experimentally that the reverberation time (T) depends directly on the volume of the room and inversely on the absorption of the sound by its various surfaces. If the fractional amount of sound energy absorbed by a surface is called the absorbing power α, the total amount of absorption is the integrated value of α over all the surfaces of the room. Thus, if 30 percent of the sound energy

incident on a particular surface is reflected back into the room, the value of α for that surface is 0.7. Each square metre thus acts as the equivalent of 0.7 square metre of a perfect absorber. The metric sabin (after Wallace Clement Sabine, a United States physicist) is the equivalent of one square metre of perfectly absorbing material. The corresponding English unit uses the square foot and is known simply as the sabin. If a is the total absorbing power of all the surfaces of the room, then integrating the fractional amount of sound energy absorbed over all the room surface, one gets $a = \int \bar{\alpha} dS = \bar{\alpha}S$, in which dS is the element of surface and S the total room surface. The average absorbing power per unit is $\bar{\alpha}$. The Sabine reverberation law may then be expressed as $T = 0.161v/\bar{\alpha}S$, in which T is the reverberation time in seconds, v is the volume of the room in cubic metres, and $\bar{\alpha}S$ is the absorption in metric sabins. The same formula applies if the volume is in cubic feet and the product $\bar{\alpha}S$ is in sabins, though the factor 0.161 then becomes 0.049. Various modifications of the Sabine formula have been proposed to meet the situation posed by extremely high surface absorption ("dead" rooms), and to take account of the absorption in the air of the room.

The sabin unit

The optimum acoustical conditions in a room depend on its intended use. For satisfactory reception of speech, a relatively short reverberation time is obviously essential for adequate articulation. With long reverberation time successive utterances overlap, unless the speech is abnormally slow. For auditoriums with volume of the order of 6×10^3 cubic metres to 14×10^3 cubic metres (2×10^5 to 5×10^5 cubic feet), experiments suggest a maximum reverberation time of about 2.5 seconds for satisfactory articulation. For best listening conditions in such rooms the reverberation time should be about one second. These figures indeed assume a low general noise level—*i.e.*, only about 30 decibels above minimum audible. In large rooms (of the order of 3×10^4 cubic metres or more) no speaker's voice can be heard without the use of a public address system, no matter how short the reverberation time. A room used primarily for music requires a longer reverberation time than one used for speech. For concert halls of the size mentioned above the optimum reverberation time for music is about 1.7 seconds, though for organ music this figure should be closer to two seconds. Inasmuch as absorption varies more or less directly with the frequency, the reverberation time varies inversely therewith. The above figures are for an average frequency of about 500 hertz.

Sabine's law indicates that the reverberation time for a room of given size can be controlled by appropriate changes in the absorption by the surfaces. Many kinds of acoustic absorbing materials exist to meet a wide variety of needs, both for permanent installation to meet the original design requirements and for portable installations to provide for flexibility in use.

In addition to the overall reverberation, and indeed as a contributing factor to it, the transmission of the sound back and forth between the walls, floor, and ceiling of a room stimulates the development of the normal modes mentioned above. Hence a complicated sound pattern can result in which certain frequencies are accentuated, leading to discomfort and difficulty in hearing. The excitation of the most troublesome normal modes can be materially reduced by the use of absorbing material, and also by introducing slight irregularities in the walls, so as to get rid of parallelism. These precautions break up the regular sound pattern and serve to diffuse the sound energy more equally throughout the space.

Modern architectural acoustics technology uses many special techniques, including the free use of sound-reinforcing systems with loudspeakers placed at strategic points throughout the space. The optimum acoustic design of auditoriums as well as schoolrooms, business offices, residences, etc., reached a high level of efficiency in the 1960s and 1970s. Much of the success of modern room acoustic design is due to the use of computers. A simulated model of a given hall can be introduced into a computer program that will then yield a faithful representation of the acoustical properties of the actual hall. Thus, the modifications necessary to give optimum conditions are relatively easy to determine while the hall is still in the design stage. Acoustic models of this kind will undoubtedly play a role of increasing importance in room acoustics and indeed in applied acoustics generally.

Computers used in acoustic design

MUSICAL ACOUSTICS

Until relatively recent times, acoustical theory had little bearing on musical instruments, even though historically much acoustical theory was developed in the attempt to explain how such instruments work. With the advent of sound recording and reproduction, the attempt to apply acoustical principles to the design of more effective instruments has led to considerable progress.

Conventional musical instruments are complicated examples of coupled vibrating systems—*i.e.*, systems with many degrees of freedom. Many of them, such as horns and woodwinds, function as acoustic filters, emitting only certain frequencies or combinations thereof. (For details see SOUND, MUSICAL; MUSICAL INSTRUMENTS; PERCUSSION INSTRUMENTS; STRINGED INSTRUMENTS; WIND INSTRUMENTS.)

The development of electroacoustics has encouraged the invention of many types of electrical musical instruments, such as electronic organs and related devices. Computer-composed music has also attracted considerable attention (see ELECTRONIC MUSIC AND INSTRUMENTS).

MILITARY APPLICATIONS

The principal military use of acoustics has been the detection of distant objects on the ground or under water. If an impulsive sound wave from a distant source like a gun is received by at least three microphones placed at precisely known intervals along a line (more than three are used in practice), knowledge of the sound velocity in air and the differences in time of arrival of the sound signal at the various microphones is sufficient to yield the bearing of the source and its distance from the microphone range. This is the basis for sound ranging in air. There are, to be sure, many complicating factors, such as the variation of sound velocity due to temperature gradients and wind velocity. Hence meteorological measurements are always important in sound ranging.

Seawater is in general a very good medium for the transmission of sound radiation. Moreover, its acoustic impedance (see the section above on properties of sound waves) matches solid sound sources much better than that of air. Hence sound ranging under water has become a very effective method for the detection of underwater objects such as submarines. Passive detection based on the sound emitted by an underwater vehicle is fundamentally similar to sound ranging in air. Active detection, called sonar (*q.v.*), is the more common method. In this, a beam of sound pulses sweeps the water horizontally; when it strikes a solid object, an echo is returned, providing a bearing on the object. The distance may be found from the knowledge of the sound velocity in water. This method is the acoustical analogue of radar. To secure a reasonably sharp beam from a not-too-large transducer surface (employing piezoelectric or magnetostrictive elements), a fairly high frequency must be used. On the other hand, because the absorption of sound in seawater increases more or less with the square of the frequency, there are limitations involved. It is usual to employ rather larger transducers with lower frequencies to secure long range. The use of sonar demands many correction techniques to take care of temperature gradients, changes in density, reflection from surface and bottom and other error-producing effects.

Sonar

In the 1970s sonar was becoming increasingly important not only as a military weapon but as a navigational device and a means for exploring the properties of the sea.

NOISE AND NOISE CONTROL

Noise or unwanted auditory experience was an important aspect of practical acoustics in the 1970s. Noise in the life of man had reached the level at which it was considered a form of environmental pollution. Noise control then developed into a major branch of acoustical engineering.

Every advance in the efficiency of the transformation of heat energy into mechanical energy has involved a noise problem, and in general this increases with power production. The large-scale turbulence of exhaust gases streaming from a jet airplane is an unusually intense source of sound in which the total acoustic power is proportional to something like the fifth power of the jet-stream velocity. The annoyance of this is evident to anyone who lives in the vicinity of an airport. When a plane moves faster than the speed of sound in air, the bow wave formed is shed as a shock wave (analogous to Cherenkov radiation in light), and this leads to the so-called sonic boom, a loud explosion as the shock wave hits the ground. This is one of the serious noise hazards of the supersonic jet plane.

Noise problems are not confined to aircraft. Factory noise has long been a hazard to workers. Control ranges from the attempt to make machines quieter by the use of absorbent mountings to the more careful construction of the moving parts in order to decrease vibrations leading to noise (see VIBRATIONS). To protect human ears when control at the source fails, ear plugs have been ingeniously designed to cut out loud sounds of certain frequencies while still permitting the hearing of conversational speech. Noise in buildings can be controlled by the use of absorbing materials on the walls or built into the partitions and floors. Traffic noise in large cities is amplified by reflection from the high-rise buildings, and this has placed emphasis on the importance of silencing motor cars. Many people are endangering their hearing by exposure to unnecessarily loud music. The noise-control problem is one of the most serious in modern society (see POLLUTION CONTROL).

Control of noise in buildings

ULTRASONICS IN INDUSTRY AND BIOMEDICAL SCIENCE

The possibility of producing ultrasonic radiation (frequency from 20 kilohertz to 10^{10} hertz) at high and controllable intensities has led to important industrial and medical applications (see ULTRASONICS AND INFRASONICS). The scattering of a thin beam of ultrasonic radiation by an irregularity or defect in an otherwise continuous solid provides an excellent means for the nondestructive testing of metals and other solids, as by the ultrasonic reflectoscope. Ultrasound can also produce emulsions more conveniently than by other means—*e.g.*, mercury in water. Many types of bacteria can be destroyed by irradiation with ultrasound. Certain types of chemical reactions are speeded up by ultrasonic radiation. Its use in underwater detection and depth finding has already been mentioned. Ultrasound has an important application to the cleaning of materials immersed in a liquid.

More recently, beams of ultrasound have been used in medical diagnosis in place of X-rays. Their use is also being explored as a therapeutic agent in arthritis. The field of bioacoustics, that is, the acoustical study of biological materials, including living tissue, has become an extensive branch of biomedical science through the use of ultrasonic radiation. Bioacoustics was also materially extended in the 1960s and 1970s by elaborate studies of the sounds produced by animals, especially marine animals, as well as the influence of sound of all frequencies and intensities on living organisms.

The practical applications of ultrasonics were much enhanced in the late 1960s and 1970s by the development of acoustical holography. In optical holography a three-dimensional object is illuminated with a beam of coherent laser light. Before the light reflected from various parts of the object is allowed to form an image on a photographic plate it is made to interfere with the original (reference) laser light. This interference makes the resultant image provide information about not only the intensity of the light from various parts of the object but also its phase (essentially difference in time of arrival). When the final record image (called a hologram) is examined in laser light the viewer sees what amounts to a three-dimensional view of the object in place of the standard two-dimensional photographic image. Acoustical holography works in essentially the same way, except that a beam of ultrasound (which, unlike ordinary light, is usually coherent) is used as the incident radiation. The sound reflected from various parts of an object must be made to form a visual image. This is then handled more or less as in the case of an optical hologram. Because ultrasound can penetrate solids opaque to light, acoustical holography can provide three-dimensional images of the constituents of solids, making it useful in nondestructive testing as well as in medical diagnosis (see HOLOGRAPHY).

Acoustical holography

BIBLIOGRAPHY. LORD RAYLEIGH, *Theory of Sound*, 2 vol. (1877–78; rev. ed., 1926; 1-volume ed., 1945), probably the most famous treatise on sound ever published, has remained an authority on nearly all aspects of theoretical acoustics since its publication. Twentieth-century accounts, particularly recommended, are F.V. HUNT, *Electroacoustics* (1954), a historical survey of the development of the application of electricity to acoustics; H.F. OLSON, *Acoustical Engineering*, 3rd ed. (1957) and *Music, Physics and Engineering* (1967), two books on various aspects of industrial acoustics, particularly with reference to sound recording and reproduction; R.B. LINDSAY, *Mechanical Radiation* (1960), an advanced treatment of the general principles of acoustical waves; L.E. KINSLER and A.R. FREY, *Fundamentals of Acoustics*, 2nd ed. (1962), a basic text in acoustics for advanced undergraduates; J.N. BRADLEY, *Shock Waves in Chemistry and Physics* (1962), a current treatise on the properties and applications of shock waves; R.W.B. STEPHENS and A.E. BATE, *Acoustics and Vibrational Physics*, 2nd ed. (1966), a current textbook for advanced undergraduates and graduate students; R.J. URICK, *Principles of Underwater Sound for Engineers* (1967), a practical treatise with emphasis on recent developments; P.M. MORSE and K.U. INGARD, *Theoretical Acoustics* (1968), an advanced and highly mathematical treatise; V.M. ALBERS, *The World of Sound* (1970), modern acoustics, popularly presented without mathematics; J. BACKUS, *Acoustical Foundations of Music* (1969), a study of acoustical principles at the basis of music, suitable for students of music; and K.D. KRYTER, *The Effect of Noise on Man* (1970), an important discussion of the modern noise problem. Important journals include the *Journal of the Acoustical Society of America* (1929–); *Acustica* (1951–), the international journal of acoustics; the *Journal of Sound and Vibration* (1964–); and *Ultrasonics* (1963–). *Soviet Physics—Acoustics* (1955–), in English translation, is available through the American Institute of Physics. These journals are mainly devoted to the publication of archival research in acoustics, but most of them contain notes on modern developments of a more popular character as well as news about acoustical scientists and engineers. Useful acoustical data will be found in the chapter on acoustics in the *American Institute of Physics Handbook*, 2nd ed. (1963; 3rd ed., 1972).

(R.B.L.)

Sound, Musical

That some sounds are intrinsically musical, while others are not, is an oversimplification. From the tinkle of a bell to the slam of a door, any sound is a potential ingredient for the kinds of sound organization called music. The choices of sounds for music making have been severely limited in all places and periods by a diversity of physical, aesthetic, and cultural considerations. This article will analyze those involved in Western musical traditions.

Characteristics of musical sound. The fundamental distinction usually made has been between tone and noise, a distinction best clarified by referring to the physical characteristics of sound. Tone differs from noise mainly in that it possesses features that enable it to be regarded as autonomous. Noises are most readily identified, not by their character but by their sources; *e.g.*, the noise of the dripping faucet, the grating chalk, or the squeaking gate. Although tones too are commonly linked with their sources (violin tone, flute tone, etc.), they more readily achieve autonomy because they possess controlled pitch, loudness, timbre, and duration, attributes that make them amenable to musical organization. Instruments that yield musical sounds, or tones, are those that produce periodic vibrations. Their periodicity is their controllable (*i.e.*, musical) basis.

The strings of the violin, the lips of the trumpet player, the reed of a saxophone, and the wooden slabs of a xylophone are all, in their unique ways, producers of periodic vibrations. The pitch, or high–low aspect, cre-

ated by each of these vibrating bodies is most directly a product of vibrational frequency. Timbre (tone colour) is a product of the total complement of simultaneous motions enacted by any medium during its vibration. Loudness is a product of the intensity of that motion. Duration is the length of time that a tone persists.

The vibrating reed

Each of these attributes is revealed in the wave form of a tone. The pattern may be visualized as an elastic reed —like that of a clarinet—fixed at one end, moving like a pendulum in a to-and-fro pattern when set into motion. Clearly, this reed's motion will be in proportion to the applied force. Its arc of movement will be lesser or greater depending upon the degree of pressure used to set it into motion. Once moving, it will oscillate until friction and its own inertia cause it to return to its original state of rest. As it moves through its arc the reed passes through a periodic number of cycles per time unit, although its speed is not constant. With these conditions prevailing, its motion through time could be charted by placing a carbon stylus on its moving head, then pulling a strip of paper beneath it at a uniform rate. The reed's

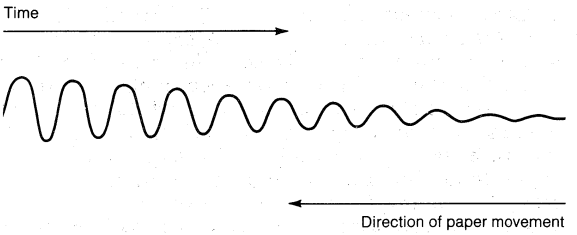


Figure 1: Visual representation of a reed's vibration.

displacement to-and-fro diminishes in a smooth fashion as time passes (decreasing intensity). Each cycle of its arc is equally spaced (uniform frequency). Each period of the motion forms the same arc pattern (uniform wave content). If this vibratory motion were audible, it could be described as follows: it grows weaker from the beginning (diminishing loudness) until it becomes inaudible; it remains at a stable level of highness (steady pitch); and it is of unvarying tonal quality (uniform timbre). If the reed were a part of a clarinet and the player continued blowing it with unvaried pressure, loudness, pitch, and timbre would appear as constants.

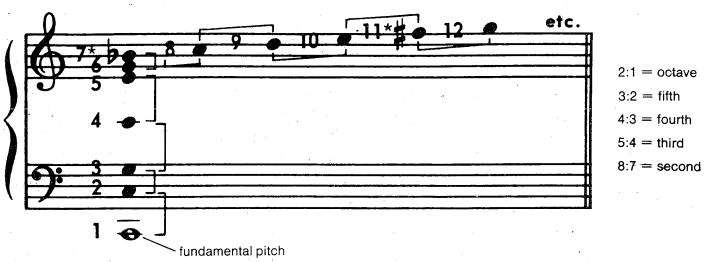
TONE

Most musical tones differ from the demonstration tone (above) in that they consist of more than the single wave form shown in Figure 1. Any material undergoing vibratory motion imposes its own characteristic oscillations on the fundamental vibration. The reed probably would vibrate in parts as well as a whole, thus creating partial wave forms in addition to the fundamental wave form. These partials are not fortuitous. They bear harmonic relationships to the fundamental motion that are expressible as frequency ratios of 1:2, 3:4, etc. This means that the reed (or string or air column as well) is vibrating in halves and thirds and fourths as well as a whole. Another way of expressing this is that half the body is vibrating at a frequency twice as great as the whole; a third is vibrating at a frequency three times greater; etc.

Overtone series

These numerical relationships also are expressible by pitch relationships as the harmonic, or overtone, series, which is merely a representation of numerical ratios in terms of pitch equivalents. Depending upon its shape and substance, a vibrating mass performs motions that are the equivalents of these partial vibrations, whether it be the mass of a string, reed, woodblock, or air column. This means that most tones are composites: they consist of partial vibrations of the vibrating body as well as the vibrations of the whole mass. Although one can develop the acuity required to hear some of these overtones within a musical tone, the ear normally ignores them as separate parts, recognizing only a more or less rich tone quality within the fundamental pitch.

Although pure tones, or tones lacking other than a fundamental frequency, sometimes occur in music, most



★ Actual pitch slightly lower than standard notation shows.

Figure 2: Overtone series.

musical tones are composites. A typical violin tone, for instance, is relatively rich in overtones while a flute tone sometimes approaches a pure tone (without overtones). What the listener recognizes as "a violin tone" or "a trumpet tone" also is a product of the noise content that accompanies the articulation of any sound

From C. Culver, *Musical Acoustics*. Copyright 1956. Used with permission of McGraw-Hill Book Company

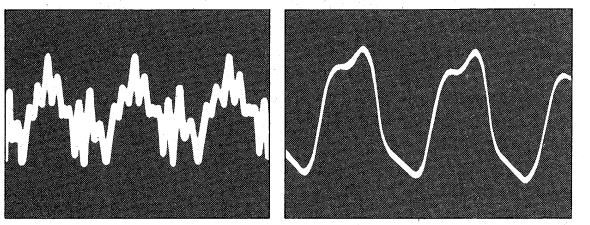


Figure 3: Comparison of (left) violin wave form, d′, and flute wave form, d′.

on the particular instrument. The friction of the bow as it is set into motion across the string, the eddies of air pressure within a horn's mouthpiece, or the hammer's impact on a piano string all add an extra dimension, a significant "noise factor," to any manually produced tone. After articulation, however, it is the presence or absence of overtones and their relative intensities that determine the timbre of any tone. The violin and flute tones are distinguishable because their articulatory "noises" are quite different; and, as Figure 3 reveals, their overtone contents are dissimilar, even when they produce the same pitch. Figure 4 shows a comparative graphing of the overtone content within those two wave forms.

Electronic sound

Musical tones of determined harmonic content can be produced by electronic vacuum tubes or transistors as

From C. Culver, *Musical Acoustics*. Copyright 1956. Used with permission of McGraw-Hill Book Company

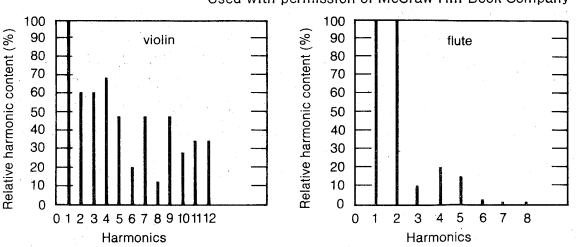


Figure 4: Relative partial content of the violin and flute tones.

well as by traditional manual instruments. Some electronic organs, for example, use single vacuum tubes whose frequency output can be varied through control of an adjustable transformer. Through ingenious mixing circuits a compound tone consisting of any predetermined overtone content can be produced, thereby imitating the sound of any traditional instrument. Composers of electronic music have utilized this capability to synthesize tones quite different from any available on traditional instruments, as well as tones similar to natural sounds. Electronic computers are capable of complete imitation of such sounds; the tone is broken down into its component parts, then synthesized through an auditory output circuit.

MOVEMENT

Once an audible oscillation is produced by a vibrating body, it moves away from its source as a spherical pressure wave. Its rate of passage through any medium is determined by the medium's density and elasticity; the denser the medium, the slower the transmission; the greater the elasticity, the faster. In air at around 60° F, sound moves at approximately 1,120 feet per second, the rate increasing by 1.1 feet per second per degree of rise in temperature.

Sound waves move as a succession of compressions through the air. The wavelength is determined by frequency; the higher the pitch, the shorter the wavelength. A pitch of 263 cycles per second (middle C of the piano) is borne as a wavelength of around 4.3 feet (speed of sound ÷ frequency = wavelength).

By the time a wave has moved some distance it has changed in some of its characteristics. The journey has robbed it of intensity, which is inversely proportional to the square of the distance. Its timbre has been altered slightly by objects within its path that disrupted an equitable distribution of frequencies, particularly the high frequency waves, which, unlike the low, move in relatively straight paths from their sources.

Resonance

The area within which a sound occurs can have considerable effect upon what is heard. Just as a string or reed or air column has a natural resonance period (or rate of vibrations), any enclosure—whether an audio speaker cabinet or the nave of a cathedral—imposes its resonance characteristics on a sound wave within it. Any tone that approximates in frequency the characteristic resonance period of an enclosure will be reinforced through the sympathetic response, or natural resonance, of the air within the enclosure. This means that tones of frequencies differing from the resonance of the enclosure will be less intense than those that agree, thereby creating an inequity of sound intensities.

Fortunately, most rooms where music is performed are large enough (wall lengths greater than about 30 feet) so that their natural resonance periods are too slow to fall within the range of pitches of the lowest musical tones (usually no lower than 27 cycles per second, although some organs have pipes that extend to 15 cycles per second). Smaller rooms can produce disturbing sympathetic resonance unless obstructions or absorbent materials are added to minimize that effect. (Bathroom singers revel in this phenomenon because the band of resonance sometimes lies close enough to the pitches of the male voice to support it, making it appear richer and more powerful.)

Reverberation period

In addition to resonance, any enclosure possesses a reverberation period, a unit of time measured from the instant a sound fills the enclosure (steady state) until that sound has decayed to one-millionth of its initial intensity. Anyone who has spoken or clapped his hands inside a large, empty room has experienced prolonged reverberation. There are two reasons for such protracted reverberation: first, the space between the surfaces of the enclosure is so great that reflected sound waves travel extended distances before decaying; and, second, the absence of highly absorbent materials precludes appreciable loss of intensity of the wave during its movement.

The reverberation period is a crucial factor in rooms where sounds must be heard with considerable fidelity. If the period is too long in a room where speech must be understood, spoken syllables will blend into each other and the words will be mumbled confusion. If, on the other hand, the reverberation period is too brief in a room where human "presence" and music each contribute to the acoustics, only a "cold" and "dull" feeling will persist, because no reverberative support of the prevailing sounds can be provided by the enclosure itself.

Wise architects plan auditoriums to fulfill particular needs by tailoring reverberation to the expected speech and music activities. A magnificent pipe organ placed in a church designed to meet only the reverberation needs of speech cannot fulfill its tonal potential because the room's dead characteristics will provide no support. Since speech and music demand slightly different reverberation periods, a compromise must be effected to serve the needs of both. A reverberation period should be greater in a church where music plays a significant role than in a church where the spoken word is dominant and music is incidental.

From C. Culver, *Musical Acoustics*. Copyright 1956. Used with permission of McGraw-Hill Book Company

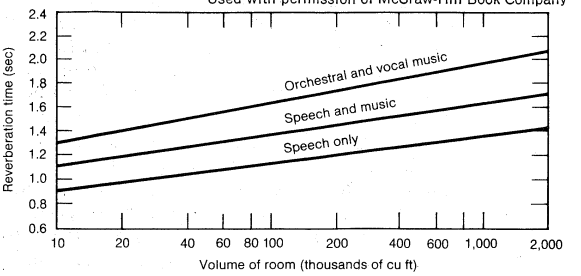


Figure 5: Optimum reverberation periods of 500 cycles per second as a function of total enclosure volume.

Although all sound waves, regardless of their pitch, travel at the same rate of speed through a particular medium, low tones mushroom out in a broad trajectory while high tones move in straight paths. For this reason listeners in any room should be within a direct path of sound propagation. Seats far to the side at the front of an auditorium offer occupants a potentially distorted version of sound from its source. Thus the high frequency speakers (tweeters) in good audio reproduction systems are angled toward the sides of the room, ensuring wider coverage for high frequency components of all sounds.

Sites of musical performance in the open demand quite different acoustical arrangements, of course, since sound reflection from ceilings and walls cannot occur and reverberation cannot provide the desirable support that would be available within a room. A reflective shell placed behind the sound source can provide a boost in transmission of sounds toward listeners. Such a reflector must be designed so that relatively uniform wave propagation will reach all locations where listening will occur. The shell form serves that purpose admirably since its curved shape avoids the right angles that might set up continuous reflections, or echoing. Furthermore, sound waves are reflected more uniformly over a wide area than with any other shape, diffusing them equally over the path of propagations. (The needs here are similar to those of the photographer who wishes to flood a scene uniformly with flat light rather than focus with a spotlight on a small area.)

PITCH AND TIMBRE

Just as various denominations of coins combine to form the larger units of a monetary system, so musical tones combine to form larger units of musical experience. Although pitch, loudness, duration, and timbre act as fourfold coordinates in the structuring of these units, pitch has been favoured as the dominating attribute by most Western theorists. The history of music theory has to a great degree consisted of a commentary on the ways pitches are combined to make musical patterns, leaving loudness and timbre more as the "understood" parameters of the musical palette.

Music terminology, for example, recognizes loudnesses in music in terms of an eight-level continuum of nuances from "extremely soft" (ppp, or pianississimo) to "extremely loud" (fff or fortississimo). (The musical dominance of Italy from the late 16th to the 18th century—when these Italian terms first were applied—explains their retention today.)

The timbres of music enjoy an even less explicit and formalized ranking; other than the vague classifications "shrill," "mellow," "full," and so on, there is no standard taxonomy of tone quality. Musicians for the most part are content to denote a particular timbre by the name of the instrument that produced it.

Divisions of the pitch spectrum

Pitch is another matter. A highly developed musical culture demands a precise standardization of pitch, and Western theory has been occupied with this task from as early as Aristoxenus (4th century BC). Especially since the Renaissance, when instruments emerged as the prin-

cipal vehicles of the musical impulse, problems of pitch location (tuning) and representation (notation) have challenged the practicing musician. When at least two instrumentalists sit down to play a duet, there must be some agreement about pitch, or only frustration will result. Although the standardization of the pitch name a^4 (within the middle of the piano keyboard) at 440 cycles per second has been adopted by most of the professional music world, there was a day—even during the mid-18th century of Bach—when pitch uniformity was unknown.

Man's perception of pitch is confined within a span of roughly 15 to 18,000 cycles per second. This upper limit varies with the age and ear structure of the individual, the upper limit normally attenuating with advancing age. The pitch spectrum is divided into octaves, a name derived from the scale theories of earlier times when only eight (Latin *octo*) notes within this breadth were codified. Today the octave is considered in Western music to define the boundaries for the pitches of the chromatic scale. The piano keyboard is a useful visual representation of this 12-unit division of the octave. Beginning on any key, there are 12 different keys (and thus 12 different pitches), counting the beginning key, before a key occupying the same position in the pattern recurs.

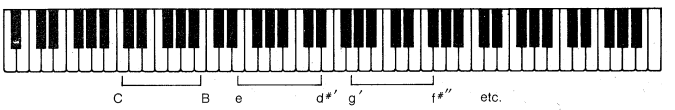


Figure 6: Piano keyboard.

One must keep in mind that the chromatic scale, within the various octave registers of man's hearing, is merely a conventional standard of pitch tuning. Performers like singers, trombone and string players, who can alter the pitches they produce, frequently make use of pitches that do not correspond precisely to this set of norms. The music of many non-Western cultures also utilizes distinct divisions of the octave. Furthermore, some contemporary music makes use of pitch placements that divide the octave into units smaller than the half-step. This music, called microtonal, has not become standard fare in Western cultures, in spite of its advocates (Alois Hába, Julian Carillo, Karlheinz Stockhausen) and even its special instruments that provide a means for consistent performance.

Western music history is dotted with systems formulated for the precise tuning of pitches within the octave. From a modern viewpoint all suffer from one of two mutually exclusive faults: either they lack relationships (intervals) of uniform size, or they are incapable of providing chords that are acceptable to the ear. Pythagorean tuning provides uniformity but not the chords. Just tuning, based on the simpler ratios of the overtone series, provides the chords but suffers from inequality of intervals. Meantone tuning provides equal intervals but gives rise to several objectionable chords, even in simple music. All three of these systems fail to provide the pitch wherewithal for the 12 musical keys found in the standard repertoire (see also TUNING AND TEMPERAMENT, MUSICAL).

The compromise tuning system most widely accepted since the mid-19th century is called "equal temperament." Based on the division of the octave into 12 equal half-steps, or semitones, this method provides precisely equal intervals and a full set of chords that, although not as euphonious as those of the overtone series, are not offensive to the listener.

The semitone is the smallest acknowledged interval of the Western pitch system. The sizes of all remaining intervals can be calculated by determining how many semitones each contains. The names of these intervals are derived from musical notation through a simple counting of lines and spaces of the staff. Just as the overtone content of a single tone determines timbre, the relationship of the constituent pitches of an interval determines its quality, or sonance. There is a long history of speculations in this area, but the subjectivity of the data indicates that little verifiable fact can be sorted from it.

Consonance and dissonance

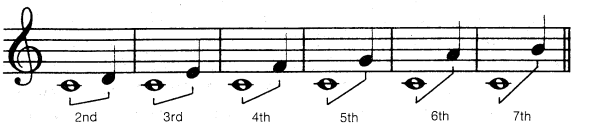


Figure 7: Generic names of pitch intervals.

Until the 20th century, music theorists were prone to concoct tables that showed an "objective" classification of intervals into the two opposing camps of consonant and dissonant. But only the person who utters these terms can know with assurance what he means by them, although many attempts have been made to link consonant with pleasant, smooth, stable, beautiful and dissonant with unpleasant, grating, unstable, and ugly. These adjectives may be reasonably meaningful in musical contexts, but difficulty arises if one attempts to pin a singular evaluation on a particular interval per se.

Theorists have noted that the character of an interval is altered considerably by the sounds that surround it. Thus the naked interval that sounds "grating," "unstable," and lacking in fusion might within a particular context create an altogether different effect, and vice versa.

Recognition of the power of context in shaping a response to the individual pitch interval has led some music theorists to think more in terms of a continuum of sonance that extends from more consonant to more dissonant, tearing down the artificial fence once presumed to separate the two in experience.

Helmholtz theory

The explanation of consonance and dissonance offered by H.L.F. von Helmholtz in *On the Sensations of Tone* (1863) is perhaps as helpful as any. An initial theory was based on the notion that dissonance is a product of beats, which result from simultaneous tones or their upper overtones of slightly differing frequencies. Another explanation, offered later by Helmholtz, held that two tones are consonant if they have one or more overtones (excluding the seventh and ninth) in common.

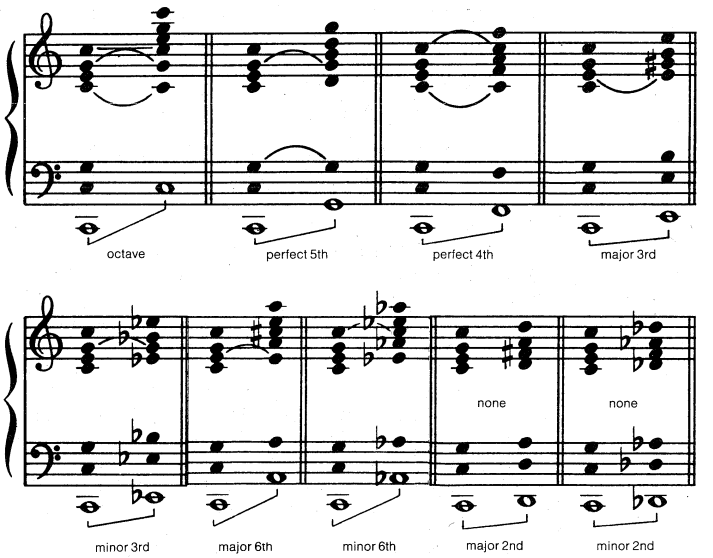


Figure 8: Common overtones (incomplete series, excluding the seventh) at various pitch intervals.

Music in which a high degree of dissonance occurs has rekindled interest in this old problem of psychoacoustics. The German composer Paul Hindemith (1895–1963) provided one explanation of harmonic tension and relaxation that depends upon the intervals found within chords. According to his view a chord is more dissonant than another if it contains a greater number of intervals that, as separate entities, are dissonant. Although Hindemith's reasonings and conclusions have not been widely accepted, the absence of any more convincing explanation and classification often leads musicians to use his ideas implicitly.

Although the complete pitch spectrum can be tuned in a way that provides 12 pitches per octave (as the chromatic

scale), pitch organization in music usually is discussed in terms of less inclusive kinds of scale patterns. The most important scales in traditional Western theory are seven-tone (heptatonic), which, like the chromatic, operate within the octave. These scales are different from one another only in the intervals formed by their constituent pitches. The major scale, for instance, consists of seven pitches arranged in the intervallic order: tone–tone–semitone–tone–tone–tone–semitone.

Called major because of the large (or major) third that separates the first and third pitches, this scale differs from the minor scale mainly in that the latter contains a small (or minor) third in this location. Since three variants of the minor scale are recognized in the music of the Western repertoire, it is important to note that they share this small interval between their first and third pitches.

Scales and modes

Major and minor scales formed the primary pitch ingredients of music written between 1650 and 1900, although this is a sweeping generalization for which exceptions are not rare. Other scales, called modes, possess greater representational power for music of earlier times and for much of the repertoire of Western folk music. These too are heptatonic patterns, their uniqueness produced solely by the differing pitch relationships formed by their members. Each of the modes can most easily be reproduced by playing successive white keys at the piano.

The modes and the major and minor scales best represent the pitch structure of Western music, though they do not utilize the total complement of 12 chromatic pitches per octave. They are abstractions that are meaningful for tonal music; *i.e.*, music in which a particular pitch acts as a focal point of perception, establishing a sense of repose or tonality to which the remaining six pitches relate. Major and minor scale tonality was basic to Western music until it began to disintegrate in the art music of the late 19th century. It was replaced in part by the methods of Arnold Schoenberg (1874–1951), which used all 12 notes as basic material. Since that revolution of the early 1920s, the raw pitch materials of Western music have frequently been drawn from the complete chromatic potential. By contrast, the music of several Eastern cultures, a number of children's songs, and occasional Western folk songs incorporate pitch materials best classified as *pentatonic* (a five-pitch scale).

SOUND PRODUCTION OF MUSICAL INSTRUMENTS

Classification of musical instruments

Excluding electronic tone synthesizers, which employ vacuum tubes or transistors to produce tones, musical instruments can be classified within three groups: (1) chordophones, or strings; (2) aerophones, or winds; and (3) idiophones and membranophones, nearly all of which are percussion instruments. Each category is further divisible into groups according to the way the vibrating medium is set into motion.

Chordophones. Three means of eliciting sounds determine three categories within the family of chordophones. They are bowing, plucking, and striking. Most common of the first category are the violin, viola, violoncello, and double bass of the orchestra, all of which use a horsehair bow for setting their strings into motion. Essentially a resonant box bearing strings of four different fundamental frequencies, members of this group have not changed appreciably in construction since the 17th century, except for the 20th-century advent of the electrified bass, which is in fact a close cousin of the amplified guitar.

Violins and the larger members of its group are sounded by plucking (pizzicato) on occasion, which provides a brittle tone of extremely brief duration. The harp is the best known orchestral instrument whose tone depends upon the noise components added by plucking. Other plucked instruments are the guitar, banjo, mandolin, ukelele, zither, lyre, lute, and the harpsichord. The latter differs from the piano in that its strings are actuated by the plucking action of a tiny plectrum.

The piano is most notable of the struck string instruments, employing a hammer mechanism linked with the keyboard for producing its wide range of sounds. Other instruments of this group are the clavichord and the dulcimer.

For all chordophones pitch is proportional to string tension and inversely proportional to length, thickness, and density. Since string length is the most readily altered of these factors, all chordophones provide a means for altering the resonating length of strings (as with the violin and guitar) or a set of many string lengths and masses (as with the piano and harp) for producing a variety of pitches.

Aerophones. This category covers everything from the piccolo to the pipe organ and is best understood by consistent reference to the nature of the air column employed in the various types of instruments, as well as the way this air column is set into motion.

Brass instruments consist of a long tube whose cross section is proportionately small. Coupled with a mouthpiece that, in response to vibrations of the performer's lips, helps to create eddies of air pressure that set an enclosed air column into motion, these instruments produce a range of pitches corresponding to the overtone series. The bugle is a primitive kind of brass instrument in that it is limited to only one overtone series, while the modern trumpet, cornet, French horn, trombone, tuba, flügelhorn, and various kinds of euphoniums utilize valves or a slide to lengthen the air column and thus provide up to seven different overtone series. Pitch on these instruments is primarily a function of tube length, the wavelength of the instrument's fundamental pitch equal to twice the length of the tube, plus a so-called end correction that accommodates variations of bore. Timbre is a product of mouthpiece shape, bore (whether cylindrical or conical), and material, aside from the important role performed by the player himself in obtaining desired overtones.

Woodwinds prior to the 20th century were made for the most part of wood. Today the flutes and clarinets are classified in this group only because of this heritage, while the saxophones, always built of metal, share only the reed mouthpiece and similar fingering technique with the clarinet. All are, nonetheless, called woodwinds, and they consist of an air column set into motion by one of two means: (1) through high pressure eddies produced by the wind of the performer blown directly into the instrument (as with a recorder or whistle) or over it (as with the flute and piccolo), or (2) by means of a vibrating reed that is set into motion by air pressure from the performer. The clarinets and saxophones utilize a single reed fixed at one end, while the oboe, English horn, and bassoon use two thin reeds that are connected laterally and vibrate jointly. For all of these instruments, either keys or the fingers of the performer directly open holes, with the effect of shortening the enclosed air column of the instrument and thereby producing higher fundamental pitches. Through overblowing and various fingering procedures, the overtone series provides the wealth of pitches available on these instruments.

Free reed instruments utilize a single, freely vibrating reed, different in nature from that of a woodwind. The category includes the accordion, harmonica, and harmonium and their relatives. In these instruments the reed vibrates, causing periodic vibrations in the air; but the reed's size, rather than the air enclosed by the instrument, determines the pitch.

Pipe organs are of the aerophone (wind) category, too, although their keyboard mechanism and literature link them closely with the piano and harpsichord. Like a grand synthesis of woodwinds and brasses, organs produce their tones by means of tuned air columns that are formed with pipes of varied length, cross section, and shape (called flue pipes) or by means of a vibrating brass reed actuated by forced air (called reed pipes). Flue pipes range in length from under an inch to 32 feet.

Idiophones and membranophones. Idiophones are instruments whose body vibrates to produce sound. The class contains most of the pitched percussion instruments. These include instruments made of wood or other organic material, such as xylophones. They also include pitched percussion instruments that are struck or plucked and are made of metal or other inorganic material (triangle, glockenspiel, vibraphone, celesta, tubular bells,

gongs, steel drums, cymbals, glass harmonica, etc.). Idiophones without pitch consist of such instruments as the percussion board, castanets, and rattles, all of which are made of wood or other organic material and are struck, scraped, rubbed, brushed, or shaken.

Membranophones produce sound by a vibrating membrane. The group consists most notably of the timpani or kettledrum, which can be tuned by increasing or decreasing the tension of the membrane that forms the head of the enclosed cavity. Other membranophones consist of drums without fixed pitch, such as side drums, bongos, and various non-Western types of fixed and indefinite pitch. Tone quality and character are the result of the player's skill in controlling intensity and overtone character of the sound.

BIBLIOGRAPHY. This listing proceeds from the less technical, general references to those aimed at the technical reader of specialized interests.

WILLEM A. VAN BERGEIJK, J.R. PIERCE, and E.A. DAVID, *Waves and the Ear* (1960); ARTHUR H. BENADE, *Horns, Strings and Harmony* (1960); CHARLES A. CULVER, *Musical Acoustics*, 4th ed. (1956); JOHN BACKUS, *The Acoustical Foundations of Music* (1969); JOHN MILLS, *A Fugue in Cycles and Bels* (1935); MAX F. MEYER, *How We Hear: How Tones Make Music* (1950); LLEWELYN S. LLOYD, *The Musical Ear* (1940); ALEXANDER WOOD, *The Physics of Music*, 6th ed. (1962); HERMANN VON HELMHOLTZ, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*, 4th ed. (1877, with suppl. to 1885; Eng. trans., *On the Sensations of Tone As a Physiological Basis for the Theory of Music*, 2nd ed., 1885, reprinted 1961); RENE DESCARTES, *Musicae compendium* (1650; Eng. trans., *Compendium of Music*, 1961); CORNELIS J. NEDERVEEN, *Acoustical Aspects of Woodwind Instruments* (1969); HARRY F. OLSON, *Music, Physics and Engineering*, 2nd ed. (1967); LEO LEROY BERANEK, *Music, Acoustics and Architecture* (1962); JAMES MURRAY BARBOUR, *Tuning and Temperament: A Historical Survey*, 2nd ed. (1953); PAUL HINDEMITH, *Unterweisung im Tonsatz*, vol. 1 (1937; Eng. trans., *The Craft of Musical Composition*, vol. 1, 1941).

(W.E.T.)

Sound Reception

Sound waves are a particular kind of mechanical activity consisting of vibrations in a gas, liquid, or solid medium. If an animal possessing an auditory mechanism comes in suitable contact with a medium vibrating at a frequency and intensity within its range of aural (hearing) sensitivity, it may hear the sound. For land animals, the usual vibrating medium is the air; for fishes and other aquatic creatures, it commonly is the water. Yet, under suitable conditions, all hearing animals can perceive sound waves transmitted by media other than the one in which they live; thus, humans can hear noise while underwater (see SOUND).

This article is divided into the following major sections:

I. Sound reception in invertebrates
  Hearing in insects
  Hearing in spiders
II. Sound reception in vertebrates
  Hearing in fishes
  Hearing in amphibians
  Hearing in reptiles
  Hearing in birds
  Hearing in mammals

In the course of evolution, animals have developed a variety of sense organs that respond to mechanical stimuli. There are at least ten of these mechanoreceptors in vertebrates and perhaps as many in advanced invertebrates. Not all of these structures respond to sound, however, for among them are the simple touch endings of the skin and the motion receptors that serve (mediate) bodily equilibrium (see MECHANORECEPTION). Although the different ways of registering mechanical changes in the environment or within the body represent various structural specializations, it is not feasible to identify any one of them simply in terms of its structure; many different mechanisms, cells, or organs may perform similar functions. Ears, for example, take many forms in the lower animals and often have little resemblance to these organs in man and other higher vertebrates. Yet the service that they perform in sound reception is similar enough that they may be called ears.

Although there is no fossil record of the origin and development of auditory structures, in animals with ears the evolutionary process in every instance appears to have been a conversion to an auditory function of structures that previously mediated a simpler form of mechanoreception. Indeed, any mechanoreceptor, even though best adapted to respond to some other form of mechanical stimulation, will respond to vibrations within some region of the sound frequency range if the vibrations have a sufficiently high level of intensity.

Attempts to define hearing

Many attempts have been made to define hearing, often with indifferent success. The task is difficult, and in certain respects the lines of distinction are arbitrary. The ear cannot be identified by any standard structure, nor can it be identified in terms of the stimulus as simply a receiver of sound vibrations. As noted above, mechanical receptor organs will respond to sound vibrations within some region of the frequency range if a sufficiently high level of intensity is provided. Moreover, the ear cannot be characterized in terms of the physical principles by which it operates because these principles vary among the ears of different animal species.

A definition of hearing, therefore, must be sought in terms of the ear's specialization of function and the relative effectiveness with which it performs this function. Thus, hearing may be characterized as the reception of sound vibrations by an organ, the ear, that has developed for this particular purpose and that has reception of sound as its primary function. This definition excludes the reception of sound vibrations by touch (tactual) endings in the skin, for example, because these structures respond most readily to direct pressure. Before such receptors will respond to sound waves, the vibrational intensity of the sound must be relatively great. Also excluded are the hair sensilla, of which arthropods have many types, whenever it can be shown that these organs respond with greater sensitivity to another stimulus (most often a simple direct deflection of the central hair).

Principles of sound reception

Theoretically, several aspects of vibration might serve in its detection by an ear. These characteristics include the amplitude (extent) of the motion of particles (*e.g.*, molecules) in a medium, the velocity and acceleration of the motion, the pressure exerted upon an obstacle in the path of the sound waves, and temperature changes occasioned by the vibrations. All of these manifestations have been utilized in attempts to design microphones for the detection and measurement of sound, but only two (pressure and velocity effects) have proved to be of any practical value. Thus, those devices that employ these two effects are known as pressure and velocity microphones.

It seems more than coincidence that these same two aspects of sound, pressure and velocity, are the only stimulus characteristics on which the evolution of ears appears to have been based. Moreover, just as the pressure microphone is the most practical type designed by man, among ears the pressure type is the most widespread and the most highly developed. Ears that distinguish changes in velocity have appeared only in a few lower animals—as an elaborated hair organ in some insects and perhaps spiders and in two special forms among fishes. All other ears are pressure receptors that have taken two lines of evolutionary development, one in most of the insects and another in vertebrates above fishes.

Types of animals that have ears. Considering the usefulness of the sense of hearing to such highly organized animals as man, it may seem surprising that this sense is so limited in its appearance and development among animals. It is found only in two major groups of animals: arthropods (*e.g.*, insects and crabs) and vertebrates (*e.g.*, amphibians, birds, and mammals). The condition that probably limited the development of hearing in other species was the lack of sufficient advancement and flexibility of the nervous system.

Biological value of ears

In those animals with auditory structures, hearing serves purposes of great biological value: in its more primitive forms, it is used to sense danger and enemies, to detect

prey, and to identify prospective mates; at a more complex level, hearing is involved in communication within social groups and in emotional expressions of various kinds. The cry of an infant mouse that has strayed from the nest elicits a response by the mother to retrieve it. The singing of a male thrush asserts a claim to its territory, attracts a female to the area, and warns off other males. Among higher mammals (*e.g.*, monkeys and apes) vocalizations show even greater variety and express a range of meanings that may be interpreted in human terms as expressions of such concepts as danger, aggression, love, and the availability of food. In man, the elaborations of auditory communication can be even more symbolically complex, extending to speech and music (see SPEECH, PHYSIOLOGY OF; EAR AND HEARING, HUMAN). The significant features in complicated sounds that people perceive and differentiate correspond to the physical dimensions of frequency (the number of waves, cycles, or vibrations per second), intensity, phase, complexity of wave form, and temporal pattern. The variety of distinguishable acoustic forms is enormous.

Among the most highly refined applications of the auditory sense are those found in such animals as bats and dolphins. These creatures are able to discern objects around them by a process called echolocation; the animal sends out a cry and, by the nature of the echo, is informed of the presence of obstacles or potential prey. For these animals, the sense of hearing provides a service in the dark that closely approaches the reliability of vision in the perception of objects and spatial relations.

I. Sound reception in invertebrates

HEARING IN INSECTS

It has long been believed that at least some insects can hear. Chief attention has been given to those that make distinctive sounds (*e.g.*, katydids, crickets, and cicadas) because it was naturally assumed that these insects produce signals for communication purposes. Organs suitable for hearing have been found in insects at various locations on the thorax and abdomen and, in one group (mosquitoes), on the head.

Among the many orders of insects, hearing is known to exist in only a few: Orthoptera (crickets, grasshoppers, katydids), Homoptera (cicadas), Heteroptera (bugs), Lepidoptera (butterflies and moths), and Diptera (flies). In the Orthoptera, ears are present, and the ability to perceive sounds has been well established. The ears of katydids and crickets are found on the first walking legs; those of grasshoppers are on the first segment of the abdomen. Cicadas are noted for the intensity of sound produced by some species and for the elaborate development of the ears, which are located on the first segment of the abdomen. The waterboatman, a heteropteran, is a small aquatic insect with an ear on the first segment of the thorax. Moths have simple ears that are located in certain species on the posterior part of the thorax and in others on the first segment of the abdomen. Among the Diptera, only mosquitoes are known to possess ears; they are located on the head as a part of the antennae.

Location of insect ears

All the insects just mentioned have a pair of organs for which there is good evidence of auditory function. Other structures of simpler form that often have been considered to be sound receptors occur widely within these insect groups as well as in others. There is strong evidence that some kind of hearing exists in two other insect orders: the Coleoptera (beetles) and the Hymenoptera (ants, bees, and wasps). In these orders, however, receptive organs have not yet been positively identified.

Types of auditory structures. Four structures found in insects have been considered as possibly serving an auditory function: hair sensilla, antennae, cercal organs, and tympanal organs.

Hair sensilla. Many specialized structures on the bodies of insects seem to have a sensory function. Among these are hair sensilla, each of which consists of a hair with a base portion containing a nerve supply. Because the hairs have been seen to vibrate in response to tones of certain frequencies, it has been suggested that they are sound receptors. It seems more likely, however, that the sensilla primarily mediate the sense of touch and that their response to sound waves is only incidental to that function.

Antennae and antennal organs. Many sensory functions have been attributed to the antennae of insects, and it is believed that they serve both as tactual and as smell receptors (see CHEMORECEPTION). In some species, the development of elaborate antennal plumes and brushlike terminations has led to the suggestion that they also serve for hearing. This suggestion is supported by positive evidence only in the case of the mosquito, especially the male, in which the base of the antenna is an expanded sac containing a large number of sensory units known as scolophores. These structures, found in many places in the bodies of insects, commonly occur across joints or body segments, where they probably serve as mechanoreceptors for movement (kinesthetic organs). When the scolophores are associated with any structure that is set in motion by sound, however, the arrangement is that of a sound receptor.

Structure of the scolophore

The basic structure of the scolophore is shown in Figure 1. Four cells (base cell, ganglion cell, sheath cell, and terminal cell), together with an extracellular body called a cap, constitute a chain. Extending outward from the ganglion cell is the cilium, a hairlike projection that, because of its position, acts as a trigger in response to any relative motion between the two ends of the chain. The sheath cell with its scolopale provides support and protection for the delicate cilium. Two types of enclosing cells (fibrous cells and cells of Schwann) surround the ganglion and sheath cells. The ganglion cell has both a sensory and a neural function; it sends forth its own fibre (axon) that connects to the central nervous system.

In the mosquito ear (Figure 1) the scolophores are connected to the antenna and are stimulated by vibrations of the antennal shaft. Because the shaft vibrates in response to the oscillating air particles, this ear is of the velocity type. It is supposed that stimulation is greatest when the antenna is pointed toward the sound source, thereby enabling the insect to determine the direction of sounds. The male mosquito, sensitive only to the vibration frequencies of the hum made by the wings of the female in his own species, flies in the direction of the sound and finds the female for mating. For the male yellow fever mosquito, the most effective (*i.e.*, apparently best heard) frequency has been found to be 384 hertz, or cycles per second, which is in the middle of the frequency range of the hum of females of this species. The antennae of insects other than the mosquito and its relatives probably do not serve a true auditory function.

Cercal organs. The cercal organ, which is found at the posterior end of the abdomen in such insects as cockroaches and crickets, consists of a thick brush of several hundred fine hairs. When an electrode is placed on the nerve trunk of the organ, which has a rich nerve supply, a discharge of impulses can be detected when the brush is exposed to sound. Sensitivity extends over a fairly wide range of vibration frequencies, from below 100 to perhaps as high as 3,000 hertz. As observed in the cockroach, the responses to sound waves up to 400 hertz have the same frequency as that of the stimulus. Although the cercal organ is reported to be extremely sensitive, precise measurements remain to be carried out. It is possible, nevertheless, that this structure, which is another example of a velocity type of sound receptor, is primarily auditory in function.

Tympanal organs. The tympanal organ of insects consists of a group of scolophores associated with a thin, horny (chitinous) membrane at the surface of the body, one on each side. Usually the scolophores are attached at one end by a spinous process to the tympanic membrane (eardrum); the other ends rest on an immobile part of the body structure. When the membrane moves back and forth in response to the alternating pressures of sound waves, the nerve fibre from the ganglion cell of the scolophore transmits impulses to the central nervous system. Because the tympanic membrane is activated by the pressure of sound waves, this is a pressure type of ear.

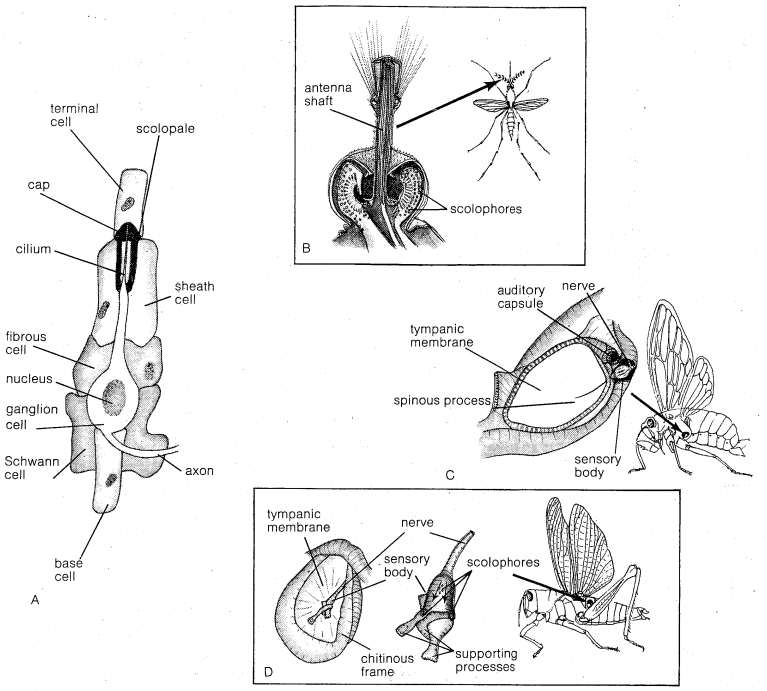


Figure 1: *Auditory mechanisms in insects.*
(A) A scolophore organ. (B) The mosquito ear. (C) The ear of the cicada *Magicicada septendecim.* (D) The ear of the grasshopper.

By courtesy of (B, right) the U.S. Department of Agriculture, and (B, left) *Acustica*, (1953) pages 335–343, and (C, right) H. Weber, *Grundriss der Insektenkunde* (4th ed. Ed. Gustav Fischer, Stuttgart, 1966), and (D, right) reprinted from Anna Botsford Comstock: *Handbook of Nature Study*, Copyright, 1939, by Comstock Publishing Company, Inc. Used by permission of Cornell University Press, and (D, left and centre) E.G. Gray, *Philosophical Transactions* B. 243, 77 (1960)

Simple tympanal organs, such as those found in moths, contain only two or four elements, or scolophores. In cicadas, on the other hand, these organs are highly developed; they include a sensory body (a number of scolophores in a capsule) that may contain as many as 1,500 elements. Shown in Figure 1 is the structure of the tympanal organ in the seventeen-year cicada (*Magicicada septendecim*).

With 80 to 100 scolophores, the grasshopper ear, which has been studied more thoroughly than any other insect ear, is structurally between that of moths and cicadas. Ordinarily, the tympanic membrane is hidden beneath the base of the insect's wing cover. A bundle of auditory nerve fibres runs from one side of the sensory body, which lies on the inner surface of the membrane (Figure 1), and joins other nerve fibres of the region to form a large nerve extending to a ganglion (nerve centre) in the thorax.

Evidence of hearing and communication in insects. *Behavioral observations.* That the insect ear serves an auditory purpose has been proved by a large number of experimental observations, particularly those that have dealt most extensively with katydids and crickets. Males of these groups produce sounds by stridulation, which usually involves rubbing the covers of the wings together in a particular way. One wing has a serrated surface (a "file") that runs along an enlarged vein; the other wing has a sharp edge over which the file is scraped. The scraping causes the wing surfaces to vibrate; the natural resonances of the vibrations and the particular rhythm and repetition rate of the scraping movements determine the nature of the song, which varies with each species. Most females are silent, but those of a few species have a poorly developed stridulatory apparatus, and weak sounds have been reported. Both males and females have tympanal organs for sound reception.

Stridulation

The observation that the males of many insect species produce repeated stridulatory sounds during the mating season led to the inference that the primary purpose of these noises was to attract a female. That this is indeed the case was first established by the extensive observations of the Yugoslavian entomologist Ivan Regen, who worked over the period 1902–30 mostly with a few species of katydids and crickets. In one of his earliest experiments, Regen proved (1913–14) that a male katydid of the species *Thamnotrizon apterus* responds to the sound of another male by chirping. The first male responds in turn to the second male's chirp, and the two insects then set up an alternating pattern of chirping. Although this pattern had been observed earlier, Regen was the first to prove by a series of well-designed experiments that it depends upon the sense of hearing. After removing the forelegs, on which the tympanal organs are located, of certain males, he found that even though these insects continued to stridulate, they did so only in purely individual rhythms that were not affected by the sounds of other males. Any alternation of chirping between deafened males, or between a deafened and a normal male, occurred only rarely, for brief times, and by chance.

A long series of check experiments by Regen showed that other stimuli, such as light, odours, and surface vibrations, did not affect the chirping behaviour. In these experiments the insects were placed in separate rooms, and their sounds were transmitted by telephone.

Chirping and mating behaviour

Further experiments carried out by Regen on field crickets (*Liogryllus campestris*) demonstrated the reactions of females to chirping males. In the most elaborate of these experiments, 1,600 sexually receptive females were released around the periphery of a large enclosed area in the middle of which had been placed a cage containing one or more chirping males. Precise data concerning the frequency with which the females moved toward the cage were obtained by surrounding the cage site with an array of traps in which the females were caught as they moved inward. The results were statistically significant. Normal females (those with intact tympanal organs) moved toward the cage and eventually reached it. The removal of one foreleg and its tympanal organ, however, caused difficulty; the movements were more random and the approaches fewer, although some females did succeed in reaching the cage. When both tympanal organs were removed or if the male failed to chirp, the performance of the females was reduced to chance. They also failed to exhibit the seeking performances if the male's stridulatory organ was modified, as by removing the file, so that little or no sound was produced.

In 1926 Regen returned to his study of the alternating chirping pattern of katydids and succeeded in having males react to an artificial sound, one that Regen himself produced. He also found that the alternation could be demonstrated with a suitably active male by using a variety of sounds—whistles, percussion noises, and sounds made with his mouth. It was never altogether clear, however, what changes Regen had made in his signals that finally brought success; probably the secret lay in the particular rhythm and timing of the signals. At any rate, this method made possible a study of the general nature of the auditory sensitivity of these insects and the range of sound frequencies to which they responded. It was shown that katydids are most sensitive to the very high frequencies, those that are beyond the limit of the human ear. The instruments available to Regen at the time, however, did not permit a precise measurement of intensity thresholds. (A threshold is the lowest point at which a particular stimulus will cause a response in an organism.)

Although the work of Regen and others established the basic character of sound reception in insects and its role in communication and mating, other details had to await the introduction of electrophysiological methods in this field as well as the development of electronic methods for the precise production, control, and measurement of sound stimuli.

Electrophysiological observations. When making electrophysiological observations of an auditory mechanism, an electrode (one terminal, generally a fine wire, in an electric circuit) is placed on a nerve or some other sensory structure in the mechanism. Sounds, presented at different frequencies and intensities, produce neural or sensory changes, which are actually electrical discharges or changes in electrical potential of extremely small magnitude. The impulses are picked up by the electrode and transmitted to an instrument with which they can be amplified, observed, and recorded. In both behavioral and electrophysiological observations, the auditory sensitivity of an animal to sounds of different frequencies can be illustrated by a curve.

The electrophysiological method was first used in research on the insect ear in 1933, with observations mainly on two katydid and one cricket species. The tympanal organ of these insects is located on one of the segments of the foreleg; its nerve goes to a ganglion in the thorax. When an electrode is placed on this nerve, its threshold sensitivity and overall frequency range can be determined by varying the intensity and frequency of the sounds applied to the tympanic membrane. It has been found that the tympanal organ of these insects responds poorly to low tones (those of low frequency) but improves rapidly as the frequency increases to a maximum sensitivity around 3,000 to 5,000 hertz. For higher frequencies the sensitivity declines, until a limit is reached at 30,000 hertz. It is likely that the insect's identification of its own species by means of song is primarily in terms of intensity and time patterns, with the rapid changes of intensity playing a prominent part. The possibility of frequency also entering into the pattern, however, cannot be ruled out.

A further question concerns the perception of the direction of a sound source. Clearly, if a female is to seek out and find a chirping male, the effectiveness of her performance depends upon an ability to localize the sound. Experiments indicate that the magnitude of electric responses from the tympanal nerve in katydids varies in a systematic manner when a given sound is presented at different angles while the distance is held constant. The insects continue to exhibit this directional pattern even after one of the tympanal organs has been removed. As was mentioned earlier, Regen found that female crickets deprived of one tympanal organ were still able to locate a chirping male, though less effectively than when both organs were intact.

HEARING IN SPIDERS

Whether spiders have a sense of hearing has long been debated. Early anecdotal observations concerning this matter have now been reinforced with both behavioral and electrophysiological evidence showing without doubt that spiders are sensitive to mechanical vibrations and also to aerial sounds. Whether this sensitivity should be regarded as hearing is considered later in this section, after a review of the anatomical and behavioral evidence.

Anatomical evidence. The bodies of spiders contain many slitlike openings, called lyriform organs, that have been considered as sensory in nature. Most of these organs probably have a kinesthetic function and thus provide information on local movements of body parts. There is one type of lyriform organ, however, that differs from the others in its location and in certain structural details. It is found on the metatarsal (next to last) segment of each of the eight legs, close to the joint that this segment makes with the tarsus (the last segment, or foot), and consists of a number of slits—about ten in the common house spider—that partially encircle the leg. Each slit contains a fluid chamber the inner wall of which is pierced by a tubule through which a thin filament runs to one of the two side walls (lamellae) that enclose the slit. This filament is evidently the termination of a ganglion cell that lies deeper in the leg. It has been suggested that an alternating compression of the lamellae stimulates the terminal filament.

Lyriform organs

The responsiveness of the common house spider to aerial sounds and mechanical vibrations includes a wide range, from below 20 to as high as 45,000 hertz. Within this range the sensitivity, as measured by electrical potentials, varies widely for aerial sounds; in some experiments narrow regions of frequency have been found in which no responses could be obtained at the highest intensities available. These variations of sensitivity are ascribed to mechanical resonances in the lyriform structure.

The tarsus evidently plays an important part in responses to sounds. Removal of portions of the tarsus reduces the responses about in proportion to the amount removed; immobilization of the tarsus greatly impairs the sensitivity. It appears, therefore, that the tarsus serves as a sensing element that transmits vibrations to the lyriform organ, which thus is a velocity type of ear.

Behavioral evidence. It has been reported that spiders react in characteristic ways to a buzzing insect caught in their web. The spider apparently locates the insect at once, runs to it, and attacks it. An inactive object, however, such as a small pebble enmeshed in the web, produces a different response: the spider manipulates the strands of the web, locates the object, and cuts away the filaments surrounding it so that the object drops to the ground. The reactions of a house spider to a mechanical vibrator applied to a point on the web have been observed. Such a stimulus elicits a response similar to that of an active insect if the vibratory frequency is between 400 and 700 hertz. For frequencies above 1,000 hertz, however, the spider reacts either by running to a secluded corner of the web or, if the intensity is too great, by abandoning the web altogether. From this and similar evidence it has been concluded that the spider has the ability of pitch (tone) discrimination between low and high ranges and perhaps can distinguish between tones of the lower range.

Spiders also react to aerial tones from an artificial source, such as a loudspeaker. These stimuli elicit an orientation response, in which the spider faces the source and reaches out with the two front legs. Thus, in view of the high level of sensitivity to both aerial and mechanical stimuli, the reception of sounds in the spider can probably be regarded as true hearing, and the lyriform organ as a form of ear. It is evidently a velocity type of ear, for there is no tympanic surface to respond to sound pressures, and the small leg segments seem to respond to the oscillatory motions of the air particles.

II. Sound reception in vertebrates

The ear of vertebrates appears to have followed more than one line of evolutionary development, but always from the same basic type of mechanoreceptor, the labyrinth. All vertebrates have two labyrinths that lie deep in the side of the head, adjacent to the brain. They contain a

The labyrinth

number of sensory endings the primary functions of which are to regulate muscle tonus (a state of partial muscular contraction) and to determine the position and movements of the head and body.

Generalized sketches of vertebrate labyrinths are shown in Figure 2, with the usual locations of the sensory endings indicated for the different vertebrate classes. Two

From *Biological Review* (1936)

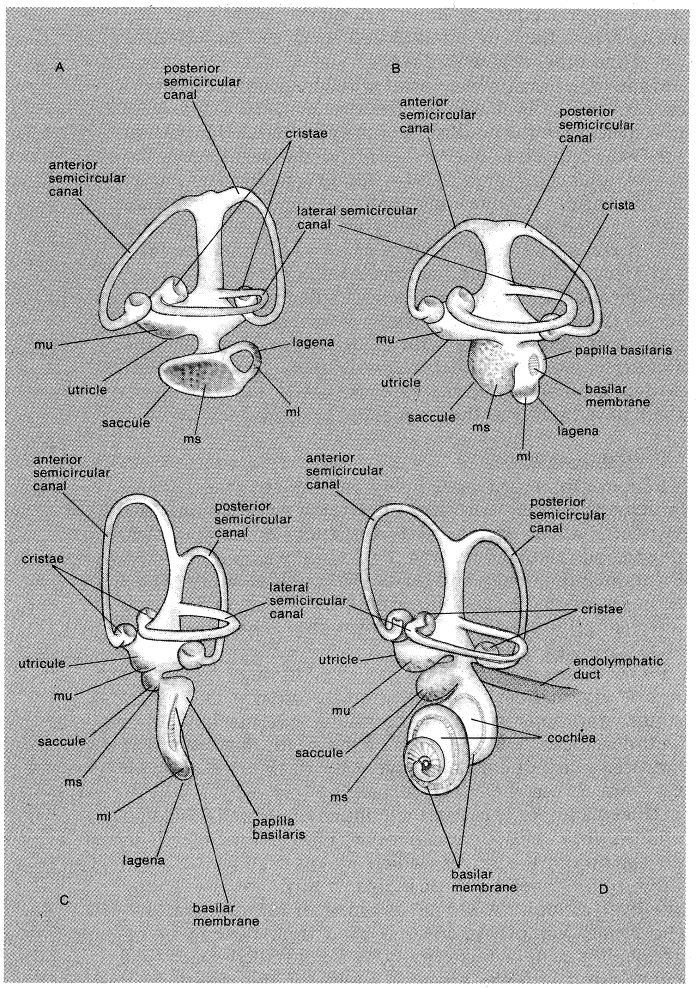


Figure 2: Generalized labyrinth of (A) fish, (B) turtle, (C) bird, and (D) mammal. (Abbreviations: mu, macula utriculi; ms, macula sacculi; ml, macula lagenae.)

main divisions of these endings are distinguished: a superior division, which includes the three semicircular canals, the organs associated with the sense of balance, and the utricle, a small sac into which the semicircular canals open; and an inferior division, which includes the saccule (also a small sac) and its derivatives. Arising at or near the connection between the utricle and the saccule is the endolymphatic duct, which ends in an endolymphatic sac; this structure probably regulates fluid pressures in the labyrinth and aids in the disposal of waste materials.

The superior division of the labyrinth (Figure 2) is remarkably constant in form throughout the vertebrates except in the cyclostomes (*e.g.*, hagfishes and lampreys), in which the canals and endings are reduced in number. The utricle contains a macular ending, the macula utriculi, and each semicircular canal ends in a crista. In all vertebrate classes except the placental mammals and a few other scattered species, a papilla neglecta is present. It is usually located on the floor of the utricle or near the junction of the utricle and the saccule.

The inferior division of the labyrinth always contains a saccule with its macula, the macula sacculi, but the derivatives of the saccule vary greatly in the different vertebrate classes. In teleosts (bony fishes), amphibians, reptiles, and birds there is a lagena (a curved, flask-shaped structure), with its macula, the macula lagenae. Only the amphibians have a papilla amphibiorum, which is located near the junction of the utricle and the saccule. In some amphibians and in all reptiles, birds, and mammals, there is a papilla basilaris, which is usually called a cochlea in the higher forms, in which it is highly detailed. The elaborate sensory structure of higher types of ears, containing hair cells and supporting elements, is called the organ of Corti.

The macular endings consist of plates of ciliated cells (cells with short, hairlike projections) along with accessory cells, all surmounted by an otolith (a calcareous mass containing numerous particles of calcium carbonate embedded in a gelatinous matrix) or, in teleosts, by one large mass of calcium carbonate. The crista endings contain moundlike groups of sensory cells with supporting cells; the sensory cells have elongated cilia that are embedded in a gelatinous body, the cupula, which forms a sort of valve across an expanded portion of each semicircular canal. The papillae contain plates or ribbons of ciliated cells in a structural framework that lies on a movable membrane, except in amphibians, in which the papillae are on a solid base. These ciliated cells are not surmounted by an otolithic mass or a cupula, but some of the cilia are attached either directly or indirectly to a tectorial membrane (a membrane with one edge fixed to a stationary base, thus anchoring the cilia) or to an inertia body (a mass lying over the ciliated cells and restraining the movements of the cilia).

Functions of labyrinthine endings

The endings have different functions: the macular organs serve primarily as gravity receptors and detectors of sudden movements; the crista organs serve for the perception of rotational acceleration; and the papillae serve for hearing. As structural relations suggest, the auditory endings are derived either from the other labyrinthine receptors or from the primitive labyrinthine epithelium.

HEARING IN FISHES

The cyclostomes and the elasmobranchs (*e.g.*, sharks and rays) possess a labyrinth with maculae and cristae but have no auditory papillae. There are, nevertheless, two possible ways by which some of these cartilaginous fishes, especially the sharks, react to sounds in the water: by means of the macular organs and by means of the lateral-line apparatus. (For details on these mechanisms, see MECHANORECEPTION: *Reception of external mechanical stimuli.*) It is in the bony fishes (teleosts) that a true ear whose function is hearing first appears among the vertebrates. This ear, which occurs in a number of forms, has varying degrees of effectiveness as a sound receiver; some fishes hear well, others poorly. The differences arise, at least in part, from the accessory mechanisms that aid in the utilization of sound energy.

The basic auditory mechanisms in teleosts. In most fishes, especially in many marine forms, the auditory mechanism is relatively simple, consisting of macular endings that evidently have been diverted from their primitive functions as detectors of gravity and motion. The important change is not in the structure of the end organ but in its innervation—the nerve supply has connections that transmit auditory information. It is thought that in most teleosts the change to an auditory function has occurred in the saccular macula, and probably the lagenar macula as well, and that the utricular macula continues as a receptor for gravity and motion.

Inertia principle in the teleost ear

The simple macular ending of the teleost ear is stimulated by sound through the operation of an inertia principle. Sound waves pass readily through the water and into the body of the fish, causing most of the tissues to vibrate in a uniform manner. The macular otolith, however, represents a discontinuity; because its density is greater than that of the other tissues, it exhibits an inertia effect (resistance to movement). Its motions not only lag behind those of the surrounding tissues but are probably of lesser amplitude as well. Accordingly, a sound creates a relative motion between the otoliths and the other tissues. More specifically, there is relative motion between the bodies of the hair cells, which rest on a tissue base, and the cilia of these cells, the tips of which are in

contact with the otolith. This method of stimulating the auditory hair cells is inefficient, however, because of the relatively small difference in density between the body tissues and the otoliths.

Special stimulation mechanisms. In certain groups of teleosts the efficiency of hair-cell stimulation has been increased by a discontinuity that is nearly 1,000 times greater than the one between tissue and otolith; this is the discontinuity between the otolith and a gas bubble. Although there are varying anatomical methods of achieving it, the simplest arrangement, which is found in clupeids, mormyrids, labyrinthine fishes, and a few others, consists of a gas-filled sac that lies against one wall of the labyrinth. In clupeids (*e.g.*, herring), a group in which the utricular macula rather than the saccular or lagenar maculae has an auditory function, long anterior extensions of the swim bladder form air sacs, one adjacent to each utricular macula. In the mormyrids, which include the elephant-nosed fish, a similar condition exists in early life; during adult development, however, the connections with the swim bladder disappear, leaving the air sacs connected with the saccular and lagenar endings. The gas content of these sacs is then maintained by special glands that extract gas from the blood. Air sacs arise in various other ways.

Air sacs

One large group of fishes, referred to as the Ostariophysi (*e.g.*, catfishes, minnows, and carps), has no air sac adjacent to the labyrinth, but a possibly equivalent condition is achieved through a mechanical connection between the swim bladder and fluid chambers adjacent to the labyrinth. A chain of three or four small bones, known as the Weberian ossicles, extends from the anterior wall of a part of the swim bladder to a fluid-filled chamber called the atrium, which in turn connects by fluid passages with the two labyrinths in the region of the saccule-lagena complex (Figure 3). In this arrangement

Adapted from *Copeia* (1943) and (bottom) *Nature* (1938)

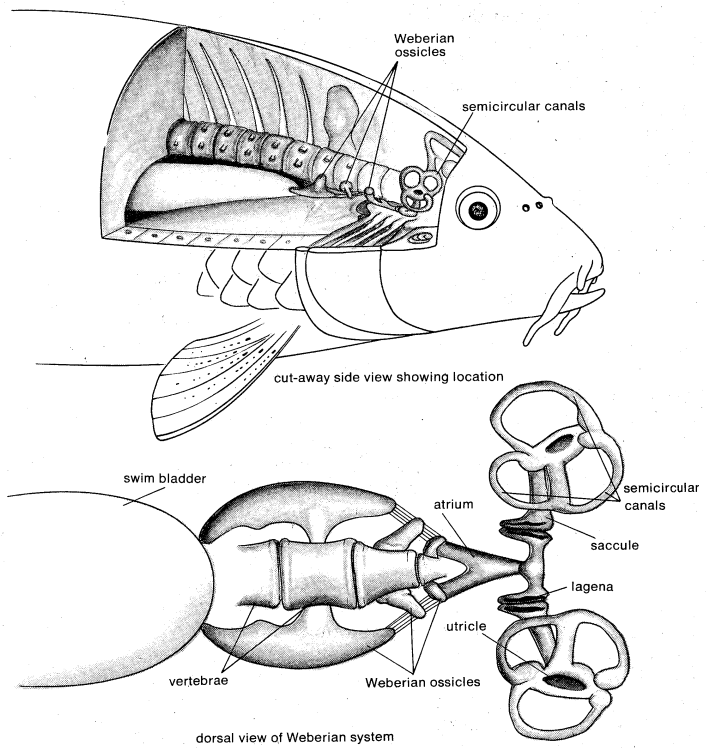


Figure 3: The Weberian ossicles and labyrinths of an ostariophysan fish.

the discontinuity is between the air of the swim bladder and the chain of ossicles in contact with it; the relative motion arising from sound stimulation is communicated through the ossicular (bony) chain and the fluid channels to the macular endings.

Regardless of the mechanism employed, however, the ear of all teleost fishes is basically a macular organ. Because it is stimulated by sound that is transmitted to tissues adjacent to the sensory cells and that acts differentially on these cells, this ear is of the velocity type.

Auditory sensitivity of fishes. Although only limited experimental data are available, it appears certain that, in general, fishes with the accessory mechanisms described above have greater sensitivity and a higher frequency range than do those lacking such mechanisms; while upper frequency limits are about 1,000 hertz for many fishes, they are about 3,000 hertz for the Ostariophysi and other specialized types.

Many experiments have dealt with the problem of auditory sensitivity in fishes, but the species most extensively tested has been the goldfish, a variety of carp belonging to the Ostariophysi. In one well-controlled investigation, the sound intensities required to inhibit respiratory movements, after conditioning with electric shock, were studied. The greatest sensitivity was found to be around 350 hertz; above 1,000 hertz sensitivity declined rapidly.

In view of the simple anatomical character of the ear, the question of whether fishes can distinguish between tones of different frequencies is of special interest. Two studies dealing with this problem have shown that the frequency change just detectable is about four cycles for a tone of 50 hertz and increases regularly, slowly at first, then more rapidly as the frequency is raised.

HEARING IN AMPHIBIANS

There are three orders of living amphibians: the Apoda, which are legless, wormlike types such as caecilians; the Urodela, which are tailed forms such as mudpuppies, newts, and salamanders; and the Anura, which are tailless forms including frogs and toads. Although members of all three orders have ears, the structures vary greatly in the different groups, and little is known about them except in such advanced types as frogs.

The auditory mechanism in frogs. Although the frog has no external ear (structures on the outside that direct sound vibrations inward), the middle-ear mechanism is well developed. On each side of the head, flush with the surface, a disk of cartilage covered with skin serves as an eardrum. From the inner surface of this disk, a rod of cartilage and bone, called the columella, extends through an air-filled cavity to the inner ear. The columella ends in an expansion, the stapes, which makes contact with the fluids of the inner-ear (otic) capsule through an opening, the oval window. A second opening in the otic capsule, the round window, is covered by a thin, flexible membrane; it is bounded externally by a fluid-filled space that can expand into the air-filled cavity of the middle ear. When the alternating pressures of sound waves cause the eardrum to vibrate, the vibrations are transmitted along the columella and through the oval window to the inner ear, where they are relayed to the round window in a path across the otic capsule by movements of the inner-ear fluids. Along this path are two auditory endings, the amphibian and basilar papillae, the sensory hair cells of which are stimulated by the fluid movements. These movements are transmitted to the ciliary tufts of the sensory cells by a tectorial membrane, which is suspended from the hair cells in such a way that it can be moved by the oscillations of the inner-ear fluids.

As sense organs for hearing, the papillae, which appear for the first time in amphibians, have cells like those in lower vertebrates that serve the same purpose. There are two types of papillae: the amphibian papilla, which is found in all amphibians, and the basilar papilla, which is found in some amphibians. Because they are located in different places in the inner ear, the papillae probably represent two distinct evolutionary developments. Moreover, they operate on a mechanical principle found in no other animal group: a tectorial membrane, moving in response to sound vibrations that have been transmitted to it by the inner-ear fluids, stimulates the sensory hair cells directly through connections to the cilia of these cells. In all higher types of ears, on the other hand, the sensory cells themselves are set in motion by the sound vibrations, while the tips of the ciliary tufts are restrained in one of several ways.

Papillae

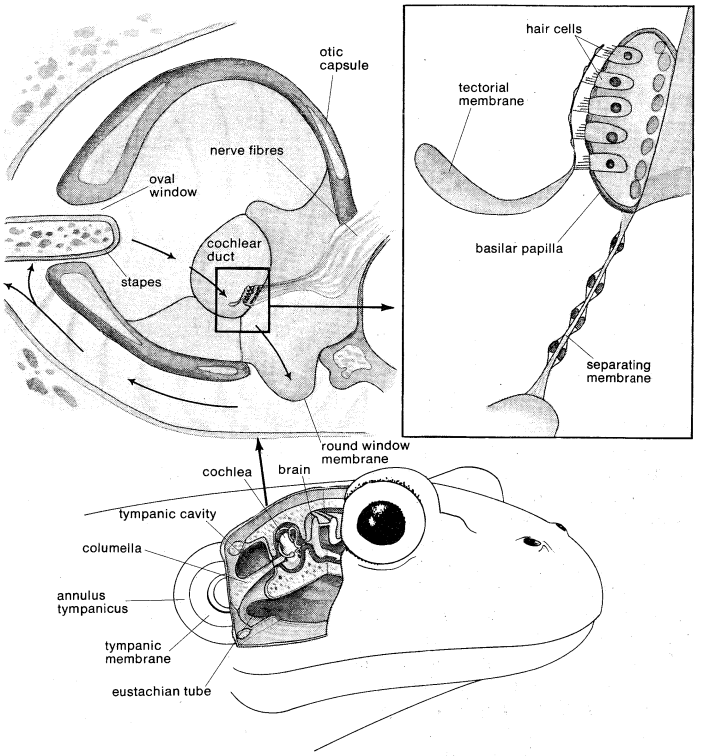


Figure 4: The inner ear of a frog, showing one of the papillae. The arrows indicate the path along which sound vibrations are transmitted by movements of the inner-ear fluids. At the right is an enlarged view of the separating membrane and the basilar papilla.
From (bottom) H.W. Rand, *The Chordates*, copyright 1950; used with permission of McGraw-Hill Book Co.

Auditory sensitivity of amphibians. Although it is presumed that all amphibians possess hearing of some kind, the evidence is sparse; only salamanders other than anurans have been studied experimentally. Salamanders trained to come for food at the sound of a tone responded only at low frequencies, up to 244 hertz in one specimen and to 218 hertz in three others.

Frogs, which are of special interest because they first live in the water as tadpoles and then undergo a metamorphosis that equips them for life on land, have been studied more extensively. Considerable modifications of the middle-ear mechanism occur during metamorphosis. Presumably, the tadpole larva has an aquatic ear that is later transformed into an aerial type.

Croaking

Interest in the hearing of adult frogs has been stimulated by their active and often loud croaking during the breeding season. Evidently, their vocalizations assist in the location and selection of mates. The first experimental study of auditory sensitivity in frogs, carried out in 1905, showed that leg movements in response to strong tactual stimuli may be enhanced or even inhibited by sounds.

Somewhat later, following some unsuccessful attempts to train frogs to make behavioral responses to acoustic stimuli, two other methods were employed to determine the sensitivity and range of their hearing. One of these was the recording of changes in the electrical potentials of the inner ear and auditory nerve; the other was the observation of changes in the potentials of the skin (electrodermal responses) to acoustic stimuli. As a result of these investigations, inner-ear potentials and electrodermal responses in the bullfrog have been recorded over a range from 100 to 3,500 hertz. In the treefrog, these same responses have been found in a range that extended from 50 to 3,000 hertz, with the greatest sensitivity from 600 to 800 hertz, and again at 2,000 hertz.

The recording of impulses from single fibres in the auditory nerve of bullfrogs and the green frog indicates that two types of auditory nerve fibres are present. This has led to the suggestion that they represent the different characteristics of the amphibian and basilar papillae. It is believed that the amphibian papilla is more sensitive to low tones and that the basilar papilla is more sensitive to high tones.

HEARING IN REPTILES

The living reptiles belong to four orders: the Squamata (lizards, snakes, and amphisbaenians), the Rhynchocephalia (one rare species, the tuatara of New Zealand), the Chelonia (turtles), and the Crocodylia (crocodiles and alligators). The reptile ear has many different forms, especially within the suborder Sauria (lizards), and variations occur in all elements of its structure—the external ear is often absent or may consist of an auditory meatus (passage) of varying length; the middle ear shows several forms in the different groups; and the inner ear varies in the degree of development of the auditory papilla and also in the ways by which the sensory cells are stimulated by sound.

Lizards. *Auditory structure.* There are about 20 families of lizards, ranging from the chameleon, a divergent type, to the gecko, certain species of which have the most highly developed ears found in the group. The chameleons, of those species studied thus far, have only a few sensory hair cells (40 to 50) in the auditory papilla. The geckos, on the other hand, have several hundred hair cells, and the *Gekko gecko* has about 1,600, the largest known number of hair cells in any saurian. Other lizard species fall between these two extremes in inner-ear development, with the iguanids, the most common lizards in the Western Hemisphere, having from 60 to 200 hair cells, according to the species.

Middle ear of lizards

What may be regarded as the standard type of middle-ear structure in the lizards consists of a tympanic membrane and a two-element ossicular chain that extends from the inner surface of this membrane to the oval window of the otic capsule. As shown in Figure 5, the

From *Journal of Auditory Research* (1965)

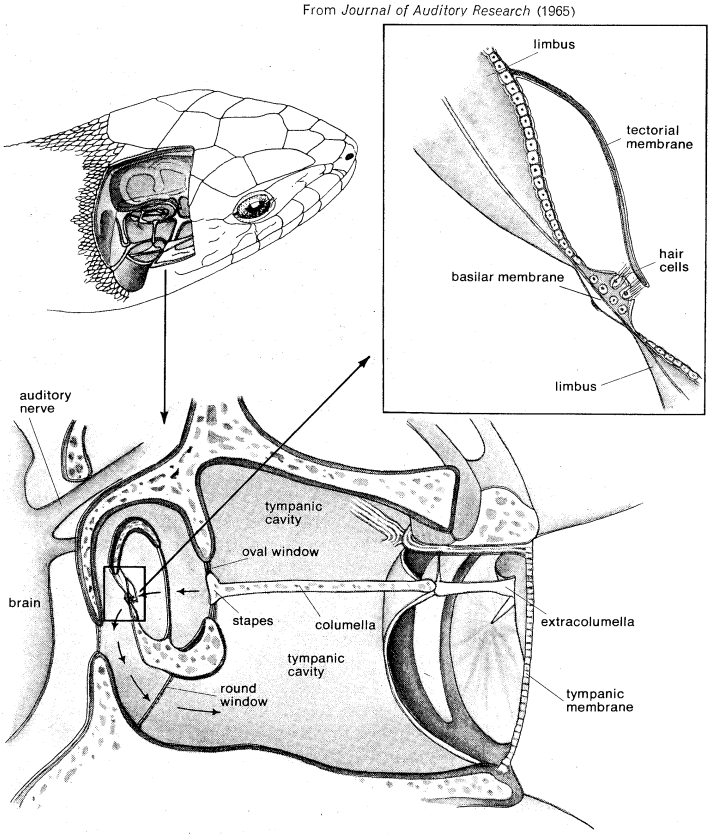


Figure 5: The auditory structures in the right lizard ear as seen from behind and from the left (see text).

ossicular chain is made up of two parts: the osseous (bony) columella, whose expanded innermost end (the stapes) fills the oval window, and the extracolumella, a cartilaginous extension that usually spreads out in two to four processes that are embedded in the fibrous layer of

the tympanic membrane. Geckos have a single middle-ear muscle attached to the lateral part of the extracolumella; evidently, contractions of this muscle stiffen the extracolumella, thereby dampening the ossicular motions and protecting the ear against excessively intense sounds.

The auditory part (cochlea) of the inner ear (Figure 5) consists of a basilar membrane lying in an opening in the limbus, which is a plate of connective tissue. The form of the basilar membrane, which is unlike the structure of the same name in amphibians and is clearly of different origin, varies from a simple oval in iguanids to a long, tapered ribbon in gekkonids. In many species the middle portion of the basilar membrane is greatly thickened, especially in some regions of the cochlea. Over this thickening, which is called the fundus, lies the auditory papilla proper—*i.e.*, that part of the cochlea in which the sensory hair cells are held in a framework of supporting tissues and cells. The hair cells usually occur in regular transverse rows, with the number of cells in a row varying along the cochlea. They have a tuft of cilia, the so-called sensory hairs, of graduated lengths, the longest of which are usually attached either directly or indirectly to a tectorial membrane. This membrane arises from a region of the limbus that is usually elevated, often strikingly so, and runs as a thin web or sheet to the region of the hair cells. Only rarely does the free edge of the tectorial membrane connect directly with the cilia of the hair cells; usually there are intermediate connecting structures that take a variety of forms, from simple fibres to relatively massive plates.

The function of the tectorial membrane and its connections to the ciliary tuft of a hair cell is to immobilize the tuft when the body of the hair cell moves in unison with the basilar membrane on which it rests. This produces a relative motion between the ciliary tuft and the body of the cell and stimulates the cell. All auditory stimulation depends ultimately upon this relative motion, and the means just described for achieving it can be regarded as the most fundamental process by which sounds are perceived. Although it is employed in the great majority of ears, it is not the only mode of stimulation. Another mode is that in the ears of fishes, in which an otolith lies upon the ciliary tufts and, by its inertia, reduces and alters the motion of the tuft relative to the cell body. Still another method is the one in the frog papilla, in which the tectorial membrane is moved by the cochlear fluids while the body of the sensory cell remains at rest.

Modes of auditory stimulation

In some lizards the inertia principle has a form different from that found in fishes. In the former, a body called a sallet lies upon the ciliary tufts of a group of hair cells and, by its inertia (or by an equivalent means), restrains the movement of the cilia when the cell body is made to move. The result is a relative motion and a stimulation of the hair cells, like the more common restraint by a tectorial membrane.

The ears of two lizard families show only the inertial restraint method of stimulation; in several other families this method functions in some regions of the cochlea for certain hair cells. Hair-cell stimulation by two or more different arrangements within the same cochlea, however, is the rule rather than the exception because of its many advantages. Although the tectorial-restraint method provides great sensitivity for individual cells, the sallet system also attains good sensitivity, but in another way: by causing many cells—those in common contact with a given sallet—to work in parallel, thus producing a spatial summation. The sallet system has the advantage of being more resistant to damage by overstimulation from intense sounds. In such lizards as the geckos, for example, in which the hair cells are divided nearly equally between tectorial and sallet systems, an exposure to excessive sound has been observed to break all the tectorial connections to the hair cells while leaving the sallet connections intact. But even though the most sensitive hair cells are inoperative, the animal can respond to sounds, although with lesser acuity.

Hearing abilities of lizards. The lizards are the lowest vertebrates to have a well-developed spatial differentiation of the cochlea in which different regions respond to different frequencies of tone. The problem of tonal discrimination has been somewhat solved in frogs, in which the differential responses to tones by the two papillae may provide some information concerning the pitch of sounds. The mechanism in frogs, however, is a poor one, as it can give only crude and uncertain cues at best.

Tonal discrimination

In some lizards, such as iguanids and agamids, a minimum of structural variation occurs along the cochlea; in others (*e.g.*, geckos, which have very extensive differentiation along their extended basil membranes) the differentiation is almost as great as that in higher vertebrates, including man. Most geckos are nocturnal in habit and use vocalizations to maintain individual territories and probably to find mates.

Although it has been possible to train two species of lizards (*Lacerta agilis* and *Lacerta vivipara*) to make feeding movements in response to a variety of sounds, including tones between 69 and 8,200 hertz, most attempts to train lizards to respond reliably to tonal stimuli have failed. The one useful method thus far developed to study the sensitivity of these animals to sounds involves recording electrical responses in the ear and in the auditory nervous system. Although such observations have provided information about peripheral response to sounds, they do not reveal anything about other processes in the nervous and behavioral systems.

Electrical responses in the cochlea of many lizard ears show considerable variations: in absolute sensitivity, in the tonal regions in which responsiveness is best, and in the extent of the frequency range. It has been concluded that most lizards have good auditory sensitivity over a range from 100 to 4,000 hertz and relatively poor hearing for lower and higher tones. This auditory range is not very different from that of man, although somewhat more restricted than that of most mammals.

Snakes. Without much doubt, snakes developed from some types of early lizards but lost their legs when they adopted habits of burrowing in the ground. Although some snakes burrow, others have taken up different habits: many species live on the surface of the ground, several are largely aquatic, and some live in trees. All, however, show drastic ear modifications that reflect their early history as burrowers; for example, there is no external ear—*i.e.*, no opening at the surface of the head for the entrance of sound. This fact, together with a seeming indifference to sounds, has led to the supposition that snakes are deaf or that they can perceive only such vibrations as reach them through the ground on which they crawl.

Auditory modifications of snakes

This supposition is incorrect; snakes can hear and are able to receive aerial sounds through a mechanism that serves as a substitute for the tympanic membrane. This mechanism consists of a thin plate of bone (the quadrate bone) that was once a part of the skull but that has become largely detached and is held loosely in place by ligaments. It lies beneath the surface of the face, covered by skin and muscle, and acts as a receiving surface for sound pressures. The columella, attached to the inner surface of the quadrate bone, conducts the received vibrations to its expanded inner end, which lies in the oval window of the cochlea. If the columella is severed, the sensitivity of the ear is significantly reduced.

Although the sensitivity of the snake ear varies with the species, it is appreciably sensitive only to tones in the low-frequency range, usually those in the region of 100 to 700 hertz. For this low range the large mass of the conducting mechanism and the presence of tissues lying over the quadrate bone are not of any great consequence. Moreover, while the sensitivity of most snakes to the middle of the low-tone range is below that of most other types of ears, it is not seriously so. In a few snakes, however, the sensitivity is about as keen as in the majority of lizards with conventional types of ear openings and middle-ear mechanisms.

That the ears of the snake receive aerial sound waves instead of depending on vibrations transferred from the ground has been proved by recording the potentials in the cochlea of one ear while rotating the animal in front

of a sound-wave source so that the ear being studied was sometimes facing the source and sometimes directed away from it. The recorded potentials were significantly greater when the ear was facing the source. There would have been no difference in the responses if the sound first set up vibrations in the ground and these were then transmitted to the body. This observation also shows that the ears of the snake can determine the direction of a sound in terms of its relative intensity in the two ears. Although snakes can perceive vibrations from the ground that are present at a sufficient intensity, this ability is not peculiar to them; all ears respond to vibrations transmitted to the head.

Amphisbaenians. The amphisbaenians form a little-known group of reptiles. Because they are burrowers and live almost entirely underground, they are seldom seen. The one species in the United States, *Rhineura floridana*, is found in some parts of Florida; a number of species occur in other regions of the world, especially in South America and Africa.

The animals construct a maze of underground tunnels, which they patrol in search of such food as grubs and worms. Although small eyes below the body surface can receive light through a transparent scale, amphisbaenians evidently make little use of vision. There is reason to believe, however, that they use hearing to locate their prey.

Auditory structures in the face of amphisbaenians

Amphisbaenians, like snakes, have no surface indication of an ear; a receptive mechanism below the surface and different from that in snakes conveys vibrations to the inner ear. In the oval window, which occupies the entire lateral surface of the otic capsule, is a stapes. The head of the stapes in most species is directed laterally and forward; it is united by a joint with a rod of cartilage (the extracolumella) that extends forward along the face, in the line of the lower jaw. The extracolumella lies below the surface, where it makes close contact with and finally enters a dense layer of the skin. When the facial region is exposed to sounds, the vibrations are transmitted through the dense layer of the skin to the extracolumellar rod and then through it to the stapes, finally reaching the fluid of the inner ear. That this is the route of sound conduction has been proved by cutting the extracolumella at different places and observing the reduction of recorded responses in the ear.

The auditory mechanism of amphisbaenians varies somewhat according to species but is substantially as described above. The sensitivity, which also varies with species, is surprisingly high in some species, considering the unusual nature of the mechanism involved. Studies similar to those described for snakes have proved that this ear receives aerial sounds and that it can determine the direction from which the sound originated. As expected, this ear also responds to mechanical vibrations communicated directly to the skull.

Turtles. It is sometimes supposed that the turtle's ear is a degenerate organ, largely or even completely unresponsive to sound. Although the turtle's ear is unusual in some respects, and can be regarded as specialized in its manner of receiving and utilizing sounds, it is not a degenerate organ. There is good evidence that turtles are able to hear and that some species have excellent acuity in the low-frequency range.

A plate of cartilage on each side of the head serves as a tympanic membrane. Leading inward from the middle of this plate is a two-element ossicular chain consisting of a peripheral extracolumella and a medial columella the expanded end (the stapes) of which lies in the oval window of the otic capsule. Within the otic capsule are the usual labyrinthine endings, including an auditory papilla. As shown in Figure 6, the auditory papilla lies in a path between the oval window and an opening (the round window) in the posterior wall of the otic capsule. Unlike the round window in most ears, that in turtles has no membranous covering for transmitting pressure changes to the air-filled cavity of the middle ear. Instead, the opening leads to a fluid-filled chamber, the pericapsular recess, that extends laterally and anteriorly to enclose the external portion of the stapedial expansion of the columella. A pericapsular membrane separates the perilymph (fluid) of the otic capsule from the fluid of the recess. When the stapes is moved inward by the columella at one phase of a sound vibration, the fluid of the otic capsule is displaced, causing a pressure change that, after passing through the sac containing the auditory endings, continues in a circuitous course to the external surface of the stapes. This circuit is indicated by the arrows in Figure 6. When the columella moves outward, the fluid circuit reverses itself. Hence the result of a continuous sound wave is a surging back and forth of the fluids in the otic capsule and the pericapsular recess at the same frequency as that of the sound.

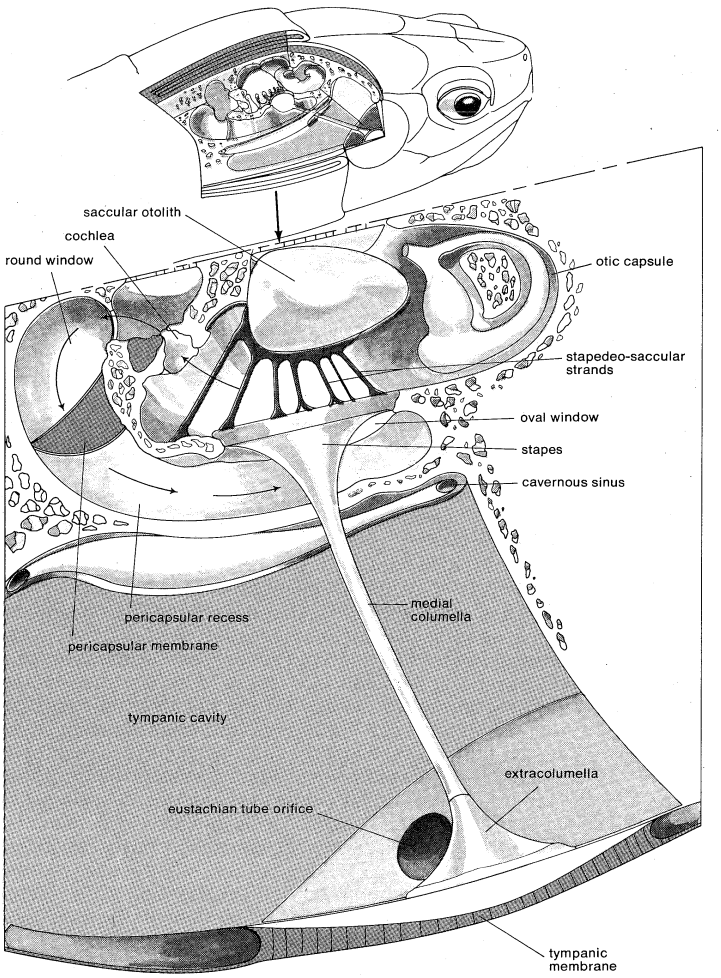


Figure 6: *Structure of the turtle ear.*
(Top) View of the head, showing the location of the otic capsule; (bottom) detail of the otic capsule. Arrows indicate the path along which vibrations are transmitted by movements of the cochlear fluids.
From *Proceedings of the National Academy of Sciences* (1956)

The special mechanical arrangement in the turtle ear is fully effective within the low-frequency range. Indeed, the relatively large mass of tissue and fluid involved in the response to sounds is in part responsible for the efficiency of the ear at low frequencies and also for the rapid loss of sensitivity as frequency increases.

This type of cochlear response to sounds is not peculiar to turtles; it is also found in snakes, through a structural arrangement of similar form. Although it also occurs in amphisbaenids, the fluid path in these animals is entirely different: it proceeds through the perilymphatic recess into the brain cavity and then by an anterior passage across the head to the lateral surface of the stapes.

Certain experiments involving the turtle's sensitivity to sounds have used training methods (conditioned responses); only a few have met with success. It has been found that turtles of the species *Pseudemys scripta*, trained to withdraw their head, respond to sound over the

low-frequency range, with the greatest sensitivity in the region of 200 to 640 hertz. This result is in close agreement with electrophysiological observations in which it has been found that impulses could be obtained from the auditory nerve of *Chrysemys picta* for tones between 100 and 1,200 hertz, with highest sensitivity for tones below 500 hertz. Similar results have been obtained by additional observations of this kind with several other species of turtles, some of which are very sensitive to a narrow band of frequencies in the low-tone range. Evidently, the type of receptor mechanism in the turtle can achieve great sensitivity through mechanical resonance at a particular region of the low-frequency scale.

Low-frequency sensitivity in turtles

Evidence has also been obtained that these responses are to aerial waves and not to vibrations set up in the ground. The sensitivity to surface vibrations was considerably poorer than that to aerial sounds. In addition, cutting the columella seriously impaired the responses to aerial sounds but hardly affected responses to mechanical vibrations applied to the turtle's shell.

Crocodiles. The order Crocodylia includes four groups of closely related forms: crocodiles, alligators, caimans, and gavials. The crocodile ear, although clearly reptilian in general structure, has a number of peculiar features. Leading to a tympanic membrane on each side of the head is a shallow external passage the outside opening of which is protected by an earlid that is closed when the animal enters the water and dives. Beyond the tympanic membrane is a middle-ear cavity, with the one on the right connected to the one on the left by an air passage that runs across the head above the brain. A sound presented to one ear, therefore, reaches the other ear about equally well. A columellar system connects the tympanic membrane to the oval window of the otic capsule, as in other reptiles. The inner ear is highly developed and bears many similarities to the cochlea of birds, described in the next section. Elongated and slightly curved, the cochlea contains about 11,000 sensory hair cells, about seven times as many as found in that of the most advanced lizard (*Gekko gecko*).

In comparison to some lizards, the cochlea of *Caiman crocodilus*, which has been most extensively studied, exhibits only a moderate degree of structural differentiation. Yet in this cochlea fibre bundles that extend from the root portion of the tectorial membrane separate into fine fibres that form individual connections with the ciliary tuft of each hair cell. This arrangement is not a common one, though present in certain lizards, such as the chameleons, and also in some degree in birds. It probably provides a high level of specificity in the stimulation process or as much specificity as the overall mechanical pattern permits.

The hearing of crocodilians has not been studied very extensively. It has been noted that the breathing rate in a crocodile accelerates in response to loud sounds, such as the firing of a gun, and it has been observed that specimens of the Mississippi River alligator produce vocalizations of roaring or hissing when low-frequency sounds are made by blowing a horn or by plucking a metal rod. Studies of the electrical potentials in the ear of *Caiman crocodilus* show that it is sensitive to frequencies ranging from 20 to 15,000 hertz.

HEARING IN BIRDS

Ears of birds show considerable uniformity in general structure.

The avian auditory structure. The ears of birds are similar in many respects to those of reptiles. The outer ear consists of a short external passage, or meatus, ordinarily hidden under the feathers at the side of the head. Most birds have a muscle in the skin around the meatus that can partially or completely close the opening.

The tympanic membrane of birds

The tympanic membrane bulges outward as in most lizards. In the songbirds, however, it consists of two separate membranes, with the outer one apparently serving to protect the inner one from injury. From the inner surface of the tympanic membrane an ossicular chain transmits vibrations of the cochlea. As in lizards, the chain consists of an osseous inner element, the columella, and a cartilaginous extracolumella that extends the columella peripherally and connects with the tympanic membrane.

The cochlea of birds is similar to that of crocodiles, consisting of a short, slightly curved bony tube within which lies the basilar membrane with its sensory structures. The length of the basilar membrane varies between 2.5 and 4.5 millimetres (0.1 and 0.2 inch) in most birds, but in the owls it may reach ten millimetres (0.4 inch) or more. At the end of the cochlea is another ending with a different function, the lagena and its macula.

Auditory sensitivity in birds. Using the conditioned-response method to study auditory sensitivity in a small songbird, the bullfinch, responses over a frequency range from 100 to 12,800 hertz have been observed. The electrophysiological method was first applied to the study of hearing in birds in 1936. In this study impulses from the cochlea of pigeons were recorded for tones usually up to 10,000 hertz and occasionally as high as 11,500 hertz. Although this method has been used by many investigators since 1936, few detailed and quantitative results have been obtained; nevertheless, one striking characteristic revealed by these studies has been the high degree of sensitivity in the low and middle range and the very rapid decrease in the high tones.

Uses of hearing in birds. Like other animals, birds use hearing to warn them of enemies and other kinds of danger. To a degree hardly equalled in lower species, they also use hearing in social relations and communication. Many male birds sing to hold their territories and to attract mates. Some birds also use vocalizations to identify their mates or group members. During the breeding period of the emperor penguin, for example, the male leaves his mate for a journey taking many days in order to obtain food. Upon returning to the general area where his mate has remained with a pack of hundreds of birds, the male is able to locate and to recognize his partner by an interchange of calls.

There is good reason to believe that certain birds, including the swiftlets (*Collocalia*) of Asia and Australia, the oilbirds (*Steatornis*) of tropical America, and possibly a few others, are able to use echolocation when flying in the dark caves that they inhabit. Moreover, it is well established that many owls locate and catch their prey by auditory cues. On a dark night, an owl perched in a tree can hear the rustling sounds made by a mouse in the grass and leaves on the ground below; by accurately localizing this signal, he can make his strike and capture the prey without any visual aid. This performance has been reproduced in a completely dark laboratory.

HEARING IN MAMMALS

Auditory structure of mammals. In the mammals the ear reaches its highest level of development, with well-differentiated divisions of outer ear, middle ear, and inner ear. Except in some of the sea mammals, in which certain modifications and degenerations have taken place, these structures carry out their functions in a remarkably regular manner.

The pinna

The outer ear consists of pinna (or auricle) located behind the ear opening and partially enclosing it and an auditory meatus that leads inward. The pinna varies greatly in size relative to the size of the animal, being large enough in many species to serve a useful purpose in the collection and reflection of sounds. Many mammals can move the pinna back and forth to regulate in some degree the entrance of sounds to the auditory meatus, which transmits the sounds inward to the tympanic membranes. In some mammals, such as many of the marine types, the external opening can be closed to keep out water when the animal dives, and in certain species of bats the tube itself contains a valve that can be closed to protect the ear against undesirable sounds.

The middle ear of mammals consists of a tympanic membrane, an ossicular chain of three elements, and two tympanic muscles. The tympanic membrane bulges inward, unlike the usually outward-bulging membrane of reptiles and birds. The elements in the ossicular chain are the malleus (hammer), incus (anvil), and stapes (stir-

rup), so named because of the resemblance of the bones to these objects. The malleus is attached to and partly embedded in the fibrous layer of the inner surface of the tympanic membrane. It connects to the incus, which connects in turn to the stapes, the footplate of which lies in the oval window of the cochlea.

One tympanic muscle extends from an attachment to the skull to an insertion on the malleus. Another muscle has its insertion on the neck of the stapes. By their contractions, both muscles add friction and stiffness to the ossicular chain, thereby reducing its mobility and protecting the inner ear from excessive sounds. The contraction of the muscles is a reflex action and occurs in both ears at the same time in response to loud sounds.

The coiled cochlea

The inner ear is called the cochlea because in man this structure is a complex tube coiled into about 2.5 turns, thus bearing some resemblance to a snail's shell, from which the term is derived. The name cochlea has now been extended to include the auditory portion of the labyrinth in all animals, even when the structure is not coiled, as in reptiles, birds, and egg-laying mammals. In the mammals in which it is coiled, the number of turns in the cochlea varies with species from a little less than two to as many as four. The guinea pig and its relatives have the largest number of cochlear turns. Extending along the inside of this coiled passage is the basilar membrane, bearing on its surface the sensory structure known as the organ of Corti, which contains the hair cells.

In mammals a uniform system is employed in the stimulation of the hair cells by sounds. A relatively thick tectorial membrane, anchored securely on one edge to the supporting structure (the limbus), lies with its free portion over the hair cells and with the cilia of these cells firmly attached to the lower surface of this portion. When vibratory movements of the basilar membrane cause the bodies of the hair cells to move, the tips of the cilia are restrained by their attachments to the tectorial membrane. Hence the relative motion between the bodies and cilia of the hair cells stimulates them.

The sizes, shapes, and spatial relations of many otic structures vary in the different mammalian species, but it is thought that the same basic principles of operation are involved. This uniformity contrasts with their situation in reptiles, in which different systems are present both in different species and sometimes within one ear.

Auditory sensitivity among mammals

A number of features are of particular significance in determining the sensitivity and frequency range, which vary with species. Because large masses involve great resistances when moved at high frequencies, the size and mass of the moving parts determine to some degree the variations of sensitivity with frequency and the frequency limits within which the ear operates. The ossicular chain is a mechanical lever, and its lever ratio and the difference in area between the tympanic membrane and the stapedial footplate determine the efficiency of sound transmission from air to the cochlear fluid. The mechanical characteristics of the cochlea and the degree of variation of these characteristics along its extent determine the frequency range of hearing and the degree to which different tones can produce different response patterns. Finally, the numbers and distribution of hair cells along the basilar membrane and the density and specificity of innervation of these cells determine the delicacy and precision with which their periodic activity and spatial patterns are registered by the central areas of the auditory nervous system.

These anatomical features have been studied in detail in a few animals: among the mammals, mainly in cats, guinea pigs, and to a lesser degree in man. The functional aspects, as shown in responses to sounds and to discriminations among different sounds, have been considered principally in man and to a much more limited extent in other mammals. (The characteristics of human hearing are treated at length in the article EAR AND HEARING, HUMAN.) Some of the auditory characteristics of mammals below man are described in the sections that follow.

Hearing in subhuman mammals. *Primates.* The hearing of other species in the division of mammals to which man belongs has always been of special interest. A number of species have been studied, including monkeys, marmosets, and chimpanzees among the primates considered as the most advanced, the anthropoids; and tree shrews, lemurs, and lorises among the more primitive.

By using a variety of training methods with chimpanzees, monkeys, and marmosets, behavioral thresholds have been recorded in response to sounds of different intensities and frequencies. When compared with each other and with man, it has been found that the hearing sensitivity of these animals and man is remarkably similar over a range of frequencies from 100 to 5,000 hertz, after which the sensitivity begins to differ. The differences observed at the higher frequencies, however, may be partly attributed to variations in experimental procedures. Thus, the results for the chimpanzee stop at 8,192 hertz because this was the highest tone used in the tests. Other observations have shown that chimpanzees can hear tones up to about 33,000 hertz and that young human subjects often hear tones as high as 24,000 hertz. It is also evident that monkeys and marmosets of the species studied can hear still higher tones.

Common laboratory animals. Certain mammals have long been favourite subjects for various kinds of biological studies in the laboratory, largely because of their convenient size, hardiness under caged conditions, and gentle temperament. Familiar among these are cats, dogs, guinea pigs, rats, mice, rabbits, and, more recently, hamsters, chinchillas, and gerbils. Auditory sensitivity functions have been obtained in these animals by a variety of behavioral and electrophysiological methods.

When measured behaviorally by conditioned responses and then plotted on a curve, the auditory threshold sensitivity of cats, guinea pigs, and chinchillas is much the same—a progressive improvement in sensitivity as the frequency is raised until the middle tones (about 500 to 5,000 hertz) are reached, at which point sensitivity tends to remain the same, and then shows a rapid loss in the upper frequencies. There are differences, however, in the maximum sensitivity attained in the middle region, with the guinea pig the least sensitive and the cat the most sensitive of the three species.

Differences between behavioral and electrophysiological observations

Sensory responses in the cochlea of mammals have been measured electrophysiologically by placing an electrode on the round window membrane. Unlike behavioral curves, however, the curves obtained by plotting the sound required to produce an arbitrary amount of electrical potential of the cochlea do not represent auditory thresholds. Instead, their usefulness is largely in their shapes, which indicate in a relative way the regions of good and poor sensitivity. In addition, these curves represent the performance of the peripheral portion of the auditory mechanism up to the point at which the sound stimulus activates the sensory hair cells in which the potentials are generated. Hence, unlike the curves obtained by behavioral responses, those obtained by cochlear potential methods do not indicate the performance of the central auditory nervous system (the nerve connections between the ear and brain and those parts of the brain in which neural impulses from the ear are processed to produce behavioral responses).

In the simpler animals, the two types of curves are much alike, judging from the very limited evidence available. In mammals, however, the behavioral curves differ from the cochlear potential curves in three ways. In the behavioral curves there is (1) an exaggerated gain in sensitivity to tones of low frequency, (2) a greater sensitivity to the medium-high tones, and (3) a more rapid loss of sensitivity to the extreme-high tones and a lower frequency of the upper limit. These differences are believed to arise mainly through the elaborate neural processing that takes place in the more highly developed mammalian nervous system, a processing that improves the sensitivity to high-frequency tones but reaches a limit of effectiveness and finally fails above some frequency limit. With these conditions in mind, the electrophysiological curves can be used to predict reasonably well an animal's behavioral responses to sound waves.

Large mammals. Because most of the mammals in which hearing has been studied by laboratory methods

are small, much less is known about the auditory capabilities of large ones, even of such domesticated animals as horses and cows; nevertheless, it is usually assumed that the auditory capabilities of these animals are much like those of humans. At least they hear sounds in man's vocal range because they seem to respond to verbal signals. Elephants, for example, trained as working animals, are said to obey as many as 30 different commands. A number of wild animals of medium and large size—raccoons, opossums, and several members of the cat and dog families—have been studied electrophysiologically by the cochlear-response method. Their sensitivity curves are fairly similar in form and in the upper limits attained.

Marine mammals. Of special interest are the sea mammals, which have been derived from early land species and which have undergone certain changes in order to adapt themselves to at least a partially aquatic existence. In the course of adapting to marine conditions, however, some sea mammals, such as seals and sea lions, seem to have made only limited alterations in their ear structures. In addition to being able to close the meatus when diving, their pinnas have been greatly reduced or essentially lost, a feature of streamlining for rapid progress through the water.

Adaptations for aquatic hearing

There are three possible ways that the hearing of marine mammals might be adapted to an aquatic environment: (1) unchanged aerial hearing, with no aquatic adaptation; (2) conversion to an aquatic type of hearing with loss of good hearing for aerial sounds; and (3) development of some kind of double system, with at least serviceable reception of both aerial and aquatic vibrations. In a study of hearing in the common seal, in which responses to aerial and aquatic stimuli were compared, it was found that this animal has a greater sensitivity to aquatic sounds, especially in the upper frequencies, which extended to the remarkably high frequency of 160,000 hertz. Yet, although the seal has made an adjustment for hearing in water, it has not sacrificed the quality of its aerial hearing, which remains at an excellent level, especially for one frequency around 2,000 hertz and another around 12,000 hertz. These differences in auditory sensitivity suggest that the mechanisms in this animal for aerial and aquatic hearing are somehow different, but no complete explanation of the adaptations has yet been found.

Whales, on the other hand, have converted their ears to a truly aquatic form, apparently with some sacrifice of aerial reception. The study of their ears and hearing has been carried out in only a few species of the toothed whales, which produce sounds and use their ears in the process of echolocation (see next section).

The ear of whales has undergone extensive changes. The pinna is absent and the external ear opening has been reduced to such a minute size, almost a pinhole in some species, that it no longer serves as a path for the entrance of sound. The eardrum, although present in a modified form, seems to serve no useful purpose; it is connected to the malleus only by a ligament, and this connection can be cut without an ensuing loss of sound reception. The usual three ossicles of the middle ear are present, with the footplate of the stapes resting in the oval window. These ossicles are much more massive than the ordinary mammalian ossicles.

It appears that the whale ear has been converted to a true aquatic type, functioning according to principles similar to those found in the ears of fishes, as described earlier. Sound vibrations in the water readily pass through the tissues of the head and reach the deep-lying middle- and inner-ear structures. Probably the ossicles represent an inertial mass in somewhat the same way that the otolithic body does in fishes. Because of their inertia, the ossicles tend to move with smaller amplitudes and in different phase relations when the tissues of the head, including parts of the cochlea, are set in vibration. This difference in relative motion produces an alternating displacement of the cochlear fluid, which is in contact with the footplate of the stapes and which can be set in motion because of the presence of a pocket of gas in the region of the round window. The performance of the whale ear has been measured in an exact manner throughout the frequency range in one species, the bottle-nosed dolphin (*Tursiops truncatus*). By a conditioned-response method, it has been found that this animal possesses excellent auditory sensitivity that extends well into the high frequencies.

Echolocation in bats. Bats are divided into the large bats and the small bats. With one or two exceptions, the large bats live on fruits and find their way visually. The small bats feed mostly on insects, catching them on the wing by a process known as echolocation. As was mentioned earlier, echolocation is a process in which an animal produces sounds and listens for the echoes reflected from surfaces and objects in the environment. From the information contained in these echoes, the animal is able to perceive the objects and their spatial relations.

How echolocation works

Bats produce sounds with the larynx, an organ in the throat that has undergone certain adaptations that make it unusually effective in producing intense, high-frequency sounds. The character of the sounds varies with the species and also with the particular activity. On striking a small object such as a flying insect, the emitted sounds are reflected with only a small fraction of their original energy; the sound is further weakened before reaching the ears of the bat when it must travel some distance through the air.

Although the frequency of bat cries varies with species, their cries usually occur in a range between 80,000 and 30,000 hertz. In most species, such as *Myotis lucifugus* and *Eptesicus fuscus*, the cry is a frequency-modulated pulse of sound; it begins at a high frequency, say, of 70,000 hertz, and in about 0.2 second declines in frequency to about 33,000 hertz. The starting frequency may vary, even in successive cries; a second pulse might begin at 60,000 and end at 30,000 hertz. The greatest energy in the cry is usually in the middle of this frequency range, perhaps around 50,000 hertz in the species mentioned above.

The use of such high frequencies is an essential feature of the bat's sonar system. In order to determine the nature of objects by reflected sounds, it is necessary that the wavelengths of the sound be small in relation to the dimensions of the objects—indeed, as small as possible if fine details are to be represented.

An important problem of echolocation is how the bat is able to detect reflected sounds, often in the presence of disturbing noises, and to obtain the information necessary for tracking and catching an insect as well as discriminating between this object and others in the environment. This problem involves considering first the structure of the auditory mechanism in bats and then the nature of their hearing.

The external ear of bats is usually well developed. In most species the pinna is large relative to the size of the head, and in those species called the whispering bats, because they make such faint sounds, this structure is huge. With its large surface, the pinna acts as an efficient collector and resonator of high-frequency sounds. It is also freely movable and can be rotated and inclined in various ways. The meatus leads inward to the eardrum and, as already mentioned, contains a valve that can be closed to reduce the entrance of sounds. The middle ear of bats is of the usual mammalian pattern—a three-part ossicular chain—but its structure is impressive in the extraordinary delicacy of the moving parts. The two tympanic muscles, however, are relatively large.

The cochlea of bats also shows the general mammalian form, but there are variations that may be significant for the special functions that are performed by this ear. The basilar membrane is not particularly well developed; it is short in comparison with that of most mammals, and its structural variation from basal to apical ends is only moderate in extent. Whereas most basilar membranes are rather strongly tapered in width, being narrow at the basal end of the cochlea and several times broader near the apical end, in the bat there is only a slight taper, between twofold and threefold. Another curious feature in the cochlea of bats is the presence of local thickenings

of the basilar membrane that may add to the stiffness of the cochlear structure.

The auditory portion of the nervous system has undergone extraordinary development in bats. The regions concerned with hearing are relatively enormous, which is in accord with the great predominance of hearing over the other senses in this animal.

Auditory sensitivity of bats

The hearing of bats has been studied by both electrophysiological and behavioral methods. In the species *Myotis lucifugus*, electrophysiological measurements of cochlear potentials indicate that response is poor in the low frequencies but improves fairly steadily until the range of 2,000 to 5,000 hertz is reached, at which it tends to level off. Beyond 15,000 hertz there are many irregularities but, in general, the sensitivity declines at a rapid rate. The results of similar studies on a specimen of *Eptesicus fuscus* are much the same as those for *Myotis*, though the observations were not extended into the lowest frequencies. The most sensitive range for this species is around 4,000 to 15,000 hertz, after which there is a fairly rapid decline in the upper frequencies.

The behavioral threshold curve for *Eptesicus* has a markedly different form. There is a rapid improvement in sensitivity from 2,500 to 10,000 hertz, but the greatest sensitivity is in two peak areas, from 10,000 to 30,000 hertz and from 50,000 to 70,000 hertz, with a separation by a moderate reduction around 40,000 hertz.

There are other peculiarities of the behavioral sensitivity of *Eptesicus* to sound stimuli that are of particular interest. The rapid loss of sensitivity to tones around 40,000 hertz may be caused by a failure of neural processing for these tones. The slope of the low-frequency end of the curve is unexpectedly steep, and this appears in a region where the cochlear response is passing through its maximum. Nothing like this has been observed in other animals; it seems to be a peculiarity of the bat.

When the cochlear responses of bats are compared with similar responses in other small mammals—as, for example, the rat—there is a general similarity in the results. The rat, however, has better sensitivity as measured by this method, reaching a level of especially good acuity in the range from 20,000 to 60,000 hertz, the range in which the bat sensitivity falls off rapidly. As mentioned previously, it must be kept in mind that the sensitivity indicated by the cochlear potentials is mediated in the peripheral mechanism, before involvement of the central auditory nervous system. When the behavioral response is considered, however, the contribution made by the bat's central auditory nervous system can be appreciated: the region of greatest auditory sensitivity, extending from 10,000 to 70,000 hertz, is the same region as the frequency of the echolocation cries and the one in which bats have the greatest need of acute hearing.

The failure of these bats to exhibit a behavioral response to tones below a frequency of 10,000 hertz can perhaps be explained also in relation to their peculiar use of hearing. This is a region of frequency that has little or no value for echolocation. More than that, it often contains noises of various kinds that, if heard, might be detrimental to this essential function. It has often been observed that bats are not easily disturbed by extraneous sounds of low frequencies, even extraordinarily intense ones. This peculiarity of hearing in bats may account for their resistance to masking sounds. The slight degree of structural differentiation found in the cochlea of bats may represent another aspect of the limitation of their hearing to that part of the sound spectrum most useful in echolocation. Therefore, it appears that the ear of the bat, which is a rather ordinary type of mammalian structure so far as level of auditory sensitivity and degree of tonal differentiation are concerned, has been developed for a particular purpose—namely, the reception of high-frequency sounds within a limited range.

Echolocation in other mammals. Among the mammals possessing echolocation are the toothed whales. These animals probably produce sounds in the water in two ways: with the larynx and with the complex system of passages connected to the blowhole, which is a nostril in the top of the head. Although many different types of sounds are possible, during echolocation they consist mainly of a rapid series of clicks. These clicks contain many components, but the principal energy is in the high frequencies, from perhaps 50,000 to as much as 200,000 hertz. The use of such high frequencies by these animals is a requirement for effective echolocation in water. Because the velocity of sound is greater in water than in air, the wavelengths are longer; therefore, in order for echolocation to attain the same effectiveness of object discrimination as that achieved by the bat with aerial sounds, an aquatic animal has to use frequencies at least five times as high.

High-frequency clicks of toothed whales

Whales have good vision when submerged, and apparently their eyes remain fairly serviceable when their heads are out of water. Dolphins can be trained to strike targets or leap over obstacles held several feet above the surface of the water. For many tasks, however, they use echolocation very effectively, such as when catching fish at night or when visibility is poor in murky water. Dolphins have been trained to make fine discriminations of objects when their vision has been completely excluded by blindfolding. Echolocation of some form and degree of effectiveness is suspected in still other animals, such as shrews and sea lions, but the evidence is meagre thus far.

BIBLIOGRAPHY. There are no general texts on sound reception. Included below are some of the specialized references dealing with this subject in a technical manner.

J. SCHWARTZKOPFF, "Mechanoreception," in MORRIS ROCKSTEIN (ed.), *The Physiology of Insecta*, vol. 1, pp. 509–561 (1964); M.L. WOLBARSHT, "Electrical Characteristics of Insect Mechanoreceptors," *J. Gen. Physiol.*, 44:105–122 (1960); E.G. GRAY, "The Fine Structure of the Insect Ear," *Phil. Trans. R. Soc.*, Series B, 243:75–94 (1960); C. WALCOTT and W.G. VAN DER KLOOT, "The Physiology of the Spider Vibration Receptor," *J. Exp. Zool.*, 141:191–244 (1959); E.G. WEVER and J.A. VERNON, "The Auditory Sensitivity of Orthoptera," *Proc. Natn. Acad. Sci. U.S.A.*, 45:413–419 (1959); H. and M. FRINGS, "Uses of Sounds by Insects," *A. Rev. Ent.*, 3:87–106 (1958); R.J. PUMPHREY, "Hearing in Insects," *Biol. Rev.*, 15:107–132 (1940); "Sensory Organs: Hearing," in A.J. MARSHALL (ed.), *Biology and Comparative Physiology of Birds*, vol. 2. (1961).

(E.G.W.)

Sound Recording and Reproducing

Sound is recorded and reproduced for a wide variety of purposes. Music provides entertainment; the spoken voice is recorded for business purposes (dictation), lectures, and language training; other sounds may be recorded for various reasons. Over the years, many techniques have evolved for recording and reproducing sound, but three primary ones have survived. These are mechanical (phonograph), magnetic (tape recorders), and optical (sound track on motion picture film).

Each of these systems was developed somewhat independently, and so they will be treated separately from the standpoint of both history and technology. Certain auxiliary components that function equally well with any of these three systems will be treated in a separate section. The concepts of high fidelity, frequency response, distortion, and similar facets of the technology will be treated in the last section (see below *High-fidelity concepts and systems*). A general discussion of the physical characteristics of sound appears in the article SOUND. (Ed.)

MECHANICAL SYSTEMS

A phonograph is an instrument for reproducing sound by directly or indirectly transmitting to the air the mechanical vibrations of a stylus in contact with a sinuous groove on a moving record on which sound has previously been recorded. A prototype of the phonograph appeared in 1857, when the French inventor Léon Scott constructed a device he called the phonautograph, which recorded, as a wavy line on the smoked surface of a rotating cylinder, the movements of a diaphragm or other vibrating body.

Early history. *The Edison phonograph and its successors.* The first record of vibrations that could be played back to reproduce sounds was invented in the United States by Thomas A. Edison in 1877. The record was in the form of indentations pressed into a sheet of tinfoil by

the vibrating recording tool, or stylus, which was attached to a diaphragm. The tinfoil was wrapped around a cylinder that rotated as the sounds were being recorded. A thread on a shaft moved the cylinder axially as it rotated, so that the record groove was helical. A small helical groove having the same pitch was cut in the surface of the cylinder, and the recording diaphragm and stylus were positioned to bring the groove directly under the slightly rounded point of the stylus, thus making it easier to produce the indentations in the tinfoil. The sound waves were recorded as variations in the depth of the groove, a process later referred to as hill-and-dale recording. For reproduction the groove was again run under the stylus tip, with only enough pressure to maintain contact with the groove bottom. The term embossing has been widely used to designate the method of recording in which the groove is impressed in a yielding material by a rounded point; this method is still used in some office dictating machines.

Several years later, two U.S. inventors, Chichester A. Bell and Charles Sumner Tainter, developed an arrangement much like Edison's, except that their machine cut or engraved the groove in a cylindrical surface of wax with a sharp tool carried by the vibrating diaphragm. They applied for a patent in 1885. This method gave much better sound reproduction than Edison's indented tinfoil.

First recording on a flat disc

Another inventor, the German-born Emile Berliner, was responsible for a radical departure in sound recording. Berliner's mechanism traced its wavy line as a spiral on a flat disc rather than as a helix on a cylindrical surface. The disc was of zinc, coated with a thin layer of fatty substance that protected it from the action of acid except where the scribe had traced its line. An acid bath then left a groove, the width and depth of which could be controlled by the duration and strength of the acid treatment. Berliner thus obtained a record that could be played on an appropriately designed reproducing mechanism.

Berliner did not, however, contemplate using this etched master as the record to be played; rather, a negative was made from the master by electroforming (see below), and the negative was then used to mold records in some thermoplastic material, a procedure that subsequently became standard. Berliner's disc records were far harder and more durable than the wax cylinders that were being made. The harder material, plus somewhat greater amplitudes because the recording stylus did not have to work against any cutting resistance, made somewhat greater volume possible in the reproduced sound. Of still greater importance to the success of Berliner's discs, however, were the ease of molding from a flat surface and the ability to make his grooves deep enough to guide the reproducing stylus across the record without the need for an auxiliary feed screw. Berliner went into business manufacturing and selling his Gramophones; later innovations included a spring-driven motor with a friction governor to hold the speed constant, which replaced hand cranking. He also introduced a greatly improved sound box and developed lateral recording, in which vibrations are recorded as sidewise deflections of a groove of uniform depth rather than as variations in the depth of the groove. Lateral recording attained wide acceptance both in the U.S. and in Europe starting about 1900.

As the phonograph and record business grew steadily in Europe and the United States, various technical innovations were introduced that improved the quality of sound reproduction. Methods of molding cylinder records of a relatively hard thermoplastic were introduced in 1901. In 1908 cylinders were made of a new material called Amberol, which was harder than the materials previously used. Before this time, recording grooves were spaced about 100 to the inch; by increasing the number to 200, up to four minutes of playing time was achieved. In 1912 Edison introduced a further improvement in material, called Blue Amberol, and in 1913 he began making discs and a phonograph to play them; he continued, however, to make cylinder records for cylinder machines for a number of years.

Acoustical horns. The performance of the acoustic phonograph was greatly affected by the size and shape of the horn; and while there was little in the way of technical study during the era of the acoustic phonograph, there was some progress toward better horns. Ear trumpets and megaphones had long been familiar, and the developers of phonographs were quick to take advantage of what horns could do for their machines. The horns assumed various shapes according to the ideas of those who made them. The easiest horn to construct was of simple conical shape, but after about 1920 the exponential horn (see below) was the nearly universal choice of engineers. A theoretical analysis of its characteristics was published by a U.S. physicist, Arthur Gordon Webster, in 1919, and the subject was further elucidated by two U.S. electrical engineers, Clinton R. Hanna and Joseph Slepian, in 1924.

Use of the exponential horn

An exponential horn is one the cross-sectional area of which doubles with each increase of x inches in distance along the axis from the small end; or, if it is a round horn, the diameter doubles with each increase of $2x$ in axial distance. The rate of flare is therefore represented by x; a small value yields a horn of rapid flare, while a large value yields a slowly expanding horn. Every exponential horn has a critical frequency, known as its cutoff; below this frequency the horn is of practically no use, while for higher frequencies it will reproduce very effectively. The cutoff frequency is that at which the wavelength of the sound is approximately $20x$; hence, to extend the range of response to lower frequencies, x must be made large.

From the time of the very first phonographs, it was customary to funnel the total air displaced by a diaphragm into a horn opening considerably smaller than the diameter of the diaphragm. This action results in greater loading of the diaphragm, reducing its tendency to resonate (*i.e.*, damping it) and deriving from it increased sound power. The volume of air within the sound box (the space between the diaphragm and the mouth of the horn) must be small so that appreciable compression will not take the place of airflow.

Horns can be bent or "folded" to conserve space without too seriously impeding their performance, provided the difference between the shortest and longest paths around the bend is not over about one-third the shortest wavelength to be transmitted.

Comparing a conical horn with an exponential horn of the same length and initial and final openings, the cone expands much more rapidly at its small end and much more slowly near the large end. The result is that the conical horn gives some multiplication of airflow at frequencies well below the cutoff frequency of the exponential horn, and the output climbs gradually through a long range of frequency without ever quite overtaking that of the exponential horn above its cutoff.

The horns of the early phonographs frequently ended in a rapidly flaring bell, like a trumpet. In some cases the angle between side and axis increased toward the large end, giving a shape between that of a cone and an exponential horn. Early horns were relatively small, however, and were not designed to avoid resonances caused by reflections at the bell. Some of the resonant points may have been regarded as desirable for increased sensitivity. The feeble response at low frequencies was of less consequence because there was little low-frequency content on the records. Recording equipment was usually provided with larger horns, often of simple conical shape.

Problems inherent in early horns

As experience accumulated, the sizes of horns increased, at least on the better machines. In 1906 the Victor Talking Machine Company marketed a machine with the horn built into the cabinet. This arrangement facilitated increasing the horn size and also made it desirable to keep down the size of the middle part of the horn, thus favouring a shape somewhat closer to the exponential. The slight taper of the swinging pickup arm, which constituted the first part of the horn, increased the general resemblance to the exponential shape. In later models with larger cabinets, larger and still better shaped horns were employed. It was not until the introduction of the Orthophonic design in 1925, however, that a definitely planned exponential expansion was followed, with considerable increase in length and with an end opening occupying practically the whole front of a large cabinet.

Electrical recording and reproducing. In recording for the Edison phonograph and its immediate successors, acoustical methods were employed. Sound was collected and applied without amplification to actuate the diaphragm of the record-cutting device. The great difficulty in such direct recording lay in the extremely limited amount of acoustical power that could be collected. Performers huddled close to large clumsy recording horns; only the most powerful instruments and voices recorded well.

This disadvantage was overcome shortly before World War I with the introduction by the scientist Lee De Forest in the United States of the Audion (triode) vacuum-tube amplifier, a device that could multiply sound power a thousandfold or more with little loss of quality. Though work toward phonograph applications of the Audion amplifier began as early as 1915, the war intervened, and it was not until the early 1920s that the transition from acoustical to electrical methods took place. With the advent of electrical recording, performers could work under more natural conditions; microphones collected the sound, and it was then amplified electrically to the necessary level for recording. The Audion amplifier was also applied in reproduction, leading to replacement of the conical horn by the electric loudspeaker (see below).

The Audion amplifier

An important aspect of electrical recording and reproducing-apparatus development involved the design of vibrating parts. The essential moving parts of either a recorder or reproducer have a certain amount of mass. Forces to actuate these masses must be produced magnetically in the case of the recorder or by the action of the record on the needle tip in the case of the reproducer. In order to minimize the required forces, the moving elements in these devices were separated. Thus, in the recorder, the armature of the electromagnetic motor was separated from the stylus bar and stylus; and in the reproducer, the needle with its clamp and the lever that connects it to the diaphragm constitute one part, while the diaphragm constitutes the other principal mass. By introducing an elastic connection between the two parts, it is possible to reduce almost by half the maximum force that must be applied to the needle to vibrate the masses. These improvements in design made it possible to increase the sound output from records and paved the way for later developments such as fine-groove and stereophonic phonograph records. (E.W.Ke./H.F.O.)

Record groove parameters. The dimensions of a record groove are determined by the lateral displacement of the groove resulting from the sound and by the dimensions of the reproducing stylus; a reduction in either or both of these permits a reduction in the dimensions of the groove and allows them to be placed closer together, thus making possible a smaller record or one with longer playing time. The use of less noisy record materials such as clear Vinylite permits a reduction in the lateral displacement of the groove without reducing the signal-to-noise ratio; *i.e.*, makes finer grooves possible without increasing record surface noise. The use of pickups that require a smaller amount of force for their actuation permits the use of a smaller reproducing stylus. Experimental evidence during the early 1930s demonstrated that up to 300 grooves per radial inch (about 12 grooves per radial millimetre) and a reproducing stylus with a radius of 0.001 inch (0.025 millimetre) were practical.

Distortion in a record groove

The amount of distortion in the signal reproduced from the record groove increases as the stylus moves toward the centre of the disc—that is, as the relative circumference of the groove decreases and the relative speed of the groove past the stylus decreases. Therefore, the speed at which a disc record should rotate is related to the amount of recorded material it contains, and both speed and amount of material are governed by the circumference of the innermost groove that will give the maximum permissible distortion.

In addition, there is a definite relationship between the diameters of the outermost and innermost grooves that gives the maximum playing time per unit surface area. This relationship is such that maximum playing time for a disc of given size will be achieved if the diameter of the innermost recorded groove is one-half the diameter of the outermost groove; this relationship is also governed by the amount of permissible distortion in reproduction obtained from the innermost groove. Finally, any major change in playing time (other things being equal) calls for a change in the speed of the record in order to achieve maximum playing time with the minimum amount of record material.

Based on these principles, in 1948, fine-groove records of clear Vinylite were introduced by two different manufacturers: one was 12 inches, or about 30 centimetres, in diameter rotating at 33⅓ revolutions per minute with a playing time of about 30 minutes; and the other was a seven-inch- (18-centimetre-) diameter recording with maximum playing time of eight minutes, turning 45 revolutions per minute. These new records offered such improvements in cost, convenience, and performance that they rapidly supplanted the old discs, which had been standardized at 78 revolutions per minute. Records revolving 16⅔ times per minute are used for recordings of human speech, such as plays and readings from novels. Such recordings are especially helpful for the blind, many being circulated by mail from public libraries and sometimes referred to as "talking books." (A.D.Bu.)

Stereophonic phonograph records with two separate channels of information recorded in a single groove became a commercial reality in 1958, less than a year after the adopted system had been publicly demonstrated. In a live musical performance, the sounds of various instruments and voices come from different directions. Stereophonic systems, by reproducing sound in its original spatial auditory perspective, produce a far greater sense of realism than single-channel or monaural systems (see below *Stereophonic reproduction*). A stereophonic system requires two or more independent channels, with separate microphones for recording and separate loudspeakers for reproduction. The standard stereophonic phonograph system employs a dual-channel recorder and a single stylus with its axes of movement inclined at 45° to the record surface, the pickup being designed to produce two practically independent signals in accordance with the motion components along the two (90° displaced) axes. Because the record groove is V-shaped and ordinarily has an included angle of 90°, the groove walls are at a 45° angle with respect to the record surface. Consequently, for the 45°–45° system the signal from one channel varies the position of one groove wall about a mean (no-signal) condition without changing the position of the other.

Operation at 45° provides a symmetrical arrangement that offers advantages both in design of recorders and reproducers and in quality of reproduction because electronic characteristics of the two channels are alike. In addition, good reproduction is possible over a single-channel (monophonic) system because information from both stereophonic channels is contained in the resultant lateral modulation. A record that can be played stereophonically over stereophonic equipment or monophonically over monophonic equipment is called a compatible recording. (H.E.Ro.)

Record materials and production processes. The early tinfoil and other recording media ultimately were supplanted by materials of two general types: (1) the wax disc (actually a heavy-metal soap with wax gel) intended for subsequent processing and duplication and (2) the instantaneous media intended mainly for playback directly from the embossed or engraved original record.

Requirements for groove accuracy

The necessary accuracy of the recorded groove can be judged by the requirements established for fine-groove records: that the bottom radius of the groove be less than 0.00025 inch (0.0064 millimetre) and the width about 0.0025 inch (0.064 millimetre), giving a depth of about 0.00125 inch (0.032 millimetre). This groove is less than one-half the groove size of the 78-revolutions-per-minute record.

In cutting or engraving the master record the stylus must be sharp and accurately ground so that the cut will be very clean. Lacquer-coated discs, usually on an aluminum base, were developed as an instantaneous recording medium, but because of the slightly rubbery nature of

lacquer a smooth, low-noise groove was not obtained. The development of a burnishing facet on the stylus cutting edge resulted in quieter grooves but increased cross modulation (echo) between adjacent grooves. The introduction of the heated cutting stylus, about 1950, improved lacquer recording to the point where wax discs became practically obsolete for commercial master recordings.

Electroforming. To make copies of the recording, dies must be made the surface of which is a negative replica of the master-record surface. Electroforming techniques that reproduce the surface to an accuracy of better than 0.000001 inch are used. In electroforming, the surface serves as an electrode in an electrolyte solution containing the metal to be deposited. When current of the proper polarity (positive or negative charge) is passed through the solution, a metal sheet is formed the thickness of which is proportional to the current and time and the surface of which is an exact negative replica of the master-record surface.

In order for this process to be carried out, the master-record surface must be made electrically conductive. Before 1938, wax recordings were made conductive by brushing graphite or bronze powder into the wax surface. In the 1930s gold sputtering (a process for depositing an extremely thin layer of gold) was introduced and was widely adopted by the major record manufacturers. With the improvements in lacquer discs and recording techniques, it was found that wet-silvering methods (*e.g.*, the chemical reduction of a silver solution such as that used for silvering mirrors) led to improved results. By the early 1950s, lacquer discs with wet-silvering metallizing had been universally adopted by major record manufacturers.

After separation from the recording, the electroformed metal master disc can be used in a plastic molding press to press records, and in cases in which only up to 100 or so records are desired the electroformed metal master commonly is used. For records in commercial numbers, many pressing masters, called stampers, may be needed. It has been the practice, therefore, to electroform a metal positive from the metal master. This positive, called a "mother" or mold, can be played as a record for quality approval and then used to generate a number of stampers by electroforming. By using a metal master to generate a number of molds and each mold to generate a number of stampers, it is possible to make many stampers, and thus thousands of records, from each recording.

Use of stampers for commercial records

Record materials. Commercially, records are molded from thermoplastic materials, which become soft and flow under pressure when heated but become rigid when cooled to room temperature. Shellac, a natural thermoplastic, was used in combination with a finely ground filler such as clay for many years, but after the early 1930s synthetic thermoplastic resins gradually replaced it. High-quality record compositions without filler were first used for transcription records for radio broadcast. With the introduction of fine-groove records, use of the higher-quality compound became necessary to reduce the noise caused by the filler particles. Of the synthetic materials, the vinyl plastics are most commonly used for records.

Phonographs and record changers. *Stylus shapes and materials.* Except for the Edison hill-and-dale phonograph, which had a diamond reproducer stylus, most phonographs before 1935 required frequent changes of stylus or needle. Heavy pickups, weighing from 100 to 200 grams (3.5 to seven ounces) and with the whole weight bearing upon the small area of contact at the tip of the stylus, can exert pressures far in excess of the strength of record materials. Shellac-compounded mixtures used in records included a large proportion of abrasive filler, usually fine limestone and clay, which soon ground the tip to conform to the groove shape. In this way the contact area between stylus and groove was increased materially, thus decreasing the unit pressure so that record wear was reduced. When, between 1933 and 1935, attempts were made to use sapphire styli with electrical pickups weighing 50 to 150 grams (1.8 to 5.3 ounces), record wear was found to be excessive, probably because the sapphire could not quickly be worn sufficiently to reduce the unit pressures at the stylus tip to a satisfactory level. It was not until the introduction in the late 1930s of pickups weighing less than 35 grams (1.2 ounces) that sapphire and diamond styli became practical. The lighter pickups also permitted the use of unfilled-plastic record materials.

When comparatively large forces were necessary to extract the signal from the groove, the output was affected by the type of needle used. A short, stubby steel needle resulted in high output with good high-frequency response, but the needle also brought out the surface noise of the record. With a longer, thinner shanked and more flexible needle, the overall output was reduced, but the high frequencies were attenuated to a greater degree, making the surface noise less audible. Steel needles were used most commonly during this period, although "low-noise" fibre or thorn needles, more compliant than steel, also were employed.

Effect of the needle on output

Steel needles had to be changed after every play with abrasive-filled records. In the 1930s chrome-plated needles, which could be used somewhat longer, were introduced, and needles with hard metal tips such as osmium were also available. With the advent of lightweight pickups, jewelled styli (sapphire and diamond) became predominant.

Record styli are made to precise specifications. The point is spherical, tapering off to a 50° cone. For the 78-revolutions-per-minute record, a 0.003-inch (0.076-millimetre) point radius is desirable and for the fine-groove records, a 0.001-inch (0.025-millimetre) radius. Specifications for the stereophonic records call for a stylus point with a maximum radius of 0.0008 inch (0.020 millimetre).

There is no such thing as a permanent stylus. Even a diamond point will wear with time. With light pickups a diamond stylus should perform satisfactorily for 100–500 hours of playing time, a sapphire stylus about one-fifth of that time.

The needle should make contact on the sides of the groove for best reproduction. As the stylus wears, its sides become flat, and the stylus rides more closely to the bottom of the groove. A stylus riding on the bottom of a record groove results in noisy and distorted reproduction.

(A.M.Ma.)

Reproducer or pickup. A phonograph pickup is a device that when actuated by the record groove will produce a usable signal proportional to the signal (music) recorded in the groove. Common forms of phonograph pickups produce an electric signal proportional to the amplitude or velocity of the signal in the record groove.

Signal generation by the pickup

Phonograph pickup designers have explored nearly all of the known principles of electric generation, including the piezoelectric effect from certain crystals (the generation of a voltage by the deformation of the crystals), variable resistance, variable capacity, variable magnetic flux cutting, variable magnetic reluctance, and variable light-beam photoelectric effects. In general, it is possible to design a pickup using any principle that will produce an electrical signal from a mechanical signal. The selection of the principle and the design of the pickup are usually determined by cost and environmental factors.

Because of relatively high electric output, crystals of Rochelle salts, properly cut and designed into a pickup, have been extensively used in low- and medium-priced phonographs. Barium titanate and lead titanate zirconate are even more widely used, as these materials are less affected by moisture and temperature variations than Rochelle salts (see PIEZOELECTRIC DEVICES). Mechano-electromagnetic principles have found considerable use, moving-iron and moving-coil devices being commonest. Mechano-electro-capacity principles also have been employed.

Pickup designers, for the most part, have directed their efforts toward designs of low cost, low mechanical impedance, low distortion, and high reliability. Low mechanical impedance has received considerable attention because its attainment permits a low value of stylus force. The less

the stylus force, the less will be the wear on the stylus and the record groove. Concentrated effort on the part of pickup designers reduced the mechanical impedance to the point where only a fraction of an ounce of stylus force was necessary to hold the stylus in contact with the walls of the groove. This development permitted the use of clear Vinylite record materials and a stylus radius of 0.001 inch or less.

Stereophonic pickups

Stereophonic pickup design followed the same basic principles used for monophonic pickups. The so-called universal pickups used for hill-and-dale and lateral transcription equipment were the forerunners of the 45°–45° stereophonic pickups. In the stereophonic pickup, it was the design objective to produce two independent signals using a single stylus actuated by two independent signals recorded in a single record groove. Thus the pickup has two independent electric generating units connected to and actuated by a common stylus.

Record changers. A record changer is a mechanism designed to permit a stack of selected records to be played in sequence without human intervention.

While a number of early patents were granted on automatic record-changing devices, very few machines reached the hands of the public until after the mid-1930s. The low-cost, single-play phonograph attachment, for use with a radio, stimulated interest in the playing of records. As record collections grew, the need for record changers became apparent. By 1940 many manufacturers were equipping their phonographs with record changers.

Record-changing mechanisms are of varied designs. Commonest in the early period was the edge-selector type in which the records were placed over the centre spindle and supported at their circumference on two or more shelves. Blades or knives, above the shelves, sliced in between the bottom record and the next above to drop the bottom record onto the turntable. In another design the record stack was supported at the centre on an offset portion of the spindle and at the circumference by a shelf. The top portion of the spindle was tilted slightly to provide stability to the stack. The shelf support included a mechanism designed to push the bottom record from the offset portion of the spindle. Edges and centre holes of shellac records were damaged by these types of mechanisms.

After World War II, many improvements were made in record-changer mechanisms. The use of Vinylite record material practically eliminated damage to the record by the selector mechanism. In addition, improvements were made in the turntable drive system, tone arm, and tone-arm landing and tripping mechanisms. (A.D.Bu.)

Motors and drive systems. The first experimental machines were hand cranked, usually with some steadying effect by flywheels, but in one of his early patents Edison stressed the importance of correct and uniform speed. Battery-operated electric motors were soon used, but these were sensitive to changes in load and voltage; flyball governors were used to reduce speed variations. The most satisfactory of these governors were those with a friction disk, the friction of which increased rapidly if the speed exceeded the desired value. While Edison was reluctant to give up the electric motor, a spring-driven, governed motor developed by Frank Capps, a pioneer in many phonograph improvements, was demonstrated to Edison and adopted, and spring motors also were developed by the Columbia Company and by Berliner. At the end of 1896 all three companies were using governed spring motors.

Use of flyball governors in early machines

For recording, weight-driven machines with friction governors were used for many years. The weight supplied a constant torque, and the governor was connected to the turntable spindle through a worm gear. These were precision-built machines.

Electric motors did not come into very extensive use until radio-phonograph combinations were introduced; and then for many years the motor served only to provide a reasonably uniform driving torque, and friction governors, geared to the turntable, regulated the speed. The induction disk motor (like those used in watt-hour meters) was widely used in this manner. Small, high-speed, split-phase or shaded-pole induction motors with governors were among the many forms used during the 1930s, and a true synchronous motor, resembling a clock motor, was used in some machines. The latter had a multitoothed steel disk that, once started, was kept rotating by the intermittent pull of its toothed pole pieces.

"Rumble" in motors

"Rumble" caused by vibration of the turntable or pickup pivot presents a problem. The vibration is caused primarily by the high-speed motor and transmitted through the motor supports or the speed-reduction system. Cushion mounting of the motor is therefore necessary. Problems arise because the ear is especially sensitive to slow pitch fluctuations. Successful filtering is increasingly difficult as the frequency of the disturbance is lowered. Belting between motor and turntable can give excellent results in reducing rumble. Very light, thin, flexible belts of exceptional uniformity are required, especially if one of the pulleys is small; usually, two stages of reduction are used. Friction drives with rubber tires also are widely used because the rubber tires can be ground to a high degree of concentricity. By grinding the end of the motor shaft in its own bearings (after balancing the motor), the shaft end can serve very satisfactorily for the first unit of the drive train, even though it is as small as $\frac{3}{16}$ inch (five millimetres) in diameter. The inside of the flanged turntable, machined to concentricity, can be the final friction drive surface. The rubber tires can then be on the intermediate wheels. Because the load is light, the slip, or creep, is negligible. Careful attention to these design details will also help reduce wow, a low-frequency periodic change in signal frequency resulting from variations in turntable speed. (E.W.Ke./H.F.O.)

MAGNETIC SYSTEMS

Early history. Valdemar Poulsen, a Danish inventor, in 1898 applied for a patent on a device that stored electrical information by means of magnetized particles on a wire of magnetic steel. With later improvements, this telegraphone recorded continuously for 30 minutes on a length of piano-wire steel that moved at a speed of 84 inches (2.13 metres) per second. The device did not find wide application, however, and so Poulsen's manufacturing enterprise failed, and public interest in magnetic recording lagged.

In 1927 the first United States patent was issued for magnetic tape made by drying a liquid containing magnetic particles on the surface of a strip of paper. A German patent was issued in 1928 for a similar process. Until that time the recording medium had been a wire or tape made entirely of, or plated with, magnetic materials. In England, Germany, and the United States during the 1930s work progressed on a magnetic recording system to record and synchronize sound with motion pictures. In particular, the Magnetophon Company of Germany succeeded in producing a magnetic film-coated tape that constituted a low-cost recording medium. During World War II the magnetophon, using the coated tape, served the Nazi propaganda organizations well, and following the war, its principles were copied and improved upon in many countries, with Germany, England, and the United States leaders in this work.

Recording media. Magnetic materials in almost every possible form have been employed in one magnetic recording application or another. Wire, the first such recording medium, is no longer used, but drums of nonmagnetic material, their surfaces coated with a magnetic substance to record short bursts of information, are employed in electronic computers and some types of data recording. Cylindrical surfaces of rubber or plastic, impregnated with magnetic particles, are used to record short messages in telephone answering and recording machines. Discs, coated or impregnated with magnetic material and recorded in parallel or spiral tracks, are used in such applications as office dictating machines and electronic computers.

Modern recording tape consists of a plastic film base coated with magnetic powder that is held in place by a binder. Paper, once widely used as a base material, has been abandoned because of the superior strength and no-stretch characteristics of suitable plastics. Mylar and

Magnetic tapes

polyester plastics are widely employed as the tape base. A magnetic material widely used in coating tape is the red form of iron oxide, Fe_2O_3. The way in which this powder is prepared, the impurities deliberately introduced or accidentally present, and the physical size and shape of the particles have much to do with the characteristics of the finished tape. Red iron-oxide particles, needle-shaped and roughly one micron (micrometre) in length, are widely used. The black oxide, Fe_3O_4, has also been used as a magnetic coating material; it produces a higher signal strength but is more difficult to erase. Chromium dioxide, a magnetic material providing 50 percent greater output and wider frequency range, was coming into use in the early 1970s.

The magnetic power is thoroughly mixed with a binder, the nature of the latter depending partly on the base material to which it will be applied. The polymeric vinyl chlorides, vinyl acetate, cellulose acetate, and ethyl cellulose are commonly used. The solvent employed must not attack the base material or cause adjacent layers of tape to stick to each other when the tape is tightly wound and stored at relatively high temperatures. The object is to obtain a very thin but dense layer of iron oxide on the tape in a very uniform distribution. It must adhere tightly to the base tape but not to an adjacent layer; it must be flexible and preferably its characteristics should not change appreciably over many thousands of usages or many years of storage. The magnetic coating may vary in thickness from 0.18 to 0.65 mils (thousandths of an inch), depending on requirements.

The plastic base material may be from 0.5 to 1.5 mils thick to provide the desired physical characteristics. Generally, 1,200 feet (370 metres) are wound on a seven-inch (18-centimetre) reel to provide about 30 minutes of recording time when played at a standard speed of 7½ inches (19 centimetres) per second. Thin tape allows more tape to be wound on a given size reel, but if too thin, strength is reduced and there is danger of signal being transferred from one layer to the next when wound on a reel (called print-through).

Recording and playback processes. *Recording.* The electrical signal to be recorded is usually applied to the tape by means of a "head" consisting of a coil wound around a core of magnetic iron that has a gap at the point where the tape moves across its surface. The current in the coil produces a magnetomotive force across the gap, magnetizing the particles in the tape. The strength and direction of final magnetization of a particular point on the tape will be determined by the strength and direction of the magnetic field at the moment that point crosses the gap. Thus the tape receives a magnetic record containing direction (north or south pole), amplitude, and linear dimension (along the tape) corresponding to the direction (plus or minus), amplitude, and time of the original electrical signal. The exactness of this record depends on a variety of conditions, one of the most important being the original magnetic condition of the tape. There are two conditions that give predictable results. The tape is either completely demagnetized before recording, or it is completely magnetized in one direction or the other. In early wire and tape recording, the latter condition was used. The tape was carried past a permanent magnet so arranged that it brought the magnetic particles to saturation, leaving the tape fully and uniformly magnetized in one direction. The recording head then had a direct-current component added to the signal; this direct-current bias was of the proper direction and amplitude to bring the tape just back to a neutral state. The signal, adding and subtracting from this bias, left the tape magnetized by an amount proportional to its own strength. This procedure is called direct-current- (dc-) biassed recording and gives comparatively low distortion and good amplitude. It has a higher inherent noise level and more distortion than alternating-current- (ac-) biassed recording.

In ac-biassed recording the tape is first completely demagnetized and comes to the recording head in a neutral state. There an alternating current is added to the signal current; it is much higher in amplitude than the signal current and several times greater than the frequency of the highest frequency signal to be recorded. In crossing the gap of the recording head, the magnetic particles on the tape are subjected to many cycles of the bias current and leave the head with a magnetization that is proportional to the signal. Very low distortion and high signal-to-noise ratios can be obtained with ac-biassed recording. For sound and general signal recording, this type of tape recording gives performance superior to that of any other recording method.

Losses are encountered in recording through eddy current and resistive and capacitive losses in the windings and core of the head. Space between the head and tape causes very high losses, particularly at high frequencies (short wavelengths). The gap must be very straight and true, and the surfaces in contact with the tape must be very smooth in order to obtain the best possible transfer of the signal pattern onto the tape. Optical polishing methods are commonly used for finishing the surfaces of the gaps and faces of the head that will be in contact with the tape. The gaps used in recording heads vary in length (direction of tape motion) from 0.0002 inch (0.005 millimetre) when used also as a reproducing head to 0.010 inch (0.25 millimetre) or more for specialized applications. The tolerances on the trueness of the faces and edges of the gap are usually only a few millionths of an inch. The width of the cores and the resulting recorded tracks on the tape vary from approximately 0.025 inch (0.63 millimetre) for certain types of data recording to 0.25 inch (6.3 millimetres) for professional audio recording.

Playback. In reproduction, the tape is passed over the same head used in recording or over one similar to it. To reproduce the signal in the proper time relationship, the tape speed must be identical in recording and reproduction. The magnetized portions of the tape passing over the gap in the reproducing head cause the magnetic flux in the core to change, generating a voltage in the coil proportional to the speed of the tape and the amplitude and frequency of the recorded signal. In order to assure that tapes can be freely exchanged from one recording machine to another reproducing machine, it is necessary to establish detailed industry standards, a practice usually performed by committees of manufacturers.

Erasing magnetic tape

One of the features of magnetic recording that gives it an economic advantage over other recording methods is the ability to erase and reuse the recording medium thousands of times with little or no loss in the quality of recording. Erasure may be accomplished in either of two ways: (1) a reel of tape may be erased by subjecting the whole reel to a strong alternating magnetic field and slowly reducing the field to zero; this is called bulk erasing and is similar to the method used by a jeweler when demagnetizing a watch, although the equipment must be of a different form to accommodate large reels of tape. (2) a head, very similar to a recording head, is used to erase the tape just before it is recorded again; the tape passes over the gap in the head and is subjected to a powerful, high-frequency alternating magnetic field that diminishes as the tape passes away from the gap, leaving the tape in a demagnetized or neutral state. Professional sound-recording engineers frequently use both methods to make sure all the previous recording has been removed from a tape before it is used again.

Tape-recording machines. A wide variety of mechanisms has been devised to make use of tape recording in many different applications. Since only a small amount of power is required to move the light, flexible tape over the heads for recording or for playing back, it has been possible to make battery- and spring-driven portable mechanisms to record away from established power sources; the smaller ones are pocket-sized. The miniature machines, however, sacrifice some fidelity of reproduction to attain compactness. In particular, it is difficult to attain smooth motion and uniform speed, with the result that flutter and wow (high-frequency and low-frequency changes of pitch) may be detected in the playback of audio material. Larger portable units are available that give quite high fidelity and are often used for broadcast program recording in the field. High quality professional type magnetic

tape recorders are used for recording the master programs for phonograph, pre-recorded tape, and sound motion pictures. The recorders are in the form of eight, 12, 16, and 24 tracks. A single instrument or small group of instruments is recorded on a track. In transfer to the disc, pre-recorded tape, and sound motion picture master record, the programs recorded on the tracks are mixed to produce the desired balance and stereophony or quadraphony.

A large market has developed in tape recorders for use in the home. These record and play tape at 1⅞, 3¾, and 7½ inches (47.6, 95.2 and 190 millimetres) per second, usually on one-half of the width of the ¼-inch (6.35-millimetre) tape; this arrangement doubles the amount of tape available for recording or makes it possible to have two simultaneous channels for stereophonic recording. Pre-recorded tapes are available, but the ease with which the amateur can produce acceptable recordings is a strong factor in the popularity of such recording devices. Some machines are designed to record four or even eight tracks on a ¼-inch tape, allowing even more intensive use of the tape with only a slight further sacrifice in the quality of reproduction.

The most common method for driving a tape recorder or player (tape deck) is the so-called capstan drive, in which the tape is pulled forward by the friction between a pinch wheel consisting of a cylindrical piece of rubber and a motor-driven cylinder (capstan), rounded or tapered on the top, the speed and diameter of which determine the speed of the tape. In many machines both the feed reel, which holds the tape before playing, and the take-up reel, on which the tape winds as it plays, may be motorized in order to maintain the tape in tension.

The cassette

For ease in recording and playback, magnetic tape may be packaged in a cassette or loop cartridge. The cassette consists simply of a feed reel and take-up reel plus tape enclosed in a sealed, roughly rectangular package. The tape (usually 0.15 inch [3.8 millimetres] wide) begins on the feed reel at one side, passes through guides, and then travels across the face of the cassette where it comes into contact with the tape head located in the playback unit. It passes through guides once more on its way to the take-up reel on the other side. After play is complete, the cassette can be inverted and played in the opposite direction. Cassettes, either blank or containing pre-recorded material, are primarily used in the home, in consoles, or in portable battery-operated units. They are also employed for educational purposes; *e.g.*, to provide sound to accompany filmstrips. Cassette playback units are also very popular in automobiles.

The loop cartridge has only one reel and is employed principally for playback, for recording on such equipment is difficult and expensive for amateurs. Because of its design, the loop cartridge can play continuously for about 30 minutes, though models capable of two hours continuous play have been built. In the loop cartridge, the ¼-inch-wide tape feeds from the inside of a single reel, around a guide, past the head, around a second guide, and then back on the outside of the reel. Loop cartridges are used in stores or restaurants to provide background music or to play continuous verbal presentations at conventions or in museums. Cartridges are widely used in home high-fidelity (hi-fi) systems; most are designated eight-track stereophonic, meaning that four sets of two-track stereophonic material are recorded on them. Four-track tape with two sets of two stereo channels, though largely obsolete for home use, was still employed widely in broadcasting in the early 1970s. (W.E.St.)

OPTICAL SYSTEMS

Early efforts to produce motion pictures with sound involved attempts to synchronize phonograph recordings with the images on the screen. One inventor actually tied the phonograph to the projector motor through a system of belts and pulleys that ran the length of the theatre from booth to screen. As theatres grew larger, however, it became apparent that amplification as well as synchronization was a problem.

The amplification problem was solved first with the invention of the Audion amplifier (described above) and its application to loudspeakers. De Forest turned his attention to the synchronization problem and experimented with recording sound directly on the movie film. By 1923 he had developed techniques for transcribing the varying impulses of sound waves into varying impulses of light that could be photographed on a strip of film. The developed film, when passed between a light source and a photoelectric cell in the projector, could transform these images back into natural sound. Although there were variations in detail, De Forest's methods were fundamentally duplicated by subsequent sound systems in the U.S.S.R., the United States, and Germany. (For further description of sound recording and reproducing techniques in motion pictures, see MOTION PICTURES, TECHNOLOGY OF.) (Ed.)

Sound variations photographed on film

AUXILIARY EQUIPMENT

Microphones. The first successful instrument for converting acoustic power into electric power was the electromagnetic telephone transmitter used by Alexander Graham Bell in 1876. The term microphone was first used in 1878 when the concept of the variable resistance of a loose contact was elucidated. By virtue of this principle, the loose granules of carbon in a carbon telephone transmitter, under the influence of varying sound pressure, introduce a variable resistance into the circuit, thus controlling the energy delivered by a battery. As a result, the sound waves are converted into corresponding but amplified variations in electric current. Apart from telephone transmitters, the largest fields of application of microphones are in hearing aids, sound-recording systems (principally magnetic tape recorders), and dictating machines. Other widespread applications include public-address systems and radio and television broadcasting studios.

Operation of the microphone

The conversion, within a microphone, of sound-pressure variations in the air into corresponding electrical waves occurs in two operations that usually take place simultaneously. In the first operation, the sound wave impinges on a surface, generally referred to as a diaphragm, which is capable of slight movement. Variation of air pressure on this diaphragm causes it to move to and fro in a manner corresponding to the movement of the air particles. In the second operation, the diaphragm by its movement causes a corresponding change in some property of an electric circuit. Thus, for example, the displacement of the diaphragm may cause variations in the resistance of carbon granules or variations in the motion of a coil or conductor in a magnetic field. In each case, motion of the diaphragm produces a variation in the electric current.

Thus, the sound vibrations are first converted into mechanical motion, and then the mechanical motion is converted into variations in electric current. Hence, in studying the operation of a microphone, two things must be considered: (1) the mechanical movement of the diaphragm; and (2) the method whereby this movement sets up the desired variations of electric current. Either part of the process may be taken as a basis for classifying microphones.

Any microphone that has its diaphragm exposed to the sound wave on only one side is a pressure-operated type; that is, the displacement of the diaphragm is approximately proportional to the instantaneous pressure of the impressed sound wave. Outstanding among pressure types and of great historical importance is the carbon microphone, used mostly in telephony. Another pressure type, typically consisting of a thin stretched diaphragm separated by about one-thousandth of an inch (0.025 millimetre) from a parallel fixed plate, is the condenser microphone. The diaphragm acts as one plate of a variable capacitor (device for storing electricity) that is connected in series with a resistor and a direct current source. The motion of the diaphragm in response to sound waves varies the capacitance, producing a varying electric current. Condenser microphones are used for sound measurement and for high-quality recordings. Still another pressure type is the crystal microphone, used in hearing aids, sound recording, and office dictating machines. This

The carbon microphone

microphone depends for its operation upon the piezoelectric effect (*i.e.*, the generation of a voltage by the deformation of a crystal by a moving diaphragm). Crystals of barium titanate and Rochelle salts exhibit great piezoelectric activity; they are commonly used in crystal microphones (see also PIEZOELECTRIC DEVICES).

The dynamic or moving-coil microphone falls within the general class of moving-conductor microphones—*i.e.*, those in which an electric voltage is generated by the motion of a conductor in a fixed magnetic field. In the dynamic microphone a coil is attached to the moving diaphragm; motion of the coil in a fixed magnetic field induces a voltage in the coil. Dynamic microphones are used quite widely in recording, broadcasting, and public-address systems, especially when ruggedness is essential. Among the other types of microphones are the magnetic microphone and the ribbon microphone.

By proper design a microphone can be made to pick up sound from a single direction only, resulting in what is called a cardioid pattern. Such a microphone can be useful in discriminating against reflected and unwanted sounds and noises. Bidirectional microphones, in which the pickup is essentially the same from either front or back but zero from the sides, also are useful for specific applications. Microphones with approximately equal pickup from all directions are called nondirectional or omnidirectional. (E.I.G.)

Amplifiers. Because the output signals from phonograph pickups may be very low during quiet passages, amplifiers are required to build these minute signals up to the power necessary to actuate a loudspeaker. Controls are necessary to adjust frequency response to suit the listener's desires, the acoustic environment in the home, and any deficiencies in equipment or recordings. In the early 1970s transistors had almost completely replaced vacuum tubes for sound amplification, and integrated circuitry (*q.v.*) was rapidly coming to the fore. (Ed.)

Loudspeakers and speaker systems. Speakers are classified in much the same way as microphones. The most widely used is the moving coil of wire or "dynamic" loudspeaker, introduced in 1925. An electrical coil of wire is suspended in a fixed magnetic field, usually provided by a permanent magnet. Sound currents (*i.e.*, the electrical impulses into which sound waves have been transformed) flow through the coil, causing it to move. A diaphragm fastened to the coil pushes the air in front of it, creating sound waves. The dynamic speaker may also feature a baffle to prevent air circulation between the front and back of the diaphragm and an amplifier of adequate capacity to prevent overloading. The magnetic armature, or magnetic speaker, is of historical interest only, not having been sold since World War II. The signal current passes through a fixed coil that induces a magnetic field in a movable element of magnetic material called the armature, which is attached to a diaphragm. The performance of a loudspeaker depends heavily on the enclosure in which it is mounted; a wide variety of enclosure designs have been marketed, each with its claims to superior performance. In many cases, the choice between these different designs is more a matter of personal preference than of performance.

Magnetic speakers

Sound coverage of large groups of people comprises two classifications: public-address systems, in which the principal requirement is to provide intelligible speech over large areas, usually outdoors; and sound-reinforcement systems that must reproduce both speech and music with sufficient faithfulness to ensure that the sound system is unobtrusive. Sound-reinforcement systems are usually employed indoors and aim for greatest realism.

For large outdoor gatherings, directional horns are usually employed to concentrate the sound over the occupied areas, thus utilizing the available power most efficiently. Reflections from adjacent structures may reduce the required power, but such reflections may result in echoes which decrease rather than increase intelligibility; their effect is considered carefully. Sound-reinforcement systems for indoor use require less acoustic power because of reflections from walls and ceiling, but the entire system must be capable of high-fidelity reproduction of music and speech.

Headphones are popular for hi-fi listening in the home. They have been developed to the point where they provide excellent sound reproduction and ensure privacy with no limitations on volume, as frequently is the case with loudspeaker systems. Stereophonic headphones are especially popular because they provide a true stereophonic effect without the disturbing effects of listening-room acoustics and location of the listener in the room. The two channels of the stereophonic system are directed separately to the earphones of the headphone set. Earphones may be of either the dynamic or condenser type. The dynamic type functions in essentially the same manner as the dynamic loudspeaker, described above. A condenser earphone functions somewhat as a condenser microphone, described above, except that it is connected in such a manner that the diaphragm vibrates when voice currents are fed to it, thus reproducing the desired sound. (H.S.Kn.)

HIGH-FIDELITY CONCEPTS AND SYSTEMS

The term high fidelity applied to sound reproduction designates a state of performance required to provide a high order of realism. To achieve realism in a sound-reproducing system, four fundamental conditions must be satisfied as follows:

The four conditions required for realism in a system

1. The frequency range must be such as to include without frequency discrimination all the audible components of the various sounds to be reproduced.
2. The volume range must be such as to permit subjectively noiseless and distortionless reproduction of the entire range of sound intensities.
3. The spatial sound pattern of the original sound should be preserved in the reproduced sound.
4. The reverberation envelope or acoustic ambience of the original sound should be approximated in the reproduced sound.

To satisfy condition 1, the response must be uniform over the frequency range of at least 40 to 15,000 hertz, combined with faithful response to transients.

To satisfy condition 2, the nonlinear distortion must be of the order of less than 2 percent and the signal-to-noise ratio should be at least 60 decibels; that is, the signal should be at least 1000 times stronger than any noise that is present.

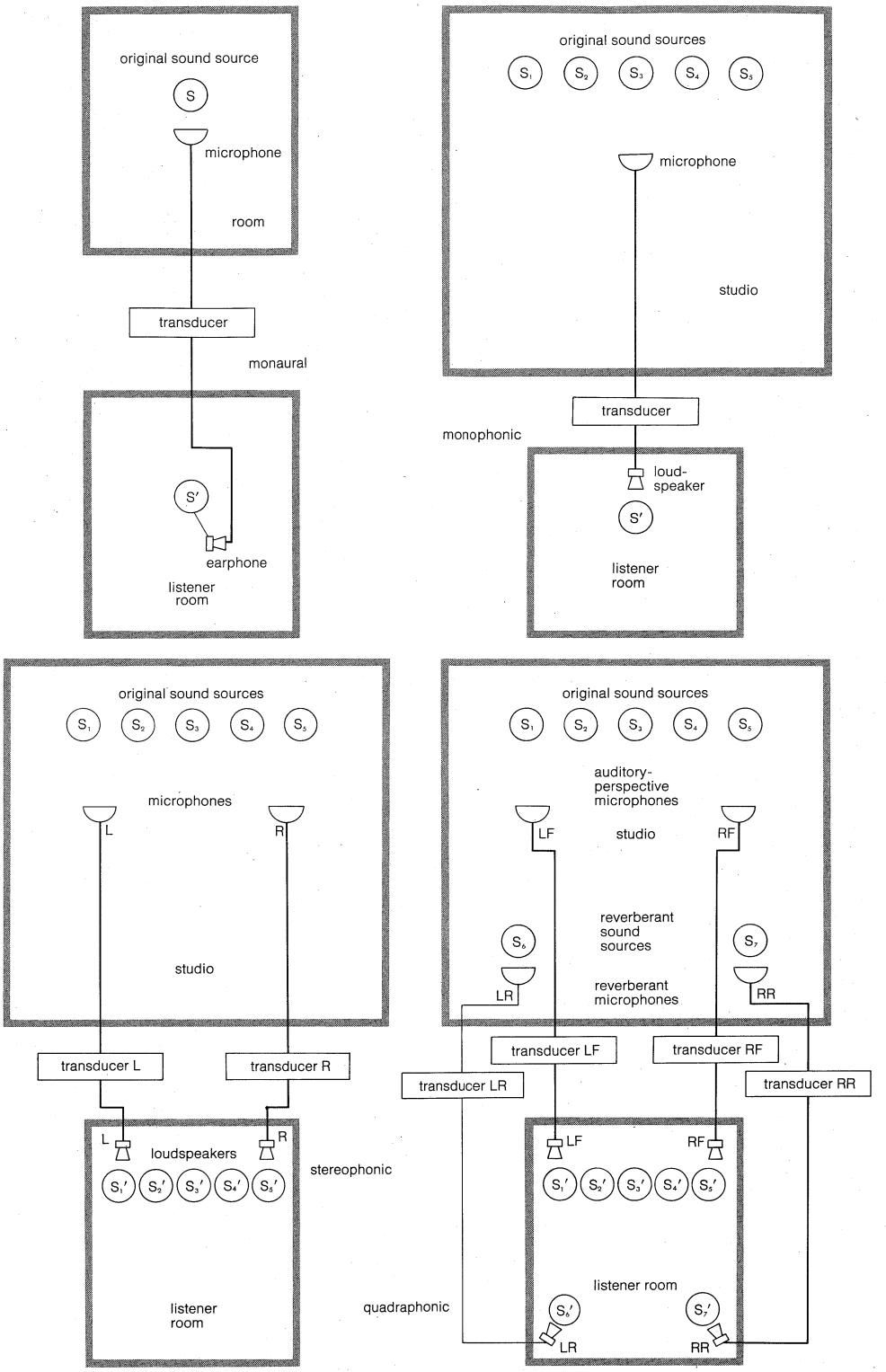
To satisfy condition 3 requires a stereophonic sound-reproducing system that provides reproduced sound in auditory perspective. The same order of spatial separation of the reproduced musical instruments and voices can be obtained in stereophonic sound reproduction as in the case of firsthand listening to the live sound sources.

To satisfy condition 4 requires the reproduction of the multireflected sound of the pickup environment as the reverberation envelope. In this way the reproduced sound simulates the reverberation envelope or acoustic ambience under live listening conditions.

Monaural reproduction. A monaural sound-reproducing system provides closed-circuit sound reproduction by means of a single channel. The action of such a system is shown at the upper left in the illustration. A microphone used to pick up the sound from the original sound source is coupled to a transducer followed by the earphone. This system may be constructed to satisfy conditions 1 and 2, but not conditions 3 or 4. Condition 4 cannot be satisfied because the microphone is usually placed very close to the original sound source and so does not pick up any of the reverberant sound.

Monophonic reproduction. A monophonic sound-reproducing system provides for the reproduction of sound by means of a single transducing channel. Its action is depicted in the upper right in the illustration. A microphone used to pick up the sound from the original sound source is coupled to the transducer, which in turn is coupled to the loudspeaker. The original sound sources are picked up by the microphone and reproduced by the loudspeaker. All the reproduced sound sources originate from a single location. This system may be constructed to satisfy conditions 1 and 2 but not condition 3. Condition 4 on acoustic ambience is only partially satisfied.

The monophonic-sound-reproducing system is the most



The four basic systems of sound reproduction.

widely used of all sound-reproducing systems. Examples are the disc phonograph, radio, sound motion pictures, television, magnetic tape reproducer, and sound-reinforcement systems.

Stereophonic reproduction. A stereophonic-sound-reproducing system provides for the reproduction of sound in auditory perspective. The action of a simple system of this kind is depicted in the illustration, lower left. Two microphones used to pick up the sound from the original sound sources are coupled to two independent transducers, which in turn are connected to two loudspeakers. The sound source 1 is picked up by the left microphone and reproduced by the left loudspeaker. Sound source 5 is picked up by the right microphone and reproduced by the right loudspeaker. Sound source 3 is picked up equally by both microphones and reproduced equally by both loudspeakers, and the reproduced sound source 3′ appears midway between the two loudspeakers. In the same way, the original sound sources 2 and 4 are reproduced as the sound sources 2′ and 4′. Thus, the geometrical configuration of the reproduced sound sources is similar to that of the original sound sources. As a consequence, the same order of separation of the individual reproduced musical instruments or voices can be obtained in stereophonic sound reproduction, as in the case of firsthand listening to live sound sources.

The stereophonic-sound-reproducing system may be constructed to satisfy conditions 1, 2, and 3, but condition 4 is only partially satisfied. Transducers for stereophonic reproduction include the disc phonograph, magnetic-tape reproducer, and frequency modulation radio.

Quadraphonic reproduction. A quadraphonic-sound-reproducing system provides for the reproduction of sound in auditory perspective and the reproduction of the reverberant sound envelope. The action of a simple system is depicted in the illustration, lower right. The LF and RF channels are identical to the L and R channels in the stereophonic system and provide the auditory perspective. The microphones in the LR and RR channels pick up the reflected or reverberant sounds. The LR and RR loudspeakers reproduce the reverberation envelope or the acoustic ambience. As a consequence, the quadraphonic sound system simulates direct listening to live sound. For example, in the concert hall, there are two major acoustical phenomena—namely, the auditory perspective and the reverberation envelope. The listener localizes the different sound sources and attributes their proper location. The listener also hears the multireflected sounds as the reverberation envelope. Thus, quadraphonic reproduction provides realism in all respects; that is, it satisfies all four conditions.

Transducers for reproducing the four channels separately and discretely include the four-channel magnetic-tape reproducer and the four-channel disc phonograph.

Systems have also been designed for matrixing and encoding from four channels to two channels and the complementary decoding and matrixing from two channels to four channels. In these systems, quadraphonic sound can be reproduced over any two-channel system. The performance with respect to auditory perspective and acoustic ambience effects, however, is inferior to that of the discrete four-channel system because of the relatively high cross talk between channels; that is, spill-over from one channel to another (see also ACOUSTICAL ENGINEERING).

(H.F.O.)

BIBLIOGRAPHY. H.F. OLSON, *Modern Sound Reproduction* (1972), provides a survey of the latest developments including quadraphonic sound. The *McGraw-Hill Encyclopedia of Science and Technology* (1971), includes information on all methods of recording. On the relationship of music and musical instruments to sound recording and reproducing, see H.F. OLSON, *Music, Physics and Engineering*, 2nd ed. (1967). The first chapter and bibliography of S.J. BEGUN, *Magnetic Recording* (1949), comprehensively cover the early history of magnetic recording. Two reference texts are H.G.M. SPRATT, *Magnetic Tape Recording*, 2nd ed. (1964); and W. EARL STEWART, *Magnetic Recording Techniques* (1958), with mathematical derivations in the appendixes and references to original sources. ROLAND GELATT, *The Fabulous Phonograph: From Edison to Stereo*, rev. ed. (1966), is a popularly written but comprehensive history of disc sound recording and reproduction through the stereo disc record; JOHN G. FRAYNE and HALLEY WOLFE, *Elements of Sound Recording* (1949), is a more technical treatise on sound recording and reproduction. See also H.M. TREMAINE, *Audio Cyclopedia*, 2nd ed. (1969); JAMES MOIR, *High Quality Sound Reproduction* (1958); and H.F. OLSON, *Acoustical Engineering* (1957).

(H.F.O./H.E.Ro./W.E.St./A.M.Ma.)

South Africa

The Republic of South Africa, the southernmost state on the African continent, has an area of 471,445 square miles (1,221,037 square kilometres). Measuring nearly 1,000 miles from north to south, as well as from east to west, it forms a fairly compact political and geographic unit; a variety of landscapes, climates, and resources are to be found within its borders. Of its four provinces, Cape of Good Hope Province, with an area of 278,380 square miles, is the largest, followed by the Transvaal (109,621 square miles), the Orange Free State (49,866 square miles), and Natal (33,578 square miles). The administrative capital is Pretoria, and the legislative capital is Cape Town.

South West Africa (*q.v.*), also known as Namibia, which is situated to the northwest and has an area of 318,261 square miles, is administered by South Africa as a part of the republic. The territory, once a German colony, was mandated to South Africa by the League of Nations; the United Nations, however, challenges South Africa's continuing administration of it on the ground that the League's mandate has been terminated. The only useful harbour of South West Africa, Walvis Bay, had become part of the Cape Province (then under British rule) during the latter part of the 19th century; later it became part of the Union of South Africa and is thus now politically part of the Republic of South Africa. South Africa also possesses two small Subantarctic islands, Prince Edward and Marion, situated in the Indian Ocean about 1,200 miles to the southeast of Cape Town.

South Africa is bordered by South West Africa to the northwest, by Botswana and Rhodesia to the north, by Mozambique and Swaziland to the east, by the Indian Ocean to the southeast and south, and by the Atlantic Ocean to the south and west. The independent state of Lesotho (*q.v.*), in the eastern part of the republic, is entirely surrounded by South African territory.

South Africa's location almost entirely to the south of the Tropic of Capricorn and within temperate climatic zones has resulted in European settlement on a scale unknown elsewhere in Africa. Although white-administered South Africa's political affiliations and most of its economic links are with the Western bloc of nations, its position—3,800 miles from South America to the west, 4,700 miles from Australia to the east, and about 6,000 miles from the markets of western Europe—is a relatively isolated one. Events in the Middle East resulting in the periodic closing of the Suez Canal have, however, emphasized the strategic location of South Africa and the importance of the sea route round the Cape of Good Hope.

South Africa's strategic location

Since World War II, South Africa has been a frequent focus of world attention. Whereas many African states have in recent decades moved from colonial to independent status, and whereas guerrilla warfare has developed in neighbouring Portuguese African territories, in South Africa there has evolved the tradition of apartheid (meaning literally "apartness"), or separation of the races, which has evoked vehement opposition from most countries in the United Nations organization. Racially, South Africa's peoples comprise a diversity of ethnic groups. Out of a population of over 21,000,000, about one-fifth are white, over two-thirds Bantu, one-tenth Coloured (of mixed descent), and the remainder Asian. (See the city articles CAPE TOWN; JOHANNESBURG, as well as articles on the individual provinces; for associated physical features, see KALAHARI; ORANGE RIVER; and VELD; and for important historical aspects, see SOUTHERN AFRICA, HISTORY OF.)

THE LANDSCAPE

The natural environment. *Relief.* Geologically, South Africa, which contains some of the oldest rocks in the world, is a stable country. Earthquakes rarely occur, and there has been no volcanic activity for about 50,000,000 years. Dominating the topography is a plateau that covers the largest part of the country and is separated from the small coastal strip by what is called the Great Escarpment. The plateau is actually a massive basin reaching heights of 8,000 feet in the basaltic Lesotho region, dropping down to heights of approximately 2,000 to 3,000 feet in the sandy Kalahari Plateau in the west. The central part of the plateau comprises the Transvaal and Free State Highveld, which is between 4,000 and 6,000 feet in altitude. The Witwatersrand Ridge forms the watershed between the Vaal and Limpopo rivers and their tributaries. The Great Karroo Plateau, south of the Orange River, averaging 4,000 feet in height, contains a number of large salt pans such as Verneukpan. For the most part, the plateau landscape is a rather monotonous plain, alternately flat or undulating.

The Great Escarpment, known by a variety of local names, is the most continuous topographical feature in South Africa. Starting in the far northeast, at altitudes of up to 7,000 feet, it is known as the Transvaal Drakensberg ("berg" meaning mountains). Continuing south-

The Great Escarpment

ward, the mountains become known as the Natal Drakensberg and reach their greatest heights, up to 11,000 feet. (The highest point in the whole southern subcontinent, Thabana Ntlenyana [11,425 feet], however, does not form part of the Drakensberg itself but is an isolated plateau peak within the borders of Lesotho.) Further to the south, the mountains become less impressive, ranging from 8,000 feet in the Sneeuberg section to 6,000 feet in the Nuweveld and to 5,000 feet in the Roggeveldberge mountain region. In the west, the escarpment comprises the Bokkeveldberge Mountains and the Kamiesberg (5,600 feet), between which both the plateau and the escarpment are not clearly defined. In the far south, but in no way forming part of the escarpment, is an area of folded mountains that includes ranges such as the Hottentots Hollandsberge, Drakensteinberg, Sederberg, Langeberg, Great Swartberg, Outeniqua, Tsitsikama, and the Hex River mountain complex, all of which vary in height between 5,000 and 7,600 feet and are geologically older than either the Alps or the Himalayas.

The narrow South African coastal plain is poorly developed; it is less than 500 feet above sea level, and only in northern Natal does it reach a mentionable width—extending about 30 miles along the west coast and broadening to about 80 miles in northern Natal. Neither is the coastal strip continuous, being interrupted by such mountain ranges as the Hottentots Hollandsberge, which extends to the sea. South Africa has a straight, unindented, and somewhat monotonous shoreline; the only worthwhile natural harbour along its 1,836-mile length is at Saldanha Bay, the use of which has been restricted by a lack of fresh water.

Drainage. Because of insufficient and inadequate rainfall, South Africa has no rivers that form navigable waterways, even for short distances. The main drainage system is that of the Orange River (*q.v.*), which drains an area of approximately 329,000 square miles; the Vaal and Caledon rivers are the main tributaries. The northern part of the plateau is drained by the Limpopo system, which empties into the Indian Ocean; the Limpopo's main tributaries are the Olifants, Marico, and Sand rivers. From the escarpment a number of smaller and shorter rivers, all of them irregular, flow seaward.

Soils. Three major soil regions may be distinguished. The soil region east of the 20-inch isohyet (*i.e.*, a line drawn on a map connecting points having equal rainfall) includes soils formed under conditions of a wet summer and dry winter climate, the more important types being laterite (red, leached, iron-bearing soil) and lateritic soils, unleached subtropical soils, and gleylike (*i.e.*, bluish-gray, sticky, and compact) podsolic soils (highly leached soils that are low in iron and lime). The second major soil region lies within the winter rainfall zone and the southern coastal area with all-season rainfall; gray sandy soils and sandy loams form the main types. West of the approximately 15- to 20-inch isohyet—an area that includes a large part of the interior and all of the western coastal area—the land is semidesert and desert, with rain falling primarily in summer and with a huge annual evaporation occurring. The soils of this region as a whole are characterized by a top sandy layer, often a sandy loam, which is underlain by a layer of lime or an accretion of silica.

Climate. The climate is controlled by three main factors.

Firstly, South Africa's location between latitudes 22° and 35° S places it—except for a very small part in the north—fully within the temperate zone, though it receives a great deal of sunshine. Apart from the southwestern Cape Province, South Africa is out of reach of the rainy westerlies and next to a subtropical high-pressure belt of descending air where condensation and rainfall do not readily develop. The greater part of South Africa thus has a dry climate.

Secondly, South Africa is flanked on the east by the warm, southward-flowing Agulhas Current, and along the west coast by the cold, north-flowing Benguela Current. This not only makes the east coast comparatively warmer but also makes the warmer and less dense air along the east coast rise more readily to facilitate the entry of rain-bringing winds.

Lastly, South Africa's inland plateau greatly tempers the influence of latitude on climate. From Cape Town in the south to the far north, the temperature hardly changes, the increasing intensity of the solar rays being moderated by the increase in altitude.

Influences of the subtropical high-pressure belt

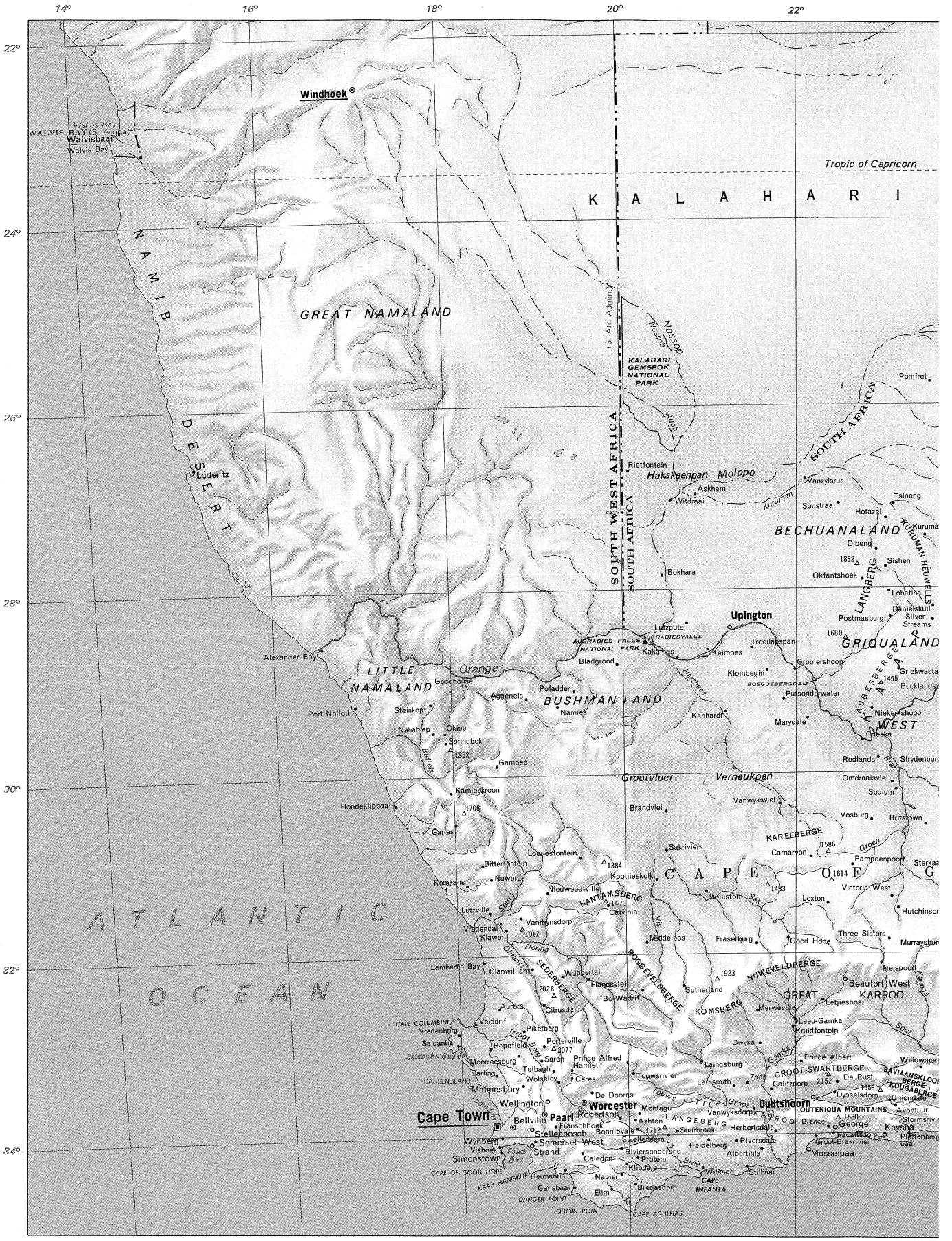
Because South Africa finds itself in the regime of the subtropical high-pressure belt, the country, as well as the adjoining ocean areas, is influenced all year around by descending, divergent upper air masses that circulate primarily eastward, generally causing fine weather and a low annual precipitation, especially to the west. Seasonal and other air movements, especially of the Atlantic high-pressure system, result in polar air masses reaching the southern part of the country and tropical air masses reaching the northern parts. During winter (from June to August), cold polar air moves over the southwestern, southern, and southeastern coastal areas, sometimes invading the southern interior part of the country from the southwest. These polar air masses are accompanied by cold fronts as well as by rain and snow. In summer (from December to February) the Atlantic high-pressure system settles semipermanently over the southern and western parts of the country. Local heating of the landmass sometimes causes low-pressure conditions to develop, resulting in tropical air masses being drawn in from the Indian Ocean over the northeastern region, bringing rain.

South Africa is a generally semiarid country where farmers are constantly faced with the problem of a water shortage. About 21 percent of the republic is arid, receiving less than 8 inches of rainfall annually; another 47 percent is semiarid, receiving 8 to 24 inches annually. Abundant rain, amounting to more than 40 inches annually, is received over only 6 percent of the country. From east to west there is a gradual decline in the rainfall. Natal receives more than 40 inches annually, Kimberley approximately 16 inches, and Alexander Bay on the west coast less than two inches. About 87 percent of the country—the inland plateau and east coast region—receives most of its rain in summer, nine percent receives most of its rain in winter (the southwestern part of Cape Province), and a mere four percent (the narrow southern strip) has its rainfall spread over the year.

Temperatures are generally moderate. The summers are moderately warm, with temperatures of between 70° and 75° F (21° and 24° C). Over the western plateau, summer temperatures are from 75° to 85° F (24° to 29° C); in the Orange River Valley it becomes very hot, with the temperatures rising to 90° F (32° C) and higher. South African winters are moderately cool; over the inland plateau July average temperatures fall below 50° F (10° C), and, in places, to less than 40° F (4° C). Along the east coast, in Natal and in the Transvaal Lowveld, even the winters are moderately warm. Influenced by the ocean currents, the temperature also declines from the east coast to the west coast. Durban, on the east coast, has an average annual temperature of 69° F (21° C), while Port Nolloth—on the west coast but approximately on the same latitude—registers 57° F (14° C). Daily fluctuations in temperature are more pronounced on the plateau and often exceed the seasonal variation; Harrismith in the Orange Free State, for example, has a seasonal fluctuation of 54° F (30° C) in range and a daily fluctuation of 61° F (34° C) in range.

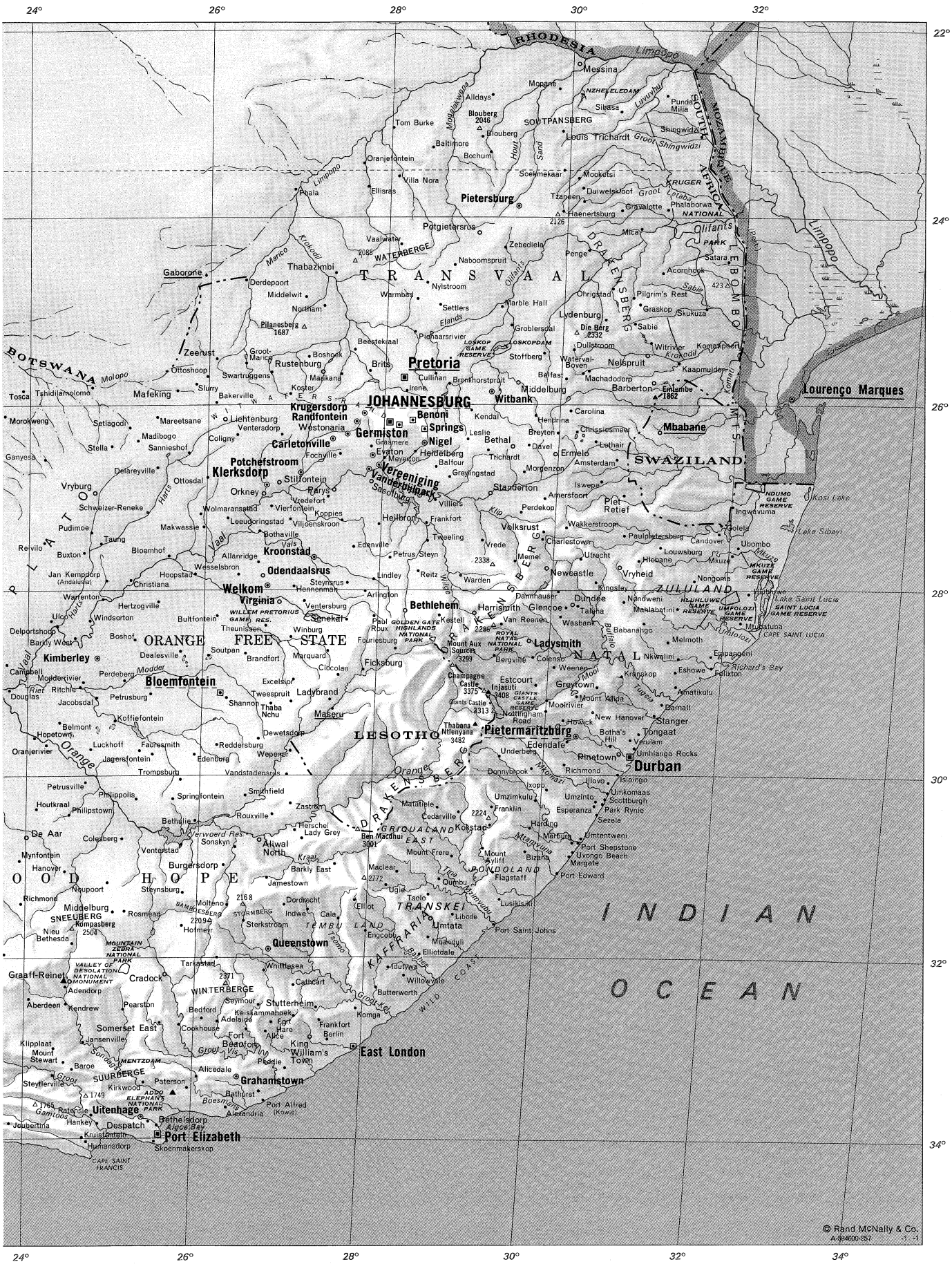
Vegetation and animal life. In a semiarid country, forests worthy of the name rarely occur; the small Knysna Forest in the southern Cape Province is perhaps the best example. Shrublike vegetation is found in the Karroo region, monotonous grassland with very few trees occurs on the Highveld; and savanna (parklike grassland) occurs in a variety of areas such as the Transvaal Bushveld and Lowveld, the Kalahari, the deep valleys of the eastern Cape, and the Natal coastal region.

When the first white settlers arrived in the 17th century, South Africa had an unbelievable wealth of animal life, including lions, elephants, rhinoceroses, and a variety of antelope types; today, however, such wildlife exists only in the most marginal areas, such as the Kalahari. To



SOUTH AFRICA Size of symbol indicates relative size of town • ○ ◉ ▣ ■ ■

Elevations in metres


24°
26°
28°
30°
32°
22°
24°
26°
28°
30°
32°
34°
RHODESIA
Limpopo
Messina
Nzhelele Dam
Luvuvhu
Pafuri
Mozambique
South Africa
Kruger National Park
Phalaborwa
Olifants
Letaba
Groot Letaba
Tzaneen
Duiwelskloof
Haenertsburg
2126
Gravelotte
Mica
Olifants Park
Lebombo Mts
Pietersburg
Potgietersrus
Vaalwater
2088
Waterberge
Zebediela
Penge
Drakensberg
Satara
Acornhoek
Sabie
423
Graskop
Skukuza
Pilgrim's Rest
Ohrigstad
Lydenburg
Die Berg
2332
Sabie
Witrivier
Komatipoort
Nelspruit
Machadodorp
Kaapmuiden
Barberton
Emlembe
1862
Lourenço Marques
Mbabane
Swaziland
Ermelo
Amsterdam
Carolina
Chrissiesmeer
Breyten
Davel
Lothair
Hendrina
Iswepe
Amersfoort
Piet Retief
Ingwavuma
Ndumo Game Reserve
Kosi Lake
Lake Sibayi
Zululand
Hluhluwe Game Reserve
Umfolozi Game Reserve
Saint Lucia Game Reserve
Mkuze Game Reserve
Hluhluwe
Mkuze
Nongoma
Vryheid
Hlobane
Louwsburg
Candover
Ubombo
Paulpietersburg
Wakkerstroom
Volksrust
Charlestown
Newcastle
Utrecht
Memel
2338
Dundee
Glencoe
Talana
Dannhauser
Ladysmith
Colenso
Weenen
Estcourt
Injasuti
3375
Giants Castle
3408
Giants Castle Game Reserve
3313
Nottingham Road
Howick
Mooirivier
Greytown
Mount Africa
Kranskop
Eshowe
Felixton
Richard's Bay
Empangeni
Nkwalini
Melmoth
Babanango
Wasbank
Mahlabatini
Nondweni
Tugela
Mooi
Buffels
Natal
Pietermaritzburg
Thabana Nlenyana
3482
Lesotho
Underberg
Edendale
Pinetown
Umhlanga Rocks
Durban
Richmond
Donnybrook
Ixopo
Umkomaas
Scottburgh
Park Rynie
Sezela
Umzinto
Umzimkulu
Franklin
Esperanza
Harding
Umtentweni
Port Shepstone
Uvongo Beach
Margate
Port Edward
Mtamvuna
Mtwalume
Mtamvuna
Pondoland
Flagstaff
Lusikisiki
Transkei
Umtata
Libode
Port Saint Johns
Indian Ocean
Kokstad
Matatiele
Cedarville
2224
Ben Macdhui
3001
Drakensberg
Griqualand East
Mount Frere
Mount Ayliff
Bizana
Maclear
2772
Ugie
Qumbu
Tsolo
Elliot
Cala
Indwe
Dordrecht
2168
Stormberg
Molteno
Barkly East
Jamestown
Rhodes
Lady Grey
Herschel
Zastron
Rouxville
Smithfield
Springfontein
Philippolis
Petrusville
Trompsburg
Vandstadensrus
Edenburg
Wepener
Dewetsdorp
Thaba Nchu
Masetu
Tweespruit
Ladybrand
Excelsior
Clocolan
Ficksburg
Fouriesburg
Golden Gate Highlands National Park
Roux
Pabl
Senekal
Winburg
Theunissen
Willem Pretorius Game Res.
Virginia
Ventersburg
Bethlehem
Harrismith
Van Reenen
Wasbank
Kestell
2286
Royal Natal National Park
Mount Aux Sources
3299
Bergville
Drakensberg
Orange Free State
Kroonstad
Allanridge
Vredefort
Parys
Sasolburg
Vaal
Vals
Vaal Dam
Villiers
Frankfort
Heilbron
Tweeling
Vrede
Edenville
Petrus Steyn
Reitz
Lindley
Warden
Wilge
Klip
Perdekop
Standerton
Greylingstad
Balfour
Trichardt
Morgenzon
Leslie
Bethal
Nigel
Heidelberg
Evaton
Meyerton
Vereeniging
Vanderbijlpark
Sharpeville
Sebokeng
Germiston
Benoni
Springs
Kendal
Witbank
Middelburg
Belfast
Johannesburg
Krugersdorp
Randfontein
Westonaria
Carletonville
Fochville
Potchefstroom
Klerksdorp
Stilfontein
Orkney
Hartebeestfontein
Coligny
Lichtenburg
Ventersdorp
Madibogo
Sannieshof
Mareetsane
Setlagodi
Morokweng
Stella
Ganyesa
Delareyville
Ottosdal
Vryburg
Schweizer-Reneke
Wolmaransstad
Leeudoringstad
Viljoenskroon
Koppies
Vredefort
Vierfontein
Bothaville
Makwassle
Pudimoe
Taung
Reivilo
Buxton
Bloemhof
Hoopstad
Wesselsbron
Odendaalsrus
Allanridge
Jan Kempdorp
(Andalusia)
Christiana
Welkom
Hennenman
Steynsrus
Arlington
Warrenton
Hertzogville
Ulco
Windsorton
Boshof
Delportshoop
Barkly West
Kimberley
Dealesville
Campbell
Perdeberg
Modderrivier
Ritchie
Douglas
Jacobsdal
Petrusburg
Bloemfontein
Bultfontein
Theunissen
Winburg
Senekal
Pabl
Roux
Golden Gate Highlands National Park
Fouriesburg
Boshof
Orange Free State
Soutpan
Brandfort
Marquard
Dewetsdorp
Thaba Nchu
Masetu
Tweespruit
Ladybrand
Excelsior
Clocolan
Ficksburg
Kimberley
Modder
Riet
Orange
Vaal
Harts
Molopo
Botswana
Mafeking
Slurry
Bakerville
Koster
Swartruggens
Rustenburg
Boshoek
Maanhana
Brits
Pretoria
Cullinan
Bronkhorstspruit
Irene
Zeerust
Groot-Marico
Derdepoort
Middelwit
Northam
Thabazimbi
Gaborone
Tosca
Tshidilamolomo
Ottoshoop
Morokweng
Setlagodi
Mareetsane
Madibogo
Sannieshof
Stella
Ganyesa
Delareyville
Ottosdal
Vryburg
Schweizer-Reneke
Reivilo
Pudimoe
Taung
Makwassle
Buxton
Bloemhof
Hoopstad
Wesselsbron
Odendaalsrus
Allanridge
Jan Kempdorp
(Andalusia)
Christiana
Welkom
Hennenman
Steynsrus
Arlington
Warrenton
Hertzogville
Ulco
Windsorton
Boshof
Delportshoop
Barkly West
Kimberley
Dealesville
Campbell
Perdeberg
Modderrivier
Ritchie
Douglas
Jacobsdal
Petrusburg
Bloemfontein
Bultfontein
Theunissen
Winburg
Senekal
Pabl
Roux
Golden Gate Highlands National Park
Fouriesburg
Boshof
Orange Free State
Soutpan
Brandfort
Marquard
Dewetsdorp
Thaba Nchu
Masetu
Tweespruit
Ladybrand
Excelsior
Clocolan
Ficksburg
Kimberley
Modder
Riet
Orange
Vaal
Harts
Molopo
Botswana
Mafeking
Slurry
Bakerville
Koster
Swartruggens
Rustenburg
Boshoek
Maanhana
Brits
Pretoria
Cullinan
Bronkhorstspruit
Irene
Zeerust
Groot-Marico
Derdepoort
Middelwit
Northam
Thabazimbi
Gaborone
Tosca
Tshidilamolomo
Ottoshoop
Pilanesberg
1687
Beestekraal
Pienaarsrivier
Loskop Game Reserve
Loskopdam
Stoffberg
Groblersdal
Marble Hall
Settlers
Warmbad
Nylstroom
Naboomspruit
Vaalwater
2088
Waterberge
Zebediela
Penge
Drakensberg
Lydenburg
Die Berg
2332
Sabie
Witrivier
Komatipoort
Nelspruit
Machadodorp
Kaapmuiden
Barberton
Emlembe
1862
Lourenço Marques
Mbabane
Swaziland
Ermelo
Amsterdam
Carolina
Chrissiesmeer
Breyten
Davel
Lothair
Hendrina
Iswepe
Amersfoort
Piet Retief
Ingwavuma
Ndumo Game Reserve
Kosi Lake
Lake Sibayi
Zululand
Hluhluwe Game Reserve
Umfolozi Game Reserve
Saint Lucia Game Reserve
Mkuze Game Reserve
Hluhluwe
Mkuze
Nongoma
Vryheid
Hlobane
Louwsburg
Candover
Ubombo
Paulpietersburg
Wakkerstroom
Volksrust
Charlestown
Newcastle
Utrecht
Memel
2338
Dundee
Glencoe
Talana
Dannhauser
Ladysmith
Colenso
Weenen
Estcourt
Injasuti
3375
Giants Castle
3408
Giants Castle Game Reserve
3313
Nottingham Road
Howick
Mooirivier
Greytown
Mount Africa
Kranskop
Eshowe
Felixton
Richard's Bay
Empangeni
Nkwalini
Melmoth
Babanango
Wasbank
Mahlabatini
Nondweni
Tugela
Mooi
Buffels
Natal
Pietermaritzburg
Thabana Nlenyana
3482
Lesotho
Underberg
Edendale
Pinetown
Umhlanga Rocks
Durban
Richmond
Donnybrook
Ixopo
Umkomaas
Scottburgh
Park Rynie
Sezela
Umzinto
Umzimkulu
Franklin
Esperanza
Harding
Umtentweni
Port Shepstone
Uvongo Beach
Margate
Port Edward
Mtamvuna
Mtwalume
Mtamvuna
Pondoland
Flagstaff
Lusikisiki
Transkei
Umtata
Libode
Port Saint Johns
Indian Ocean
Kokstad
Matatiele
Cedarville
2224
Ben Macdhui
3001
Drakensberg
Griqualand East
Mount Frere
Mount Ayliff
Bizana
Maclear
2772
Ugie
Qumbu
Tsolo
Elliot
Cala
Indwe
Dordrecht
2168
Stormberg
Molteno
Barkly East
Jamestown
Rhodes
Lady Grey
Herschel
Zastron
Rouxville
Smithfield
Springfontein
Philippolis
Petrusville
Trompsburg
Vandstadensrus
Edenburg
Wepener
Dewetsdorp
Thaba Nchu
Masetu
Tweespruit
Ladybrand
Excelsior
Clocolan
Ficksburg
Fouriesburg
Golden Gate Highlands National Park
Roux
Pabl
Senekal
Winburg
Theunissen
Willem Pretorius Game Res.
Virginia
Ventersburg
Bethlehem
Harrismith
Van Reenen
Wasbank
Kestell
2286
Royal Natal National Park
Mount Aux Sources
3299
Bergville
Drakensberg
Orange Free State
Kroonstad
Allanridge
Vredefort
Parys
Sasolburg
Vaal
Vals
Vaal Dam
Villiers
Frankfort
Heilbron
Tweeling
Vrede
Edenville
Petrus Steyn
Reitz
Lindley
Warden
Wilge
Klip
Perdekop
Standerton
Greylingstad
Balfour
Trichardt
Morgenzon
Leslie
Bethal
Nigel
Heidelberg
Evaton
Meyerton
Vereeniging
Vanderbijlpark
Sharpeville
Sebokeng
Germiston
Benoni
Springs
Kendal
Witbank
Middelburg
Belfast
Johannesburg
Krugersdorp
Randfontein
Westonaria
Carletonville
Fochville
Potchefstroom
Klerksdorp
Stilfontein
Orkney
Hartebeestfontein
Coligny
Lichtenburg
Ventersdorp
Madibogo
Sannieshof
Mareetsane
Setlagodi
Morokweng
Stella
Ganyesa
Delareyville
Ottosdal
Vryburg
Schweizer-Reneke
Wolmaransstad
Leeudoringstad
Viljoenskroon
Koppies
Vredefort
Vierfontein
Bothaville
Makwassle
Pudimoe
Taung
Reivilo

MAP INDEX

Political subdivisions

Cape of Good Hope 31·00s 23·00e
Natal 28·40s 30·40e
Orange Free State 28·30s 27·00e
Transvaal 25·00s 29·00e

Cities and towns

Aberdeen 32·29s 24·04e
Acornhoek 24·37s 31·02e
Adelaide 32·42s 26·20e
Adendorp 32·20s 24·33e
Aggeneis 29·03s 18·51e
Albertinia 34·13s 21·36e
Alexander Bay 28·40s 16·30e
Alexandria 33·39s 26·25e
Alicedale 33·19s 26·05e
Aliwal North 30·45s 26·45e
Allanridge 27·55s 26·44e
Alldays 22·44s 29·04e
Amatikulu 29·06s 31·27e
Amersfoort 27·00s 29·53e
Amsterdam 26·35s 30·45e
Andalusia, see Jan Kempdorf
Arlington 28·06s 27·54e
Ashton 33·50s 20·05e
Askham 26·59s 20·47e
Aurora 32·42s 18·29e
Avontuur 33·44s 23·11e
Babanango 28·30s 31·00e
Bakerville 26·00s 26·06e
Balfour 26·44s 28·45e
Baltimore 23·15s 28·20e
Barberton 25·48s 31·03e
Barkly East 30·58s 27·33e
Barkly West 28·05s 24·31e
Baroe 33·13s 24·33e
Bathurst 33·30s 26·50e
Beaufort West 32·18s 22·36e
Bedford 32·41s 26·05e
Beestekraal 25·23s 27·38e
Belfast 25·43s 30·03e
Bellville 33·53s 18·36e
Belmont 29·28s 24·22e
Benoni 26·19s 28·27e
Bergville 28·52s 29·18e
Berlin 32·54s 27·35e
Bethal 26·27s 29·28e
Bethelsdorp 33·52s 25·34e
Bethlehem 28·15s 28·15e
Bethulie 30·32s 25·59e
Bitterfontein 31·00s 18·32e
Bizana 30·58s 29·52e
Bladgrond 28·52s 19·57e
Blanco 33·57s 22·24e
Bloemfontein 29·12s 26·07e
Bloemhof 27·38s 25·32e
Blouberg 23·08s 28·56e
Bochum 23·17s 29·07e
Bokhara 27·57s 20·30e
Bonnievale 33·57s 20·06e
Boshoek 25·30s 27·09e
Boshof 28·34s 25·04e
Botha's Hill 29·45s 30·45e
Bothaville 27·27s 26·36e
Bo-Wadrif 32·26s 20·07e
Brandfort 28·47s 26·30e
Brandvlei 30·25s 20·30e
Bredasdorp 34·32s 20·02e
Breyten 26·16s 30·00e
Brits 25·42s 27·45e
Britstown 30·37s 23·30e
Bronkhorstspruit 25·50s 28·45e
Bucklands 29·03s 23·44e
Bultfontein 28·20s 26·05e
Burgersdorp 31·00s 26·20e
Butterworth 32·23s 28·04e
Buxton 27·38s 24·42e
Cala 31·30s 27·37e
Caledon 34·12s 19·23e
Calitzdorp 33·33s 21·42e
Calvinia 31·25s 19·45e
Campbell 28·48s 23·44e
Candover 27·28s 31·57e
Cape Town (Kaapstad) 33·55s 18·22e
Carletonville 26·23s 27·22e
Carnarvon 30·56s 22·08e
Carolina 26·05s 30·06e
Cathcart 32·18s 27·09e
Cedarville 30·23s 29·03e
Ceres 33·21s 19·18e
Charlestown 27·30s 29·55e
Chrissiesmeer 26·16s 30·13e
Christiana 27·52s 25·08e
Citrusdal 32·36s 19·00e
Clanwilliam 32·11s 18·54e
Clocolan 29·00s 27·30e
Colenso 28·50s 29·44e
Colesberg 30·45s 25·05e
Coligny 26·17s 26·15e
Cookhouse 32·44s 25·48e
Cradock 32·08s 25·36e
Cullinan 25·40s 28·32e
Danielskuil 28·11s 23·33e
Dannhauser 28·04s 30·04e
Darling 33·23s 18·23e
Darnall 29·23s 31·18e
Davel 26·24s 29·40e
De Aar 30·39s 24·00e
Dealesville 28·40s 25·37e
De Doorns 33·28s 19·41e
Delareyville 26·44s 25·29e
Delportshoop 28·22s 24·20e
Derdepoort 24·42s 26·20e
De Rust 33·30s 22·32e
Despatch 33·46s 25·30e
Dewetsdorp 29·33s 26·34e
Dibeng 27·35s 22·54e
Donnybrook 30·00s 29·48e
Dordrecht 31·20s 27·03e
Douglas 29·04s 23·46e
Duiwelskloof 23·42s 30·06e
Dullstroom 25·27s 30·07e
Dundee 28·12s 30·16e
Durban 29·55s 30·56e
Dwyka 33·02s 21·30e
Dysselsdorp 33·34s 22·28e
East London 33·00s 27·55e
Edenburg 29·45s 25·56e
Edendale 29·39s 30·18e
Edenville 27·37s 27·34e
Elandsvlei 32·19s 19·33e
Elim 34·35s 19·45e
Elliot 31·18s 27·50e
Elliotdale 31·55s 28·38e
Ellisras 23·40s 27·46e
Empangeni 28·50s 31·48e
Engcobo 31·37s 28·00e
Ermelo 26·34s 29·58e
Eshowe 28·58s 31·29e
Esperanza 30·21s 30·40e
Estcourt 29·01s 29·52e
Evaton 26·31s 27·54e
Excelsior 28·56s 27·06e
Fauresmith 29·42s 25·21e
Felixton 28·50s 31·53e
Flagstaff 31·05s 29·29e
Fochville 26·30s 27·30e
Fort Beaufort 32·46s 26·40e
Fort Hare 32·47s 26·50e
Fouriesburg 28·38s 28·14e
Frankfort 27·17s 28·30e
Frankfort 32·44s 27·27e
Franklin 30·18s 29·30e
Franschhoek 33·55s 19·09e
Fraserburg 31·55s 21·30e
Gamoep 29·55s 18·25e
Gansbaai 34·35s 19·22e
Ganyesa 26·35s 24·10e
Garies 30·30s 18·00e
George 33·58s 22·24e
Germiston 26·15s 28·05e
Glencoe 28·12s 30·07e
Golela 27·20s 31·55e
Good Hope 31·51s 21·55e
Goodhouse 28·57s 18·13e
Graaff-Reinet 32·14s 24·32e
Grahamstown 33·19s 26·31e
Graskop 24·58s 30·49e
Grasmere 26·26s 27·52e
Gravelotte 23·56s 30·34e
Greylingstad 26·44s 28·45e
Greytown 29·07s 30·30e
Griekwastad 28·49s 23·15e
Groblersdal 25·15s 29·25e
Groblershoop 28·55s 20·59e
Groot-Brakrivier 34·01s 21·46e
Groot-Marico 25·37s 26·26e
Haenertsburg 24·00s 29·50e
Hankey 33·50s 24·52e
Hanover 31·04s 24·29e
Harding 30·34s 29·58e
Harrismith 28·18s 29·03e
Heidelberg 26·32s 28·18e
Heidelberg 34·06s 20·59e
Heilbron 27·21s 27·58e
Hendrina 26·11s 29·45e
Hennenman 27·59s 27·01e
Herbertsdale 34·01s 21·45e
Hermanus 34·25s 19·16e
Herschel 30·37s 27·12e
Hertzogville 28·08s 25·33e
Hlobane 27·42s 31·00e
Hluhluwe 28·01s 32·15e
Hofmeyr 31·39s 25·50e
Hondeklipbaai 30·20s 17·18e
Hoopstad 27·54s 25·58e
Hopefield 33·04s 18·22e
Hopetown 29·34s 24·03e
Hotazel 27·15s 23·00e
Houtkraal 30·23s 24·05e
Howick 29·28s 30·14e
Humansdorp 34·02s 24·46e
Hutchinson 31·30s 23·09e
Idutywa 32·02s 28·16e
Illovo 30·05s 30·50e
Indwe 31·27s 27·23e
Ingwavuma 27·09s 32·00e
Irene 25·51s 28·13e
Isipingo 30·00s 30·57e
Iswepe 26·50s 30·31e
Ixopo 30·08s 30·00e
Jacobsdal 29·13s 24·41e
Jagersfontein 29·44s 25·29e
Jamestown 31·06s 26·45e
Jan Kempdorp (Andalusia) 27·55s 24·51e
Jansenville 32·57s 24·40e
Johannesburg 26·12s 28·02e
Joubertina 33·50s 23·51e
Kaapmuiden 25·33s 31·20e
Kaapstad, see Cape Town
Kakamas 28·45s 20·33e
Kamieskroon 30·09s 17·56e
Keimoes 28·41s 21·00e
Keiskammahoek 32·41s 27·09e
Kendal 26·04s 28·58e
Kendrew 32·31s 24·30e
Kenhardt 29·19s 21·12e
Kestell 28·19s 28·38e
Kimberley 28·43s 24·46e
Kingsley 27·55s 30·33e
King William's Town 32·51s 27·22e
Kirkwood 33·24s 25·26e
Klawer 31·44s 18·36e
Kleinbegin 28·50s 21·36e
Klerksdorp 26·58s 26·39e
Klipdale 34·19s 19·57e
Klipplaat 33·02s 24·21e
Knysna 34·02s 23·02e
Koffiefontein 29·30s 25·00e
Kokstad 30·32s 29·29e
Komatipoort 25·25s 31·55e
Komga 32·35s 27·55e
Komkans 31·16s 18·09e
Kootjieskolk 31·15s 20·21e
Koppies 27·20s 27·30e
Koster 25·57s 26·42e
Kowie, see Port Alfred
Kranskop 29·00s 30·47e
Kroonstad 27·46s 27·12e
Krugersdorp 26·05s 27·35e
Kruidfontein 32·51s 21·57e
Kruisfontein 34·00s 24·43e
Kuruman 27·28s 23·28e
Ladismith 33·30s 21·16e
Ladybrand 29·19s 27·25e
Lady Grey 30·45s 27·13e
Ladysmith 28·34s 29·45e
Laingsburg 33·11s 20·51e
Lambert's Bay 32·05s 18·17e
Leeudoringstad 27·15s 26·10e
Leeu-Gamka 32·47s 21·59e
Leslie 26·27s 28·55e
Letjiesbos 32·34s 22·16e
Libode 31·33s 29·02e
Lichtenburg 26·08s 26·08e
Lindley 28·00s 27·57e
Loeriesfontein 30·56s 19·26e
Lohatlha 28·02s 23·04e
Lothair 26·26s 30·27e
Louis Trichardt 23·01s 29·43e
Louwsburg 27·37s 31·07e
Loxton 31·30s 22·22e
Luckhoff 29·44s 24·43e
Lusikisiki 31·25s 29·30e
Lutzputs 28·03s 20·40e
Lutzville 31·33s 18·22e
Lydenburg 25·10s 30·29e
Machadodorp 25·40s 30·14e
Maclear 31·02s 28·23e
Madibogo 26·25s 25·10e
Mafeking 25·53s 25·39e
Mahlabatini 28·14s 31·30e
Makwassie 27·26s 26·00e
Malmesbury 33·28s 18·44e
Marble Hall 24·57s 29·13e
Marburg 30·44s 30·26e
Mareetsane 26·09s 25·25e
Margate 30·55s 30·15e
Marikana 25·42s 27·30e
Marquard 28·54s 27·28e
Marydale 29·23s 22·05e
Matatiele 30·24s 28·43e
Melmoth 28·38s 31·24e
Memel 27·43s 29·30e
Merweville 32·40s 21·31e
Messina 22·23s 30·00e
Meyerton 26·33s 28·01e
Mica 24·10s 30·48e
Middelburg 25·47s 29·28e
Middelburg 31·30s 25·00e
Middelpos 31·55s 20·13e
Middelwit 24·58s 27·00e
Mkuze 27·10s 32·00e
Modderrivier 29·02s 24·38e
Molteno 31·22s 26·22e
Montagu 33·45s 20·08e
Mooirivier 29·13s 29·50e
Mooketsi 23·35s 30·05e
Moorreesburg 33·08s 18·40e
Mopane 22·37s 29·52e
Morgenzon 26·45s 29·36e
Morokweng 26·12s 23·45e
Mosselbaai (Mossel Bay) 34·11s 22·08e
Mount Alida 29·09s 30·18e
Mountayliff 30·54s 29·20e
Mount Frere 31·00s 28·58e
Mount Stewart 33·10s 24·26e
Mqanduli 31·48s 28·46e
Mtubatuba 28·30s 32·08e
Murraysburg 31·58s 23·47e
Mynfontein 30·55s 23·57e
Nababiep 29·36s 17·46e
Naboomspruit 24·32s 28·36e
Namies 29·18s 19·13e
Napier 34·29s 19·53e
Nelspoort 32·07s 23·00e
Nelspruit 25·30s 30·58e
Newcastle 27·49s 29·55e
New Hanover 29·28s 30·28e
Niekerkshoop 29·19s 22·51e
Nieu Bethesda 31·51s 24·34e
Nieuwoudtville 31·23s 19·07e
Nigel 26·30s 28·28e
Nkwalini 28·45s 31·33e
Nondweni 28·11s 30·49e
Nongoma 27·58s 31·35e
Northam 25·03s 27·11e
Nottingham Road 29·22s 30·00e
Noupoort 31·10s 24·57e
Nuwerus 31·08s 18·24e
Nylstroom 24·42s 28·20e
Odendaalsrus 27·48s 26·45e
Ohrigstad 24·49s 30·33e
Okiep 29·39s 17·53e
Olifantshoek 27·57s 22·42e
Omdraaisvlei 30·08s 23·08e
Oranjefontein 23·25s 27·41e
Oranjerivier 29·40s 24·12e
Orkney 27·00s 26·39e
Ottosdal 26·58s 26·00e
Ottoshoop 25·45s 25·59e
Oudtshoorn 33·35s 22·14e
Paarl 33·45s 18·56e
Pacaltsdorp 34·00s 22·28e
Pampoenpoort 31·03s 22·40e
Park Rynie 30·25s 30·35e
Parys 27·04s 27·16e
Patensie 33·46s 24·49e
Paterson 33·26s 25·58e
Paulpietersburg 27·30s 30·51e
Paul Roux 28·18s 27·59e
Pearston 32·35s 25·08e
Peddie 33·14s 27·07e
Penge 24·22s 30·13e
Perdeberg 28·59s 25·05e
Perdekop 27·13s 29·38e
Petrusburg 29·08s 25·27e
Petrus Steyn 27·38s 28·08e
Petrusville 30·05s 24·41e
Phalaborwa 23·55s 31·13e
Philippolis 30·19s 25·13e
Philipstown 30·26s 24·29e
Pienaarsrivier 25·15s 28·18e
Pietermaritzburg 29·37s 30·16e
Pietersburg 23·54s 29·25e
Piet Retief 27·01s 30·50e
Piketberg 32·54s 18·46e
Pilgrim's Rest 24·55s 30·44e
Pinetown 29·52s 30·46e
Plettenbergbaai 34·04s 23·22e
Pofadder 29·10s 19·22e
Pomfret 25·50s 23·32e
Port Alfred (Kowie) 33·36s 26·55e
Port Edward 31·02s 30·13e
Port Elizabeth 33·58s 25·40e
Porterville 33·00s 19·00e
Port Nolloth 29·17s 16·51e
Port Saint Johns 31·38s 29·33e
Port Shepstone 30·46s 30·22e
Postmasburg 28·18s 23·05e
Potchefstroom 26·46s 27·01e
Potgietersrus 24·15s 28·55e
Pretoria 25·45s 28·10e
Prieska 29·40s 22·42e
Prince Albert 33·13s 22·02e
Prince Alfred Hamlet 33·18s 19·20e
Protem 34·16s 20·05e
Pudimoe 27·26s 24·44e
Punda Milia 22·40s 31·05e
Putsonderwater 29·09s 21·51e
Queenstown 31·52s 26·52e
Qumbu 31·10s 28·48e
Randfontein 26·11s 27·42e
Reddersburg 29·38s 26·07e
Redlands 29·51s 22·58e
Reitz 27·53s 28·31e
Reivilo 27·36s 24·08e
Richmond 29·54s 30·08e
Richmond 31·23s 23·56e
Rietfontein 26·44s 20·01e
Ritchie 29·02s 24·38e
Riversdale 34·07s 21·15e
Riviersonderend 34·09s 19·55e
Robertson 33·46s 19·50e
Rosmead 31·29s 25·08e
Rouxville 30·29s 26·46e
Rustenburg 25·37s 27·08e
Sabie 25·10s 30·48e
Sakrivier 30·54s 20·28e
Saldanha 33·00s 17·56e
Sannieshof 26·30s 25·47e

MAP INDEX (continued)

Saron 33·11s 19·01e
Sasolburg 26·48s 27·45e
Satara 24·29s 31·47e
Schweizer-Reneke 27·11s 25·18e
Scottburgh 30·19s 30·40e
Senekal 28·30s 27·32e
Setlagodi 26·16s 25·06e
Settlers 25·02s 28·30e
Seymour 32·33s 26·46e
Sezela 30·24s 30·42e
Shannon 29·08s 26·18e
Shingwidzi 23·05s 31·25e
Sibasa 22·53s 30·33e
Silver Streams 28·20s 23·33e
Simonstown (Simonstad) 34·14s 18·26e
Sishen 27·55s 22·59e
Skoenmakerskop 34·02s 25·33e
Skukuza 25·01s 31·38e
Slurry 25·49s 25·52e
Smithfield 30·09s 26·30e
Sodium 30·11s 23·09e
Soekmekaar 23·28s 29·58e
Somerset East 32·42s 25·35e
Somerset West 34·08s 18·50e
Sonskyn 30·47s 26·28e
Sonstraal 27·07s 22·28e
Soutpan 28·43s 26·04e
Springbok 29·43s 17·55e
Springfontein 30·19s 25·36e
Springs 26·13s 28·25e
Standerton 26·58s 29·07e
Stanger 29·27s 31·14e
Steinkopf 29·18s 17·43e
Stella 26·38s 24·48e
Stellenbosch 33·58s 18·50e
Sterkaar 31·05s 23·42e
Sterkstroom 31·32s 26·32e
Steynsburg 31·15s 25·49e
Steynsrus 27·58s 27·33e
Steytlerville 33·21s 24·21e
Stilbaai 34·24s 21·26e
Stilfontein 26·50s 26·50e
Stoffberg 25·29s 29·49e
Stormsrivier 33·59s 23·52e
Strand 34·06s 18·50e
Strydenburg 29·58s 23·40e
Stutterheim 32·33s 27·28e
Sutherland 32·24s 20·40e
Suurbraak 34·00s 20·39e
Swartruggens 25·40s 26·42e
Swellendam 34·02s 20·26e
Talana 28·10s 30·15e
Tarkastad 32·00s 26·16e
Taung 27·33s 24·47e
Thaba Nchu 29·17s 26·52e
Thabazimbi 24·41s 27·21e
Theunissen 28·30s 26·41e
Three Sisters 31·54s 23·06e
Tom Burke 23·05s 28·00e
Tongaat 29·37s 31·03e
Tosca 25·53s 23·58e
Touwsrivier 33·20s 20·00e
Trichardt 26·28s 29·13e
Trompsburg 30·01s 25·46e
Trooilapspan 28·40s 21·25e
Tshidilamolomo 25·50s 24·41e
Tsineng 27·06s 23·04e
Tsolo 31·18s 28·37e
Tulbagh 33·17s 19·09e
Tweeling 27·38s 28·31e
Tweespruit 29·11s 27·01e
Tzaneen 23·50s 30·09e
Ubombo 27·33s 32·00e
Ugie 31·10s 28·13e
Uitenhage 33·40s 25·28e
Ulco 28·21s 24·15e
Umhlanga Rocks 29·43s 31·06e
Umkomaas 30·15s 30·42e
Umtata 31·35s 28·47e
Umtentweni 30·42s 30·28e
Umzimkulu 30·16s 29·56e
Umzinto 30·22s 30·33e
Underberg 29·50s 29·22e
Uniondale 33·40s 23·08e
Upington 28·25s 21·15e
Utrecht 27·38s 30·20e
Uvongo Beach 30·51s 30·23e
Vaalwater 24·20s 28·03e
Vanderbijlpark 26·42s 27·54e
Van Reenen 28·22s 29·24e
Vanrhynsdorp 31·36s 18·44e
Vanstadensrus 29·59s 27·02e
Vanwyksdorp 33·46s 21·28e
Vanwyksvlei 30·18s 21·49e
Vanzylsrus 26·52s 22·04e
Velddrif 32·47s 18·11e
Ventersburg 28·09s 27·08e
Ventersdorp 26·17s 26·48e
Venterstad 30·47s 25·48e
Vereeniging 26·38s 27·57e
Verulam 29·45s 31·02e
Victoria West 31·25s 23·04e
Vierfontein 27·03s 26·46e
Viljoenskroon 27·12s 27·00e
Villa Nora 23·34s 28·01e
Villiers 27·03s 28·35e
Virginia 28·12s 26·49e
Vishoek 34·08s 18·25e
Volksrust 27·24s 29·53e
Vosburg 30·33s 22·52e
Vrede 27·30s 29·06e
Vredefort 27·05s 27·16e
Vredenburg 32·54s 17·59e
Vredendal 31·41s 18·35e
Vryburg 26·55s 24·45e
Wakkerstroom 27·24s 30·10e
Walvisbaai (Walvis Bay) 22·59s 14·31e
Warden 27·56s 29·00e
Warmbad 24·55s 28·15e
Warrenton 28·09s 24·47e
Wasbank 28·24s 30·05e
Waterval-Boven 25·40s 30·20e
Weenen 28·57s 30·03e
Welkom 27·59s 26·45e
Wellington 33·38s 18·57e
Wepener 29·46s 27·00e
Wesselsbron 27·50s 25·23e
Westonaria 26·19s 27·41e
Whittlesea 32·10s 26·50e
Williston 31·20s 20·53e
Willowmore 33·17s 23·29e
Willowvale 32·16s 28·30e
Winburg 28·37s 27·00e
Windsorton 28·16s 24·44e
Witbank 25·56s 29·07e
Witdraai 26·58s 20·45e
Witrivier 24·40s 31·00e
Witsand 34·24s 20·50e
Wolmaransstad 27·12s 26·13e
Wolseley 33·26s 19·12e
Worcester 33·39s 19·27e
Wuppertal 32·15s 19·15e
Wynberg 34·02s 18·28e
Zastron 30·18s 27·07e
Zebediela 24·19s 29·21e
Zeerust 25·33s 26·06e
Zoar 33·30s 21·28e

Physical features and points of interest

Addo Elephant National Park 33·29s 25·46e
Agulhas, Cape 34·52s 20·00e
Algoa Bay 33·50s 25·65e
Asbesberge, *mountains* 28·55s 23·15e
Atlantic Ocean 32·00s 15·00e
Augrabies Falls National Park 28·35s 20·19e
Augrabiesvalle, *waterfalls* 28·35s 20·19e
Auob, *river* 26·36s 20·36e
Bamboesberg, *mountains* 31·30s 26·10e
Bashee, *river* 32·15s 28·53e
Baviaanskloof-berge, *mountains* 33·35s 24·10e
Bechuanaland, *physical region* 27·10s 22·10e
Blouberg, *mountain* 23·01s 28·59e
Boegoebergdam, *dam* 29·02s 22·12e
Boesmans, *river* 33·42s 26·39e
Brak, *river* 29·35s 22·55e
Breë, *river* 34·24s 20·50e
Buffalo, *river* 28·43s 30·37e
Buffels, *river* 33·45s 21·11e
Bushman Land, *historic region* 29·13s 19·35e
Caledon, *river* 30·31s 26·05e
Champagne Castle, *mountain* 29·06s 29·20e
Columbine, Cape 32·47s 18·52e
Danger Point 34·40s 19·17e
Dasseneiland, *island* 33·26s 18·04e
Die Berg, *mountain* 25·12s 30·09e
Doring, *river* 31·54s 18·39e
Drakensberg, *mountains* 27·00s 30·00e
Elands, *river* 25·10s 29·10e
False Bay (Valsbaai) 34·12s 18·40e
Gamka, *river* 33·18s 21·39e
Gakarosa, *mountain* 27·50s 23·38e
Gamtoos, *river* 33·58s 25·01e
Giants Castle, *mountain* 29·21s 29·27e
Giants Castle Game Reserve, *wildlife refuge* 29·16s 29·30e
Golden Gate Highlands National Park 28·30s 28·40e
Good Hope, Cape of 34·24s 18·30e
Gourits, *river* 34·21s 21·52e
Great Karroo, *plateau* 32·25s 22·40e
Great Swartburg, see Groot-Swartberg
Griqualand East, *physical region* 30·30s 29·00e
Griqualand West, *physical region* 28·20s 23·30e
Groen, *river* 30·40s 23·17e
Groot, *river* 33·54s 21·39e
Groot-Berg, *river* 32·47s 18·08e
Groot-Kei, *river* 32·41s 28·22e
Groot Letaba, *river* 23·58s 31·50e
Groot Shingwidzi, *river* 23·15s 31·35e
Groot-Swartberge (Great Swartberg) *mountains* 33·22s 22·20e
Groot-Vis, *river* 33·30s 27·08e
Grootvloer, *salt flat* 30·00s 20·40e
Hakskeenpan, *salt lake* 26·48s 20·12e
Hangklip, Kaap, *cape* 34·26s 18·48e
Hantamsberg, *mountains* 31·22s 19·45e
Hartbees, *river* 28·45s 20·32e
Harts, *river* 28·24s 24·17e
Hluhluwe Game Reserve, *wildlife refuge* 28·05s 32·04e
Hout, *river* 23·04s 29·36e
Indian Ocean 32·00s 32·00e
Infanta, Cape 34·29s 20·51e
Injaswti, *mountain* 29·09s 29·23e
Kaap Plato, *plateau* 28·20s 23·57e
Kaffraria, *historic region* 31·30s 28·30e
Kalahari Gemsbok National Park 26·00s 20·30e
Kareeberge, *mountains* 30·53s 21·57e
Kariega, *river* 33·03s 23·28e
Klip, *river* 26·16s 27·49e
Komati, *river* 25·26s 32·00e
Kompasberg, *mountain* 31·45s 24·32e
Komsberg, *mountains* 32·40s 20·50e
Kosi, *lake* 26·55s 32·52e
Kougaberge, *mountains* 33·40s 23·50e
Kraai, *river* 30·40s 26·45e
Krokodil, *river* 25·26s 32·00e
Kruger National Park 24·00s 31·40e
Kuruman, *river* 26·56s 20·39e
Kuruman Heuwels, *hills* 27·40s 23·25e
Langeberg, *mountains* 34·00s 20·40e
Lebombo Mountains 25·30s 32·00e
Limpopo, *river* 22·25s 31·20e
Little Karroo, *plateau* 33·45s 21·30e
Little Namaland, *historic region* 29·00s 17·15e
Loskopdam *dam* 25·23s 29·20e
Loskop Game Reserve, *wildlife refuge* 25·23s 29·20e
Luvuvhu, *river* 22·40s 30·55e
Macdhui, Ben, *mountain* 30·39s 27·58e
Marico, *river* 23·35s 26·57e
Mentzdam, *dam* 33·10s 25·09e
Mkomazi, *river* 30·12s 30·50e
Mkuze, *river* 27·53s 32·29e
Mkuze Game Reserve, *wildlife refuge* 27·40s 32·15e
Modder, *river* 29·02s 24·37e
Mogalakwena, *river* 23·00s 28·40e
Molopo, *river* 28·30s 20·13e
Mooi, *river* 28·45s 30·34e
Mountain Zebra National Park 32·16s 25·29e
Mount aux Sources, *mountain* 28·46s 28·52e
Mtamvuna, *river* 31·06s 30·12e
Mzimvubu, *river* 31·38s 29·32e
Ndumo Game Reserve, *wildlife refuge* 26·53s 32·15e
Nossob, *river* 26·55s 20·37e
Nuweveldberge, *mountains* 32·13s 21·40e
Nzheleledam, *dam* 22·44s 30·06e
Olifants, *river* 23·58s 31·50e
Orange, *river* 28·41s 16·28e
Outeniqua Mountains 33·53s 22·35e
Pilanesberg, *mountain* 25·14s 27·04e
Pondoland *physical region* 31·10s 29·30e
Quoin Point 34·46s 19·37e
Richard's Bay 28·50s 32·02e
Riet, *river* 29·00s 23·54e
Roggeveldberge, *mountains* 32·17s 20·08e
Royal Natal National Park 28·45s 28·57e
Sabie, *river* 25·10s 32·02e
Saint Francis, Cape 34·14s 24·49e
Saint Lucia, Cape 28·25s 32·25e
Saint Lucia, Lake 28·05s 32·26e
Saint Lucia Game Reserve, *wildlife refuge* 28·10s 32·28e
Sak, *river* 30·52s 20·25e
Saldanha Bay 33·04s 18·00e
Sederberge, *mountains* 32·23s 19·20e
Sand, *river* 22·25s 30·05e
Sibaya, Lake 27·20s 32·40e
Sneeuberg, *mountains* 31·46s 24·20e
Sondags, *river* 33·44s 25·51e
Sout, *river* 31·35s 18·24e
Soutpansberg, *mountains* 22·55s 29·30e
Stormberg, *mountains* 31·27s 26·55e
Suurberge, *mountains* 33·18s 25·32e
Table Bay 33·53s 18·27e
Tembuland, *physical region* 31·30s 28·20e
Tina, *river* 31·18s 29·14e
Touws, *river* 33·45s 21·11e
Transkei, *historic region* 31·20s 29·00e
Tsomo, *river* 32·25s 27·50e
Tugela, *river* 29·14s 31·30e
Umfolozi, *river* 28·25s 32·26e
Umfolozi Game Reserve, *wildlife refuge* 28·19s 31·50e
Vaal, *river* 27·40s 26·09e
Valley of Desolation National Monument 32·17s 24·30e
Vals, *river* 27·23s 26·30e
Valsbaai, see False Bay
Verneukpan, *salt lake* 29·58s 21·10e
Verwoerd Reservoir 30·40s 25·40e
Vis, *river* 33·30s 27·08e
Walvis Bay 22·59s 14·31e
Waterberge, *mountains* 20·30s 17·18e
Wild Coast 32·30s 29·00e
Wilge, *river* 27·03s 28·20e
Willem Pretorius Game Reserve, *wildlife refuge* 28·16s 27·13e
Winterberge, *mountains* 32·28s 26·15e
Witwatersrand, *physical region* 26·00s 27·00e
Zululand, *historic region* 28·10s 32·00e

preserve the larger animal species from extermination, a number of wildlife reserves came into being, of which the Kruger National Park in the northeast is the best known. There is a rich bird life. Of the more than 100 species of snakes, about a quarter are poisonous.

Traditional regions. While South Africa could be divided into numerous microregions, known by a variety of local names, these may be grouped more conveniently into 11 large traditional regions known locally and internationally by the same names.

The veld regions

Three veld regions may be discerned in the north. These are the Bushveld in the far northern Transvaal west of the Transvaal Drakensberg mountains; the Lowveld east of the mountains; and the Highveld to the south, including the southern Transvaal, eastern Orange Free State, Lesotho, and an adjoining portion of the Cape Province. The Bushveld and the Lowveld are still predominantly rural, but the prairie-like landscape of the Highveld on the eastern interior plateau contains the mining, industrial, and commercial heart of South Africa and the largest urban centres.

To the west and southwest of these veld regions the arid and semiarid landscape consists of three traditional regions. These are the Kalahari (*q.v.*) in the north, which includes most of Cape Province north of the Orange River but also extends into Botswana and eastern South West Africa and which is still very much a marginal and pioneer country of rural people; the Karroo, which is south of the Kalahari and includes most of the interior of Cape Province and which is inhabited by rural Afrikaans-speaking people and a very few Bantu; and Namaqualand, a rugged rural area in the extreme west, immediately south of the Orange River.

A final category includes the more humid regions bordering the South Atlantic and Indian oceans. These are the Western Province in the southwestern corner of the country, the traditional area of the Cape Coloureds and the oldest white settlement in South Africa; the Eastern Province, centring on the port towns of Port Elizabeth and East London; the Southern Cape Districts (in South Africa itself perhaps better known as the Southwestern Districts) located between the Western and Eastern provinces; the Transkei, South Africa's largest Bantu homeland inhabited by the Xosa-speaking people, situated to the northeast of the Eastern province; and Natal (*q.v.*), a garden province, English in character and tradition and harbouring the bulk of South Africa's Indian community.

The landscape under human settlement. Approximately one-half of the total population still lives in rural areas. There has been a marked decline, however, in the white rural population in recent years. At the beginning of the 20th century, half of the white population was still rural, whereas today the proportion has dropped to one-eighth. The nonwhite rural population has declined from 85 percent to 60 percent in the same period.

In the traditional Bantu areas, settlement takes the form of farming villages; the land belongs to the whole kraal (village) community, each family receiving from the chieftain the right to build his home. Pastoral land surrounding the kraal is used communally. Among the whites of South Africa, settlement takes the form of dispersed farms. The farms are often miles apart in the arid and semiarid regions but are closer together in the better watered parts.

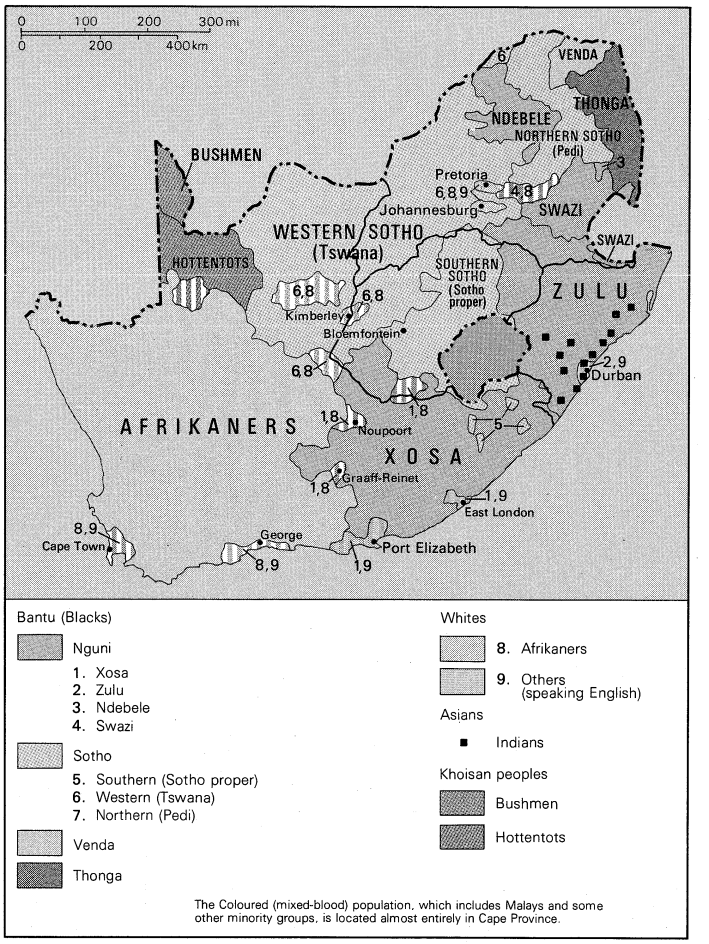
Towns and cities

Urban settlement is much denser in the east and south. All 17 towns with populations of more than 100,000 are situated in this more humid half of the country. Also located here are the four main urban concentrations—the Pretoria–Witwatersrand–Vereeniging complex in the southern Transvaal; the Cape Peninsula in southwestern Cape Province; the Durban–Pietermaritzburg area in Natal; and the Port Elizabeth–Uitenhage region in the eastern Cape Province.

Of all these, the Witwatersrand conurbation is by far the largest and is the most important for mining, industry, and commerce. Most South African towns originated as centres that supplied services to the surrounding rural areas. Others, such as Grahamstown and Simonstown, originally had a military function, while still others, such as Welkom, Virginia, and Kimberley, began as mining towns, sometimes—as in the case of Kimberley—shedding the mining function at a later stage. Some towns, like Sasolburg, where gasoline is manufactured from coal, developed around a specific industry. Railroad centres, such as Germiston and De Aar, and harbour towns, such as Durban, Cape Town, Port Elizabeth, and East London, represent another urban category; Pretoria is the main administrative centre; and Stellenbosch and Grahamstown represent towns where the educational function is outstanding. Johannesburg (*q.v.*), the economic hub of the Witwatersrand area, is the major city.

PEOPLE AND POPULATION

Groups historically associated with South Africa. *Ethnolinguistic groups.* Four racial groups are represented in South Africa, and each of these may be further subdivided into different linguistic and ethnic groups. In a broad sense, one may differentiate between whites (or Europeans) and nonwhites. The whites consist of two main groups—the Afrikaners, who are the descendants of Dutch, French, and, to a lesser degree, German forbears, and the English-speaking group. The Afrikaners comprise about 60 percent of the white population, and the English-speaking section 40 percent. Afrikaans and English are the two official languages of the country.



General ethnic composition of South Africa.

The nonwhites outnumber the whites by approximately five to one; they comprise a variety of Bantu, Coloured, and Asian peoples.

The Bantu majority

The Bantu alone form over 70 percent of the total population and may be broadly divided into four main ethnic groups. These are the Nguni, including the Xosa, Zulu, Swazi, and Ndebele peoples, who comprise some 60 percent of the Bantu of South Africa and live mainly in the eastern coastal regions; the Sotho, found mostly in the central and western Bantu areas and constituting approxi-

mately 32 percent of the Bantu; and the Venda and Thonga groups, who form 2 percent and 5 percent, respectively, of the Bantu and who are found only in the northern Transvaal. Other Bantus form 1 percent of the total. Communication with members of non-Bantu groups is often in English or Afrikaans. To ease communication among Bantu language groups working in the Witwatersrand mines, a new language, Fanagalo, has evolved.

The smaller nonwhite groups are the Coloureds (also known as the Cape Coloureds because the bulk of them are concentrated in the southwestern Cape Province), and the Asians—predominantly Indian—of whom more than 80 percent live in Natal Province.

During the early period (the 17th and 18th centuries) of the Dutch East India Company's regime at the Cape of Good Hope, miscegenation took place between the white colonists and slave women (mostly Malays from Dutch East India) as well as Hottentot women, but mostly between Hottentot women and male slaves, especially after the latter were freed in the earlier part of the 19th century; miscegenation also occurred between sailors from ships calling at the Cape and Malays. Thus came into being the Coloured community of South Africa. Varying in skin colour from almost black to almost white, they are a westernized people having higher living standards than the Bantu and are predominantly Afrikaans-speaking.

The founding of an Asian community began in 1860, when a number of contract Indians were imported to work on the Natal sugar plantations; the community grew and in the early 1970s represented about 3 percent of the population. The original Hottentots and Bushmen are almost nonexistent; a few Bushmen are still to be found in parts of the Kalahari.

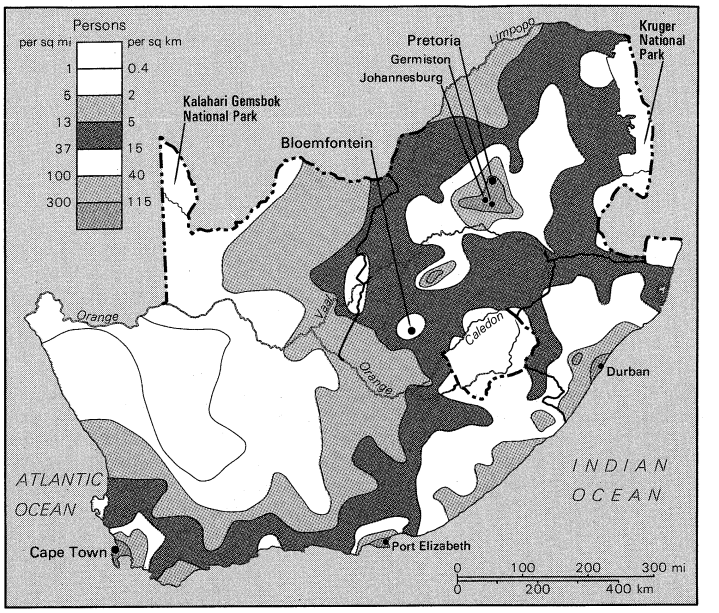
Religious groups. Among the whites and Coloureds, the Dutch Reformed Church has by far the most adherents, amounting to approximately 2,000,000 people; there are also more than 500,000 Bantu members. Among the Bantu, the Bantu Christian churches have the largest membership, with more than 2,000,000 adherents, followed by the Methodist, Roman Catholic, and Anglican churches. Most of the Asians are Hindu, but many Muslim Malays live in and around Cape Town. More than 100,000 Jews live in South Africa.

The demographic pattern

Demography. At the 1970 census, the total South African population was 21,448,169. Of these, whites numbered 3,751,328 (over 17 percent of the total); Bantu 15,057,952 (a little over 70 percent); Coloureds 2,018,453 (9 percent); and Asians 620,436 (3 percent).

The average annual increase of the total population is almost 2.4 percent, being highest for the Coloureds (more than 3 percent) and lowest for the whites (less than 2.3 percent), with Asians (2.4 percent) and Bantu (2.3 percent) in the middle category. Even when white immigration is taken into account, the rate of annual natural increase is almost twice as high for Coloureds as it is for whites—26 per thousand as against 15 per thousand. Should the growth rate of the decade of the 1960s be maintained, South Africa will have a total population of some 42,000,000 by the year 2000, of which 7,000,000 (or about 17 percent of the total) would be white, 28,000,000 Bantu (66 percent), 5,800,000 Coloured (14 percent), and 1,200,000 Asians (3 percent). The outstanding demographic feature of this projection is the relative increase of the Coloured population.

There is a definite decrease in the percentage of whites in the total population, and this relative decrease would have been even more pronounced had it not been for white immigration. Without taking into account the considerable black migrant-labour force, which comes from a variety of countries in southern and Central Africa on a temporary basis, it may be said that before 1961 immigration played a minor role. In 1961 the government established a Department of Immigration to stimulate the immigration of whites into the country. From 1961 to 1967, South Africa had an average annual immigration of more than 33,000 whites, though more than 10,000 left, resulting in an average net gain of over 23,000 whites annually. The government is attempting to attract more white immigrants, especially those with high skills.



Population density of South Africa.

The pattern of population distribution and density closely follows that of the rainfall pattern. The western arid and semiarid half of the country, with the exception of the southwestern Cape Province, has a low population density, increasing toward the south, the north, and especially toward the east. For the country as a whole, the population density is 45 persons per square mile, being lowest in the Cape Province, with roughly 15 per square mile, and highest in Natal, with 64; Transvaal (58) and the Orange Free State (33) come in between. The highest density is reached in the Bantu areas along the east coast of Cape Province, and in Natal, where it reaches 100 and more, with some Bantu districts even supporting more than 200 persons per square mile. In the greater part of the Karoo, Namaqualand, and the Kalahari, all situated in the western half of the country, the densities drop to less than five per square mile. (A.Ne.)

South Africa, Area and Population

	area*		population	
	sq mi	sq km	1960 census	1970 census
Provinces				
Cape of Good Hope†	278,380	721,001	5,363,000	4,236,000
Natal	33,578	86,967	2,980,000	2,140,000
Orange Free State	49,866	129,152	1,387,000	1,649,000
Transvaal	109,621	283,917	6,273,000	6,389,000
Bantu National Homelands				
Basotho Qwaqwa (Basotho ba Borwa)	...	...	...	25,000
Bophuthatswana (Tswana)‡	...	...	...	887,000
Ciskei‡	...	...	...	527,000
Gazankulu (Machangana)	...	...	...	268,000
Kwazulu (Zulu)	...	...	...	2,106,000
Lobowa‡	...	...	...	1,087,000
Swazi	...	...	...	118,000
Transkei‡	...	...	...	1,751,000
Vendal	...	...	...	265,000
Total South Africa†	471,445	1,221,037	16,003,000	21,448,000

*Territory of the Bantu homelands is dispersed, and often discontinuous, throughout the area of the provinces. Authoritative figures are not available.
†Excludes Walvis Bay, which is part of Cape province, but is administered by South West Africa. ‡Autonomous as at October, 1972.
Source: Official government figures.

THE NATIONAL ECONOMY

A comparative view. South Africa has enjoyed a high rate of economic growth for a long time. Since World War II the annual growth of real gross domestic product has averaged 5 percent or more. During the 1960s, South Africa experienced one of the most consistently high growth rates in the world. Between 1960 and 1969, the economy grew by 6 percent per annum in real terms.

The bases of South Africa's economy consist of agriculture, mining, and manufacturing. Mineral resources are extensive. Apart from gold and diamonds, there are nearly 70 other types of exploitable minerals. Agriculture is well developed and embraces a wide range of products. There is a large supply of cheap power; in addition there are substantial reserves of cheap labour, there is no shortage of entrepreneurial talent, and there is ready access to supplies of domestic and foreign capital.

In relative size, South Africa's national income is equivalent to about one-seventh that of the United Kingdom, but it is far larger than that of any other African country. In 1969, per capita gross domestic product was R (rand) 593. (R 0.71 equals $1 U.S.; R 1.71 equals £1 sterling as of December 1, 1970.)

The extent and distribution of resources. *Mineral resources.* Diamonds and gold are the best known and historically the most important minerals, but there are many others, such as copper, iron ore, manganese, asbestos, chrome, silver, beryllium, antimony, tin, and platinum that are of growing importance, and some resources have yet to be exploited. The only important exclusion from South Africa's list of mineral resources is oil, for which an energetic search was being made in and around South Africa in the 1970s. Meanwhile, oil is being made from coal, of which there are substantial deposits in Sasolburg. A natural gas discovery was announced early in 1969 at Plattenberg Bay on the south coast.

Diamonds and gold

Gold remains far and away the most important mineral mined, but production has probably reached its peak. A gradual decline in output is forecast from the early 1970s, the result of rising costs and declining reserves. In 1970 gold production contributed R 805,000,000 out of a total mining production worth R 1,203,000,000. Gold is produced in three main areas—the Witwatersrand, which includes Johannesburg; the Far West Rand and Klerksdorp fields; and in the Orange Free State around Odendaalsrus. Production in 1970 was 32,164,000 ounces; most of this was mined, but about one-eighth was alluvial. The centre of the diamond industry is at Kimberley. There is also diamond production in the Orange Free State and the Transvaal. Production in 1970 was 8,112,000 carats—over 70 percent more than in 1964. Known coal deposits are put at 75,000,000,000 tons, almost all of which are located in the Transvaal. The biggest use of coal is for the generation of electricity.

More than 20 companies are now producing uranium. The mining industry has established an independent company, Nuclear Fuels Corporation of South Africa (Pty.), which is responsible for processing and marketing all uranium produced. In 1970 uranium oxide production amounted to more than 3,700 tons.

Biological resources. Agriculture is of major importance to South Africa, but the supply of good agricultural land is effectively curtailed by the climate. The average rainfall for the whole country is only 18 inches, which is less than the world's average level and is very little for a subtropical country with a high evaporation rate. Runoff of water from rainfall is only 8 percent. Much of the country receives an even lower rainfall and is thus only marginally useful as agricultural land. Soil erosion is a problem in some areas, and the country is particularly subject to droughts.

Resources of timber are minimal; less than 1 percent (about 2,500,000 acres) of the land area is forested. Conifers occupy a little more than half of the forested area; eucalyptus (used mainly for mining timber and pulpwood) covers almost a quarter; and wattle (acacia), which is used for producing tanin bark, as well as in the mining and pulping industries, accounts for almost 20 percent. The forested areas are only found in the wetter regions of the south and east.

Although not as rich in animal life as East Africa, South Africa still has a large range of wild animals and birds. These play an important part as a tourist attraction and are preserved in a number of national parks.

Hydroelectric resources. As a consequence of the low annual rainfall, natural hydroelectric potential is minimal. Plans to develop hydroelectricity, however, form part of such schemes as the Orange River Project (see ORANGE RIVER). Divided into six stages of construction and to be completed over 30 years, this project involves the irrigation of large areas of some of the most arid parts of the country. It is, however, intended to import approximately 600 megawatts of electricity each year from the Cabora Bassa scheme in Mozambique when this is completed, and the Electricity Supply Commission (Escom) will also obtain power from the Oxbow scheme in Lesotho.

The Orange River Project

Sources of national income. *Agriculture, forestry, and fishing.* Agricultural production, including forestry and fishing, accounted for just under 10 percent of gross domestic product in 1970. The importance of this industry in the economy declined in the 1960s as it has grown rather more slowly than other sectors. It remains important, however, since variations in its output can seriously affect overall growth in the economy and since this sector contributes a large proportion of exports—46 percent in 1969. The agricultural sector is also of major importance in employment, giving work to approximately 25 percent of the working population. In mid-1962 there were 106,011 European-, Coloured-, and Indian-owned farms, with a total area of 66,643,100 acres. The aggregate value of agricultural production in 1969 was R 1,005,000,000, compared with R 1,231,000,000 for mining, and R 2,390,000,000 for manufacturing. Among the major products are maize, wheat, cane sugar, groundnuts (peanuts), millet and sorghum, citrus fruits, grapes, tobacco, wool, and meat. More than 1,533,000 cattle were slaughtered in 1967/68, more than 43,712,000 sheep and goats, and almost 1,700,000 pigs. Dairy production includes some 680,000,000 gallons of milk a year, 150,000,000 gallons consumed directly and 250,000,000 gallons converted into butter, cheese, and condensed milk.

Fishing areas are chiefly off the south and west coasts. Both shoal fishing (for pilchard and maasbanker) and offshore trawling for a variety of fish, including kingklip, agulha sole, stockfish (Cape hake), and kabeljou, are employed.

Mining and quarrying. The mining and quarrying sector is growing rather more slowly than is the economy in general; it is, however, expected to be of major importance in the future and figures prominently in government economic policy. Gold is the main product, after which copper, which was worth R 138,000,000 in 1970, is the next most important. Diamond production was worth R 97,000,000 in 1969, followed by asbestos (R 34,000,000), iron ore (R 29,000,000), and manganese (R 26,000,000). Coal sales reached R 116,000,000. Sales of gold are expected to decline since the Witwatersrand fields are rapidly being exhausted. By 1985 all but eight of the 35 operating mines are expected to be closed. The government has offered to subsidize the ailing mines for a period of years in order to increase production, but this postpones rather than solves the major problem of finding a replacement for gold.

Manufacturing. Between 1957 and 1970 the value of manufacturing output rose by more than 8 percent a year to a total of R 2,657,000,000. In the 1960s it contributed 21 percent to the gross domestic product and became the prime force in South Africa's rapid and sustained growth. The sector is also prominent in exports, contributing 33 percent of the total in 1970. As South Africa's main imports are manufactures, it has also helped the economy by reducing the necessity for imports. The fastest growing industries during the 1960s were transport equipment (which increased 234 percent in production between 1958 and 1969), textiles (which increased 140 percent in the same period), chemicals (136 percent), and basic metals (128 percent). The food, timber, furniture, leather, and rubber products industries have grown at rather slower rates than the average.

One of the chief reasons for the rapid growth of this sector has been the high proportion of the gross domestic product devoted to investment in South Africa. Throughout the 1960s this ran at well above 20 percent.

Energy. Consumption of oil in the early 1970s ran at 7,400,000 tons a year. Some 12 percent of this was pro-

Oil consumption

duced from coal by the state controlled coal, oil, and gas corporation. The rest was imported as crude oil and refined locally. A full year's supply of oil was stockpiled in case of an oil embargo.

Coal output, which is below colliery capacity, was almost 62,000,000 short tons in 1970. Five years before, it had been only 49,500,000 tons. Escom, the state electricity supply corporation, takes more than 50 percent of this output.

South Africa's electric power generation ranks 15th in the world in total output of electricity; South Africa produces 57 percent of all of Africa's electric power, although it has less than 7 percent of its population. In 1970 over 48,000,000,000 kilowatt-hours were generated—an increase of over 39 percent since 1965.

Financial services. South Africa has a well-developed financial system. The central bank—the South African Reserve Bank—is the sole note-issuing authority. It also exercises all the functions necessary for the implementation of monetary policy. There are nine commercial banks whose total assets and liabilities in 1970 averaged over R 3,300,000,000. Liquidity ratios (*i.e.*, the proportion of liquid capital available) tend to be rather high. In 1969 the average figure was 36 percent, against a required rate of less than 30 percent. The banking system also includes a number of merchant banks and discount houses. The discount houses deal primarily with trade bills and promissory notes rather than with treasury bills. There are also two specialist state banks—the National Finance Corporation, which was started in 1949 to develop a short-term money market, and the Land Bank, which, as its name implies, lends chiefly to agriculture.

Second in financial strength only to the commercial banks are the building societies. In 1969 their total liabilities averaged R 2,596,000,000. Installment buying, savings, and general banking are also important, as also are private pension and provident funds and insurance companies.

There is an active capital market organized around the Johannesburg Stock Exchange. Turnover increased rapidly throughout the 1960s. In 1969, however, the market suffered a severe decline, and during the early 1970s it still remained becalmed and showed no strong upward trend.

Foreign trade. Dependence upon foreign trade is relatively high. In 1969 both exports and imports were equivalent in value to 28 percent of the gross domestic product. In 1970 food, drink, and tobacco exports made up 30 percent of total exports, raw materials 24 percent, and manufactures 33 percent. Machinery and transport equipment accounted for 7 percent. South Africa's main markets are in western Europe, with the United Kingdom taking almost one-third of the total (excluding gold), and the European Free Trade Association (EFTA) and the European Economic Community (EEC) countries combined taking some 50 percent. The rest of Africa took over 17 percent in 1970, and Japan also is growing as a market. Imports of food and raw materials are insignificant compared with many industrial countries. Mineral oil imports account for 5 percent of the total, but the major items are manufactures (over 18 percent) and machinery and transport equipment (46 percent). South Africa's main suppliers are western European countries, the United States, and Japan.

Trading partners

Management of the economy. *Private enterprise and the role of government.* The economy is essentially based on private enterprise, but the state participates in a number of ways. Through the Industrial Development Corporation of South Africa, it controls several public corporations—including the South African Coal, Oil and Gas Corporation (SASOL), which produces oil from coal; the Phosphate Development Corporation (FOSKOR), which is developing phosphates; the African Metals Corporation Ltd. (Amcor), which makes metals; the Iron and Steel Industrial Corporation (Iscor), the major steel producers; SAICCOR (South African Industrial Cellulose Corporation Ltd.), which produces pulp for rayon; and Escom, the main electricity-generating authority. It also runs the railways, the national airline, harbours, telecommunications, and much of the road transport system. The government sets out targets for development in each industry in an Economic Development Plan, and, by means of a range of official bodies, encourages development of industry through financing, consultations on tariff protection, export promotion, export credit insurance, research, and a bureau of standards. The government also practices a decentralization policy, under which numerous tax and other financial incentives are awarded to concerns moving into certain designated areas of the country.

Taxation. In 1970 general government tax revenue was equivalent to 22 percent of gross domestic product —a lower proportion than in many industrialized countries. Throughout the 1960s there was a substantial surplus each year of current revenue over current expenditure. Direct taxes are more important than indirect—they contributed 54 percent of total current revenue compared with 38 percent from indirect taxes. This position has changed since the early 1960s, when the proportions were roughly equal. Direct taxes on individuals consist chiefly of state and provincial income taxes deducted at the source and taxes on dividends and interest. There is no capital gains tax. The main taxes on businesses are income tax and, in some cases, a tax on undistributed profits. Generally speaking, South Africa's tax burden is a light one.

Trade unions. There are effectively two coordinating trade union bodies in South Africa—the right-wing South African Confederation of Labour and the middle-of-the-road Trade Union Council of South Africa (TUCSA). There is also the left-wing South African Congress of Trade Unions (SACTU), which is an almost entirely African organization. Most of its leaders have left the country or are under banning orders (which prevent them from making political statements or attending or organizing gatherings, and also make it illegal to quote them in the press).

The Confederation is almost entirely composed of all-white trade unions. TUCSA membership is multiracial; the organization possesses an economic research bureau, a training and education department, a committee for unorganized workers, and an African affairs section.

The chief bone of contention between the white unions and the nonwhite unions is the question of job reservations. TUCSA has 67 unions affiliated to it, and the Confederation has 27.

Contemporary economic policies. The main aim of economic policy has been to maintain the high rate of economic growth. This growth is dependent largely on the maintenance of a high level of capital investment, a healthy balance of payments, an adequate supply of skilled labour, and control of inflation. In three of these fields considerable success has been achieved, but the labour situation has become serious. The problem is aggravated by the government's policy of trying to reverse the flow of Africans to the big cities and by the white trade unions' unwillingness to admit nonwhites to certain job categories. A range of skilled and semiskilled jobs are reserved for whites, but there is an inadequate supply of skilled white labour to fill vacancies in these job areas, and immigration from Europe has proved insufficient to fill the gap.

The labour shortage

Problems and prospects. The main problems are the self-imposed labour shortage, which has become critical, a too great reliance upon external trade, the potential decline of the gold-mining industry, and a tendency toward inflation, which, in view of the country's reliance on gold exports—for which the price is fixed—is a more serious problem than elsewhere. Future prospects depend mainly upon the government's efforts to ease the labour shortage. If this is not achieved, a serious slackening of the growth rate can be expected, as can greater inflation and a correspondingly worse balance-of-payments situation. This in turn would affect the level of business investment.

Faced with this prospect, it is likely that the government will adopt a less rigid attitude. In the longer term, it will still be necessary to find exports to replace

gold, but, within the context of South Africa's enormous resources, this is quite feasible. (E.I.U.)

Transportation. As South African rivers are unsuitable for navigation, water transport within the country is nonexistent; railroads and roads had to be built from the harbours to points of economic and population concentration in the interior.

The dense railroad network in the eastern half of the country is the result of the fact that mining, industry, and trade are overwhelmingly concentrated there, while agriculture is also better developed in the region because of higher rainfall. The South African railway system, which has a total track length of about 19,600 miles, is state property. The standard gauge, with a few minor exceptions in the Cape and Natal coastal areas, is three feet six inches, and approximately one-quarter of the republic's railroads are electrified. Steam-driven locomotives are still by far the most important.

South African Railways is the greatest single employer in the country; in 1968 to 1969, for example, it employed about 225,000 people. Approximately 500,000,000 passengers are conveyed annually by the railways, together with roughly 120,000,000 tons of goods. Of the total quantity of goods transported, coal alone makes up more than one-quarter, other minerals combined accounting for one-fifth, followed by agricultural products accounting for one-eighth.

The railroad system is augmented by an adequate network of more than 200,000 miles of road, of which about 31,000 miles are paved roads, including more than 7,500 miles of national roads connecting the country's main urban centres.

All South African harbours are state owned and, together with railways and airways, are administered by the Ministry of Transport. The shipping of the country is served mainly by South Africa's four major ports—Durban, Cape Town, Port Elizabeth, and East London—as well as by the port of Lourenço Marques in neighbouring Mozambique. Minor ports are Port Nolloth and Mossel Bay. More than 20,000 ships called at South African ports in 1969 to 1970, while 42,000,000 tons of cargo were handled during the same period. Durban, which serves the southern Transvaal hinterland, which is the economic heartland of South Africa, is the main cargo port. Cape Town, the large port nearest to Europe, is the main passenger port; while it handles only one-third the amount of cargo that Durban does, it is visited by more vessels. A large new harbour is being developed at Richard's Bay, on the Indian Ocean in Natal, to accommodate giant tankers.

Air services

Inland air transport services, both passenger and freight, are operated by state-owned South African Airways and by smaller private local companies. South African Airways also maintains regional services to such neighbouring countries as Lesotho, Botswana, Rhodesia, Madagascar, South West Africa, and Mozambique. South African Airways fly international routes to Europe, Australia, and New York via Rio de Janeiro. The air route used by South African Airways to European cities has to avoid most of the remainder of the African continent because of the antagonism of black African states to South Africa's apartheid policies. Luanda, in Angola, Las Palmas, in the Canary Islands, and Ilha do Sal, on the Cape Verde Islands, are used as intermediate stopovers. Major foreign air services also are in operation to and from South Africa.

The two international airports are Jan Smuts Airport near Johannesburg and the D.F. Malan Airport at Cape Town. Other major interior airports are the Louis Botha Airport at Durban, the J.B.M. Hertzog Airport at Bloemfontein, and others at Port Elizabeth, Kimberley, and East London. (A.Ne.)

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* The constitution of the republic may be described as written, flexible, and unitary. Its framework is charted by the Republic of South Africa Constitution Act of 1961 and the extant provisions of the original constitution, the South Africa Act of 1909, which is not specifically repealed or replaced by it. The 1961 constitution was necessitated by the transition of the state from a monarchy within the Commonwealth to a republic without, which occurred after a popular referendum among white voters in 1960 had favoured by a narrow margin the establishment of a republic. The constitution is virtually the same as the constitution of 1909, except that it does not contain such terms as cabinet or prime minister. (R.B.Ba.)

The constitution provides for a state president, who is head of the republic; an Executive Council of ministers of state appointed by him; and a Parliament comprising a Senate and a House of Assembly.

The state president is elected for a term of seven years by secret ballot of an electoral college consisting of members of both houses of Parliament. He is a constitutional ruler acting on the advice of his ministers, whom he appoints from among sitting and prospective members of Parliament having the confidence of the majority in the House of Assembly, and together they form a cabinet, which, in practice, is the Executive Council. Ministers in charge of departments may not exceed 18 in number, but the state president may also appoint not more than six deputy ministers.

Powers of Parliament

Parliament has unrestricted legislative power "to make laws for the peace, order, and good government of the republic." It consists of an upper house, the Senate, and a lower house, the House of Assembly. Ordinarily, it legislates by means of bare majorities in both houses sitting separately, but in the case of the entrenched clause of the Republic of South Africa Constitution Act, which provides for the equality between the two official languages, English and Afrikaans, it must legislate by means of a two-thirds majority of all members of both houses sitting together.

Although both houses are original legislative bodies, in practice the Senate is, in the main, a revising chamber with delaying powers in respect to bills introduced in the lower house. The Senate consists of 54 members, of whom 14 are elected in the Transvaal, 11 in Cape Province, eight each in Natal and the Orange Free State, and two in South West Africa. In addition, eight members are nominated for South Africa (four of them for their knowledge of the Coloured peoples), and three for South West Africa.

The House of Assembly comprises 160 members directly elected by white voters in the republic, and six members directly elected by white voters in South West Africa. (Coloured representation in Parliament was ended in 1971, Bantu representation having ended in 1960.) Members must be adult South African citizens of white descent, and the representatives of white voters are elected from single-member constituencies delimited by a judicial commission at intervals from five to ten years. Seats are allotted to the provinces in proportion to their adult white population. Under the 1965 delimitation, the Cape returned 54 members, Natal 18, the Orange Free State 15, and the Transvaal 73. Electoral divisions within each province have an approximately equal number of voters but may exceed or fall short of the average by between 15 percent and 30 percent under a so-called loading system. (Ed.)

The provincial councils

Regional, state, and local governments. Where the central government's powers are delegated to regional and local governments, such delegation is effected within the general framework of a separate administration for each of the four racial groups (whites, Coloureds, Indians, and Africans).

Each province has boundaries, legislative powers, a provincial council elected by white voters only, and an administrator appointed by the state president. These are defined by the Republic of South Africa Constitution Act of 1961. The councils have powers over white education, hospitals, roads, town planning, and other matters. Each province has its own system of local authorities, which vary slightly. All are elected by white voters; Coloured ratepayers had the vote in the Cape but were removed from the common voters' roll in November 1971.

The Coloured population's political role is limited to the election of 40 representatives to the 60-member Coloured Persons Representative Council, the other 20 of which are nominated by the government. It is intended that some of the powers to administer the policies of the Department of Coloured Affairs will gradually be devolved to this council. Its jurisdiction, defined in the Separate Representation of Voters Act (as amended in 1966), extends to the Coloured "Group Areas" only, as the Coloured population has no "homeland" defined as a geographic unit.

Similarly, some of the powers to administer policies held by the Department of Indian Affairs are to be devolved on the Indian Advisory Council. As its name implies, however, this is a wholly nominated body. Its jurisdiction, too, extends only to local matters affecting Indian "Group Areas," there being no Indian "homeland."

The Bantu Affairs Administration Act of 1970 completed the process of removing all major aspects of government and administration of Africans from the local authorities in "white" areas to the Department of Bantu Administration and Development. Local authorities outside the "homelands" thus no longer administer African affairs within their boundaries.

The Promotion of the Bantu Self-Government Act of 1959 recognized eight tribally based African national units or "homelands"; a ninth was announced in 1970. All but two consist of several land blocks, together comprising more than 200 scattered areas. In terms of the official policy of "separate development," these are to form the so-called Bantustans of the future; although there is no intention of consolidating each into a single geographical entity, long-term policy envisages their independence. The Bantu Homelands Citizenship Act of 1970 made every African a "citizen" of one of these national units. The Ciskei, Lebowa, Bophuthatswana, and Transkei had achieved autonomous status by October 1972. Under the Transkei Constitution Act of 1963 a partly elected legislative assembly was established, which elects a cabinet. The assembly has powers over some local matters; but all its laws are subject to veto by the state president.

The Coloured and Asian communities, for neither of which a national geographical unit is planned, were provided with elected councils under the Separate Representation of Voters Act (as amended in 1966). They are intended to exercise control over matters affecting their communities where this does not conflict with national policy. (M.J.L.)

The political process. *Elections and parties.* The parliamentary electorate of just over 2,000,000 men and women is all white save for the survivors of some 300 Coloured men who were registered on the Natal common roll at the time when the Separate Representation of Voters Act of 1951 came into force and whose names remain there. No special qualifications are required for the white franchise; the voting age is eighteen years.

The National Party

The national and parliamentary political parties are white by law as well as by custom. The National Party has been in power since 1948, its programs placing emphasis on the concepts of white South African nationalism and the right to separate development for white and black. Its only setback at the polls thus far has been the loss of nine seats to the United Party—the official opposition—at the general election of 1970. Even then, its majority in Parliament remained overwhelming, as the figures for that election show: National Party (NP) 118 seats; United Party (UP) 47 and Progressive Party (PP) 1. (R.B.Ba.)

The participation of the citizen. Most significant political developments of recent years stemmed from the implementation of the "separate development" philosophy, in terms of which the nonwhite population has been systematically deprived of political participation in the conduct of national and provincial affairs, in exchange for the establishment of their own political institutions, whether Bantu or Coloured. The last African representation in Parliament (three elected whites) was abolished in 1959 under the Promotion of Bantu Self-Government Act.

The last four "Coloured" representatives in Parliament were abolished by the terms of the Separate Representation of Voters Amendment Act of 1968. The Asian population has never been represented. The Prohibition of Political Interference Act of 1968, which outlawed multiracial political parties, forced the Progressive Party to limit its membership to whites; the Liberal Party disbanded rather than do so.

The two major national African political parties—the African National Congress and the Pan-Africanist Congress—were banned in 1960 under the Unlawful Organisations Act. The only remaining African political parties were established in the Transkei for the purposes of fighting the elections there. The Democratic Party, which won the first elections in 1963, was forced into opposition by the appointment of chiefs from the apartheid-supporting Transkei National Independence Party. The latter, however, won a majority of the elected seats in the second elections in 1968. Similarly, the first election for the Coloured Council, held in 1969, produced an elected majority for the Labour Party; but government appointees put the Labour Party into opposition. The South African Communist Party is one of the largest Communist parties in Africa; but, having been formally banned, it cannot function openly.

Justice. The common law of the republic is based on Roman-Dutch law, the uncodified law of Holland that was retained after the Cape's cession to the United Kingdom in 1814. The judiciary comprises a Supreme Court with local, provincial, and appellate divisions, the High Court, and circuit courts of South West Africa, and inferior courts. The judges, headed by a chief justice, are appointed by the state president-in-council; they cannot be removed from office except by an address of both houses of Parliament, and their salaries cannot be reduced. Appeals from the divisional courts and the high and circuit courts of South West Africa lie to the appellate division of 11 judges, with a quorum of five except when the validity of an act of Parliament is in question, when the quorum is 11.

The legal system

The inferior courts comprise magistrates' courts, Bantu courts, and water courts (dealing with the use of water resources). Most civil and criminal litigation is a matter for the magistrates' courts, each responsible for a magisterial district. Regional magistrates are appointed in populous areas with criminal jurisdiction over a number of districts. The Bantu courts do not have jurisdiction over Bantu in criminal cases or where whites are involved. In 1968 some Bantu magistrates were appointed in the Transkei with jurisdiction over Bantu only. (Ed.)

Armed forces. The defense budget for 1971/72 was about R 316,500,000—a sevenfold increase since the beginning of the decade. Full-time white servicemen number over 44,000. Nonwhites serve in the defense forces only in a manual capacity, but, owing to the shortage of manpower, consideration was being given in the early 1970s to allowing Coloured personnel to play a more active role in the navy. In addition to the regular personnel, the defense force comprises an "active citizens' force," commandos, reserves, school cadets, and auxiliary services (in which nonwhites may serve). Women were taken into the services in 1970; their training has been confined to exercises in unarmed combat and the use of light firearms.

The navy has a strength of about 4,250 men; it possesses two destroyers, six frigates designed for anti-submarine warfare, about a dozen minesweepers, and auxiliary craft. The airforce has a strength of about 8,000 men and has about 240 combat aircraft, including light bombers, fighter bombers, interceptor fighters, and transport, reconnaissance, and training aircraft.

Police activities dovetail with those of the defense forces for antiguerrilla activities. About 3,000 police have received special training for this purpose. Their equipment includes armoured personnel carriers and patrol tanks. Police manpower has been severely strained by antiguerrilla activities at the national borders as well as by their activities in Rhodesia, where that country's government has requested their help in combatting insurgence. Over

1,000 police are continuously used for border security services alone.

Administration. The complexity of the administration can be understood only in the context of its division into four almost completely separate departments, one for each of the four racial groups. In general, the usual administrative portfolios (education, welfare, health, housing, etc.) apply only to whites; all these services, however, are administered for Africans, Coloureds, and Indians by (respectively) the Ministry of Bantu Administration and Development and Bantu Education, the Ministry of Coloured Affairs, and the Ministry of Indian Affairs. For all four racial groups some powers are variously but separately delegated to regional or local bodies.

Educational services. The provinces administer primary and secondary education and teacher training for whites; the department of education, arts and sciences is responsible for special and vocational schools and universities.

The universities

For whites there are four English-language universities, five Afrikaans-language universities, and one bilingual university; for nonwhites there are three for Africans, one for Coloureds, and one for Indians. White school children are taught in a single language—English or Afrikaans according to their home language. Attendance is free and compulsory between seven and 16 years of age, and books and equipment are provided. For Asians and Coloureds education is free and compulsory where the facilities exist, which varies with each local authority. The administration of education for Coloureds and Indians is the responsibility of the Ministries of Coloured Affairs and Indian Affairs and in the early 1970s it was being partially delegated to the Coloured Persons Representative Council and to the Indian Advisory Council.

African education at all levels is administered by the Department of Bantu Administration and Development and Bantu Education. Since 1964 a process has begun of decentralizing African schooling to the "homelands" areas and delegating control and some policy matters to the territorial authorities where these have been formally constituted.

Schooling for African children is neither free nor compulsory, and parents are responsible for some of the books and equipment. The number of Africans who received some schooling doubled in the 20 years following the 1953 Bantu Education Act. Expenditure, however, increased by only about 40 percent, so that per capita expenditure fell. By 1968, per capita expenditure on African education was one-eighth of the amount spent on whites. About half the schools hold double sessions; and about a quarter of the teachers are qualified. About 85 percent of African children enter primary schools, but just under half leave after the first year; only one-tenth enter secondary schools, and less than one percent reach the final matriculation class, and of these only about half pass the examination.

Health and welfare services. For whites, health and welfare services are administered partly by the central government departments and partly by the provinces, as well as by voluntary organizations, some of them subsidized. Public hospitals and clinics are available free or partly free to the indigent.

For nonwhites, these services are provided partly by the central Ministries of Health and Welfare and partly by the Ministries responsible for the racial group concerned. In the "homelands" the Department of Bantu Administration and Development controls health and welfare services. Voluntary agencies concerned with welfare services for nonwhites must be registered to qualify for a subsidy and may not be registered if they are multiracial in constitution.

Housing. Local authorities are responsible for public housing, which is segregated by race and controlled by different departments. Most projects are not subsidized; but some subeconomic units are built mainly for nonwhites. A department of Community Development was established in 1961 to oversee a national policy. By 1970, 50 percent of new housing built by local authorities was being used for resettling Coloured and Indian communities removed from one area to another under the Group Areas Act of 1950; the consequence is a severe shortage of housing for these groups. The Department of Bantu Administration and Development initiated a large program of township development in the "homelands" during the 1960s, mainly in response to the need to provide housing for those Africans—about 1,000,000 people were removed from white areas in the implementation of the "separate development" policy.

Police. The regular police are organized nationally and comprise about 33,000 regulars and 15,000 reservists. There are about equal numbers of whites and nonwhites, the latter being forbidden to carry arms. The Security Police are organized separately. A third police force was established in 1969 in terms of the Public Service Amendment Act and the General Law Amendment Act. These acts established a Bureau for State Security, responsible only to the prime minister, and not subject to treasury approval for its spending. The latter law makes it an offense to possess or publish information about military, security, or police matters deemed prejudicial to state security; it also allows any government minister to issue a certificate preventing evidence being presented at court.

At the beginning of 1970 more than 800 people were serving sentences under the various security laws; the number detained without trial is not given officially. The daily average prison population rose sharply in the 1960s to over 90,000 in 1969 to 1970, the total number of admissions in that year being about 750,000. Of these, the majority were imprisonments for statutory offences against laws (the so-called pass laws) that define the right of Africans to live and work in "white" areas and that do not apply to other racial groups.

Social conditions. *Wages and cost of living.* The most obvious feature of social conditions is the difference between the wages and cost of living of Europeans and that of nonwhites. In 1969 the average per capita income for whites was R 95 per month, while that for Africans was R 7 per month. In the construction industry a white worker earns five times as much as an African, twice as much as an Asian, and three times as much as a Coloured worker. In the mining industry whites earn 12 times as much as Africans, four times as much as Asians, and nearly five times as much as Coloureds. The retail trades are among the least inequitable: here, a white worker earns over three times that of an African, two-and-a-half times as much as a Coloured, and one-and-a-half times as much as an Asian.

White and nonwhite income levels

Such wage differentials derive largely from two factors: the Industrial Conciliation Amendment Act of 1956, under which administrative determinations made periodically define the levels of employment that members of each racial group may occupy and ensure a racial hierarchy; and the Industrial Conciliation Act of 1956, which excludes Africans from membership in registered trade unions and, hence, excludes them from conciliatory machinery, makes it an offence for them to strike. The consequent wage differentials result in income levels below the official poverty datum line. In Soweto, an African township serving the employment area of Johannesburg, surveys show that 75 percent of families live below the poverty datum line. Income in the Bantu "homelands"—estimated in 1969 at R 53 per capita a year—is about two-thirds of the essential minimum requirements. In 1966 white households surveyed in the ten principal urban areas had an average income of R 4,637.

Health conditions. Surveys conducted in the 1950s and 1960s showed a relatively high degree of malnutrition among nonwhites, especially Africans and their children. Although infant mortality figures are not given to Africans, local authority statistics as well as research by doctors and social workers indicate figures as high as 200 to 300 per thousand in some urban areas and 400 to 500 per thousand in some rural areas. For whites the infant mortality rate has remained constant at about 27 per thousand. The rate per thousand of some notifiable diseases in 1965 was as follows: tuberculosis—whites 37, Coloured 518, Asians 248, Africans 459; kwashiorkor (a

nutritional deficiency disease)—whites 0.3, Coloureds 42, Asians five, Africans 100; typhoid fever—whites three, Coloureds 11, Asians seven, Africans 38; diphtheria—whites five, Coloureds ten, Asians six, Africans ten.

Social and economic divisions. It is the purpose of the "separate development" policy to segregate the races as far as possible in every aspect of their lives. The country is divided into one European area and eight African areas. The European area is divided in terms of the 1950 Group Area Act (as amended) into segregated areas in each of which only members of a designated race may live, own land, or occupy premises. In addition to this residential segregation, there is political, social, and economic discrimination. Marriages and sexual relations between the races are outlawed by the Prohibition of Mixed Marriages Act of 1949 and the Immorality Amendment Act of 1950. The Reservation of Separate Amenities Act of 1953 forbids persons of different races to use the same public amenities or transport facilities and also gives central or local authorities power to disallow racial mixing in places of public entertainment, cultural institutions, and sporting facilities. (M.J.L.)

Cultural life and institutions. The ten or more languages spoken by South Africa's 20,000,000 people reflect the country's cultural plurality. The largest group, the Africans, is the most eclectic in culture, having adopted, for example, a host of European ways as well as the "clicks" of the aboriginal Khoi-san languages. The Africans also form the group the least involved in the conscious perpetuation of their traditions. In custom, music, dance, and in a multiplicity of languages they retain a solid body of African culture.

South African artists, however, are generally only known and recognized when they appeal to European ideas. Painters such as Sydney Kumalo and Gerard Sekoto are still less known than Alexis Preller or Jean Welz; authors such as the black writers Ezekiel Mphahlele or Peter Abrahams (both now living abroad) are read less at home than they are abroad, whereas such white writers as Alan Paton, Nadine Gordimer, Dan Jacobson, and Laurens van der Post write for a South African as well as a world readership. Like Afrikaner authors, however, many African writers are prisoners of minority languages unknown outside South Africa. The lot of African authors, however, contrasts vividly with that of Afrikaner writers, who are laudably supported by a publishing industry, which encourages poets such as D.J. Opperman and Elisabeth Eybers and comparatively radical writers such as the "sestigers" ("men of the sixties"), Etienne Leroux, André Brink, and others. Coloured authors, also, are published in Afrikaans, whereas Africans can write only for a school textbook, and hence noncontroversial, market.

Elite African theatre and musical groups (especially choirs) exist, but for the most part the Africans perform to a localized, tradition-minded public. The vigorous cultural nationalism of the Afrikaners runs through the life of the country. The performing arts are encouraged by Afrikaner-led state bodies. Theatre, music, ballet, and opera reflect European styles, with some provincial influence, little of it stemming from African models.

Afrikaner culture

The Federation of Afrikaner Cultural Societies (FAK) is part of a closely-knit group of organizations concerned with fostering Afrikaner culture and interests. It not only organizes national activity but also struggles for higher standards in the cinema, in magazines, and in folk music and dancing.

The English community attempts no parallel activity, secure in the awareness that their language and traditions exist in an international context. The Africans find themselves the subject of much cultural activity undertaken on their behalf, mainly by the white-run Department of Bantu Education. As, however, African culture exists in a continental context, the Africans may feel their culture less in jeopardy than do the Afrikaners.

Only in the press is Afrikaner dominance successfully challenged: the English-language press, organized into strong combines of newspaper groups (such as the Argus evening group) in the main cities, dwarfs in circulation and wealth the younger, more doctrinaire Afrikaans press. The African-language newspapers are white-owned, and their standards are not high. The predominance of the English language press, however, reflects English economic, rather than political, success. The newspapers, in general, to their credit, assume a readership more cultivated than is perhaps the case.

In a country that has no television, radio constitutes an important force. Afrikaner control of Afrikaans and English broadcasting, as well as of the Radio Bantu services, which are broadcast in seven languages, reflects the underlying cultural pattern, which is characterized throughout by Afrikaner striving, the survival of the English language, and the enduring vigour of African culture. (R.V.)

Prospects for the future. South Africa's economic development outstrips by far that of any other African country. The fast rate of growth has brought with it a severe lack of skilled labour. The African population not being allowed to do skilled labour outside the Bantu homelands in South Africa, the government has instead encouraged white immigration as much as possible. The country's rich mineral resource base lacks only two important items—a sufficient water supply and natural oil. Large sums of money are being spent annually to conserve water—for example, on the Orange River Project—and to drill for oil.

Politically, South Africa's main internal problems are how to implement its policy of "separate development" and how to reverse the movement of Africans now taking place into the urban areas from the Bantu "homelands." Internationally, South Africa finds no support for its racial policies, and its stand on South West Africa has aroused the ire of a great many countries. As the country is, however, the strongest anticommunist country in Africa; as it is the world's main gold producing country, forming a vital link in the world's monetary system; and as it is the guardian of the vitally important Cape sea route, South Africa is able to command respect and even to count on protection from the major Western powers. (A.Ne.)

BIBLIOGRAPHY

General (including the physical landscape): MONICA COLE, *South Africa*, 2nd ed. (1966); R.J. HARRISON CHURCH, *Africa and the Islands* (1964); N.C. POLLOCK, *Africa* (1968); JOHN H. WELLINGTON, *Southern Africa*, vol. 1 (1955).

People and population: PAUL GINIEWSKI, *Bantustans: A Trek Towards the Future* (1961); A. NEL, *Geographical Aspects of Apartheid in South Africa* (1962); BUREAU OF STATISTICS, PRETORIA, *South African Statistics* (biennial); *The Standard Bank Annual Economic Survey*.

Economic and social conditions: NATHAN HURWITZ and OWEN WILLIAMS, *The Economic Framework of South Africa* (1962); *Report of the Commission of Inquiry into the Co-ordination of Transport in South Africa*, R.P. 32 (1969); BUREAU OF STATISTICS, PRETORIA, *South African Statistics* (biennial); *Statistical Yearbook* (1966); *Survey of Race Relations*, published annually by the South African Institute of Race Relations, Johannesburg; *State of South Africa*, published annually by the Legal and General Assurance Society Ltd., Johannesburg.

Race relations: The annual *Survey of Race Relations in South Africa*, ed. by MURIEL HORRELL, is the most comprehensive and objective record of affairs in the country. Complementary to it are the special publications of the Institute of Race Relations, such as *The African Reserves of S.A., Legislation and Race Relations*, and *Action, Reaction and Counteraction*, all ed. by MURIELL HORRELL; and *Bantu Manpower and Education* by E.G. MALHERBE (1969). A standard yearbook *State of South Africa* contains generally uncontroversial data. The United Nations regularly publishes material mostly critical of apartheid, such as SEAN GERVASI, *Industrialization, Foreign Capital and Forced Labour in S.A.*; see also UNESCO, *Apartheid: Its Effect on Education, Science, Culture and Information* (1967).

(A.Ne./R.B.Ba.)

South America

South America, the southern continent of the Western Hemisphere, is situated between longitudes 35° and 80° W and latitudes 12° N and 55° S. It is the fourth-largest of the continents, having an approximate area of

6,874,600 square miles (17,805,200 square kilometres), or about 12 percent of the land surface of the Earth. It ranks fifth among the continents in terms of population, with an estimated population of 195,000,000 people at the start of the present decade.

Political divisions

Politically, the continent and its adjacent islands are divided into 11 sovereign republics and three dependencies. The republics are Argentina, Bolivia, Brazil, Chile, Colombia, Ecuador, Guyana, Paraguay, Peru, Uruguay, and Venezuela; the dependencies are the Falkland Islands (British, but also claimed by Argentina), French Guiana, and Surinam (Dutch Guiana).

South America is massive in form and roughly triangular in shape, being broad in the north and tapering to a point in the south at Tierra del Fuego. It is bounded by the Caribbean Sea to the northwest, the Atlantic Ocean to the northeast, east, and southeast, by Drake Passage to the south, and by the Pacific Ocean to the west. In the northwest it is joined to North America by the Isthmus of Panama, which forms a land bridge narrowing to about 50 miles at one point; to the north, the islands of the West Indies roughly form a chain running from Trinidad to Florida.

In relation to the area of the continent, the South American coastline, which is 15,803 miles in length, is exceptionally short—one mile of coast for every 435 square miles of area.

The structure of South America is relatively simple. The west is dominated by the Andean mountain ranges, which run roughly north to south for about 4,500 miles; the centre and the east consist of a series of plateaus and extensive depressions filled with alluvia. Thus the margins of the continent are higher than the interior—a characteristic that produces profound effects on local topography, drainage, and land occupation.

South America rises abruptly from the ocean; it has a narrow continental shelf, especially along the west coast, where the seabed drops sharply down to deep submarine trenches, which, within 50 miles of the shore, sometimes reach depths greater than 25,000 feet. North of latitude 40° S, the west coastal plains are no wider than 50 miles, and in Peru they are largely absent; good natural harbours are missing, except for Guayaquil in Ecuador, and Buenaventura in Colombia, both situated along estuaries. The east coast follows the edge of the Brazilian plateaus. The plains are relatively narrow, but there are some fine natural harbours, such as Recife, Salvador, and Rio de Janeiro. In the south, the deep indentation of the Río de la Plata forms a great gulf, south of which the Patagonian coast is relatively regular.

Except for the fjord country of southern Chile and the Tierra del Fuego region, few islands are located off the continent. To the north, Trinidad is separated from the mainland by the Gulf of Paria, which is bordered along its northern edge by a continuation of the Venezuelan Andes of Cumaná. To the west of Trinidad are such small islands as Margarita and Curaçao. Off the easternmost extremity of the continent lies the Fernando de Noronha group; east of the southern tip of the continent are the Falkland Islands. The west coast is bordered by some very small islands close to the Peruvian coast; considerably further west are the Juan Fernández and Galápagos Islands.

Since—both to the east and the west—the lines of maximum altitude are not far from the sea, and the land, therefore, slopes toward the interior, drainage is largely away from the periphery of the continent and toward its centre. A few independent rivers have made their way to the sea; these tend to combine many streams into large waterways. Among the larger rivers are the Magdalena, the Orinoco, the São Francisco, and the Río de la Plata (fed by the Paraguay and Paraná rivers); all these are, however, dwarfed by the Amazon, in terms of volume of flow and area drained, the largest river in the world. Only Europe has more running water and less desert than South America.

An important feature of the continent is its dissymmetry. Since the Andes border the Pacific, the gigantic backbone they create divides the continent into two parts differing greatly in both size and character. This unequal division has had profound effects on climate, wildlife, and even human settlement.

Latitudinal extent

No other continent has a greater latitudinal extent or penetrates so far to the south. Though its northern part is crossed by the Equator and 80 percent of its landmass is located within the tropical zone, South America also extends into the sub-Antarctic. Because of the high altitudes and wide extent of the Andes within the tropics, extensive zones of temperate or cold climate exist in the vicinity of the Equator—a circumstance that is unique. These conditions produce a great range of climates; the unrivalled diversity of ecological zones is probably the most prominent characteristic of South American geography.

The diversity is reflected in South American wildlife and has also influenced man's activities. The diversity of rock formations is paralleled by important and extremely varied mineral resources. A good part of the continent consists of ancient formations, rich in minerals; when the Andes were uplifted, a series of important minerals were also being deposited. Because of the wide range of soils and climates, renewable resources are also highly diversified.

Originally a large part of South America—except for the arid section, the upper zones of the Andes, and the southeast—was covered with forests of several types; these forests have now been cleared to a large extent, notably in eastern and southeastern Brazil. The forests of the huge Amazon Basin are, however, still almost untouched, although clearing is now in progress. Natural conditions permit a great variation in land use, ranging from the exploitation of natural resources, such as timber and rubber, to the cultivation of a wide variety of crops, from tropical plants to temperate cereals.

The arrival of man in South America took place in relatively recent times. He certainly came south from North America but immigrants probably also arrived across the Pacific Ocean from the Malayo-Polynesian regions. In the pre-Columbian era, great empires flourished in places along the west coast and in intermontane basins where living conditions were often difficult but communications easy and techniques of managing natural resources within the reach of pre-industrial civilizations.

The other parts of South America were thinly settled by men with less advanced technical and sociological skills, who were nevertheless successful in making a series of ecological adaptations. In modern times, immigrants, mostly of Spanish or Portugese origin but also from other ethnic groups, have introduced western European civilization to South America. Negroes of African origin arrived as slaves and their descendants now constitute a substantial part of the population of some countries, particularly Brazil.

Some parts of the continent are now industrialized, though most regions still follow an agricultural way of life. The possibilities of development, offered by the wealth of mineral products and renewable resources, are considerable. The human devastation of nature and of natural resources has nevertheless already reached alarming proportions, even though South America was relatively well protected from such disasters until recent times. Ecologists find this development particularly disconcerting because the human population is increasing at a rapid rate through the whole continent; between 1963–69, for example, the annual rate of increase in tropical South America was 3 percent. In spite of the wealth of the continent's resources, the trend could prove to be an obstacle to the optimum development of most South American countries. (See the articles on individual South American countries, as well as the articles FALKLAND ISLANDS AND DEPENDENCIES; FRENCH GUIANA; and SURINAM; see also individual articles on the component states of Brazil, and the city articles BOGOTA, BRASILIA, BUENOS AIRES, CARACAS, LIMA, MONTEVIDEO, RECIFE, RIO DE JANEIRO, SANTIAGO, and SAO PAULO. There are also articles on major physical features. See also CONTINENTAL DRIFT; and CONTINENTS, DEVELOPMENT OF.)

This article is divided into the following sections:

I. Geological history
  Basic structural units
  Geological development
II. Physical geography
  Relief
    An overview
    Physiographic provinces
  Climate
    Factors influencing climate
    Climatic regions
  Drainage
    River systems
    Lake systems
    Marshes and swamps
  Soils
    Soil regions
    Soil problems
III. Vegetation and animal life
  Vegetation
  Animal life
    Fish and bird life
    Mammals
    Ecological communities
    Protection for animal life
IV. Natural resources
  Mineral resources
    Mineral fuels
    Metallic deposits
    Nonmetallic deposits
  Biological resources
    Botanical resources
    Animal resources
V. Human resources
  Ethnic origins and migrations
    Indians
    Iberians
    Africans
    Postindependence overseas immigrants
  Population and ecological distribution
    The present population
    Ecological distribution
    Sociocultural factors
    Linguistic patterns
    Sociological changes
VI. Political geography
  Historic evolution
    The nature of European influences
    The cultural heritage
    Economic and social factors
  20th-century modifications
VII. Resource development
  Industries
  Electric power and irrigation
  Agriculture and land use
VIII. Commerce
  Transportation
  Internal trade
  External trade
IX. Demographic patterns
  Current trends
    The growth rate
    The fertility rate
  Effects of rapid population increase

I. Geological history

The geological history of South America is complex and to some extent still unknown. The present geological structure is the result of the movement that uplifted the Andean mountain ranges and made itself felt throughout the entire continent.

BASIC STRUCTURAL UNITS

The five structural units

The geological structure of South America is dominated by five geotectonic units (structural rock masses of the Earth's crust), the history of which can be traced from the end of the Precambrian Era (570,000,000 years ago). The evolution of these units has followed the same general tendencies throughout subsequent geological time. The five units are the continental shields (stable masses of Precambrian rock), the epicontinental basins (*i.e.*, those situated on the continental plateau), the insular shields, the belt of pericratonic basins (*i.e.*, the basins surrounding large relatively immobile parts of the Earth's crust), and the Andean Geosyncline (*i.e.*, the gigantic linear trough in which sediments accumulated and which runs from north to south on the western side of the continent).

The continental shields consist of three old landmasses that together form the vast Guiana–Brazilian Shield that constitutes the greater part of northeastern South America. The complex, as a whole, is of crystalline rock, and—unlike the Andes—has long been quiescent. Its three parts are, firstly, the Guiana Shield, situated between the Orinoco and Amazon rivers; secondly, the Central Brazilian Shield, between the Amazon, the Parnaiba–São Francisco Basin, and the Paraná Basin, together with a long spur extending to central Paraguay; and, thirdly, the Eastern Brazilian Shield, covering eastern Brazil and Uruguay.

The epicontinental basins consist of three huge basins—the Amazon, Parnaiba–São Francisco, and Paraná basins—situated in weaker areas between large and relatively immobile masses. They have great unity of structure and have undergone uniform sedimentation, having been at first invaded by seas of the Paleozoic Era (from 570,000,000 to 225,000,000 years ago), then filled with continental sediments and volcanic deposits. These sediments have not been subjected to folding in Recent (Holocene) time (during the last 10,000 years); only tectonic movements (movements of the Earth's crust) resulting from volcanism have changed their structure.

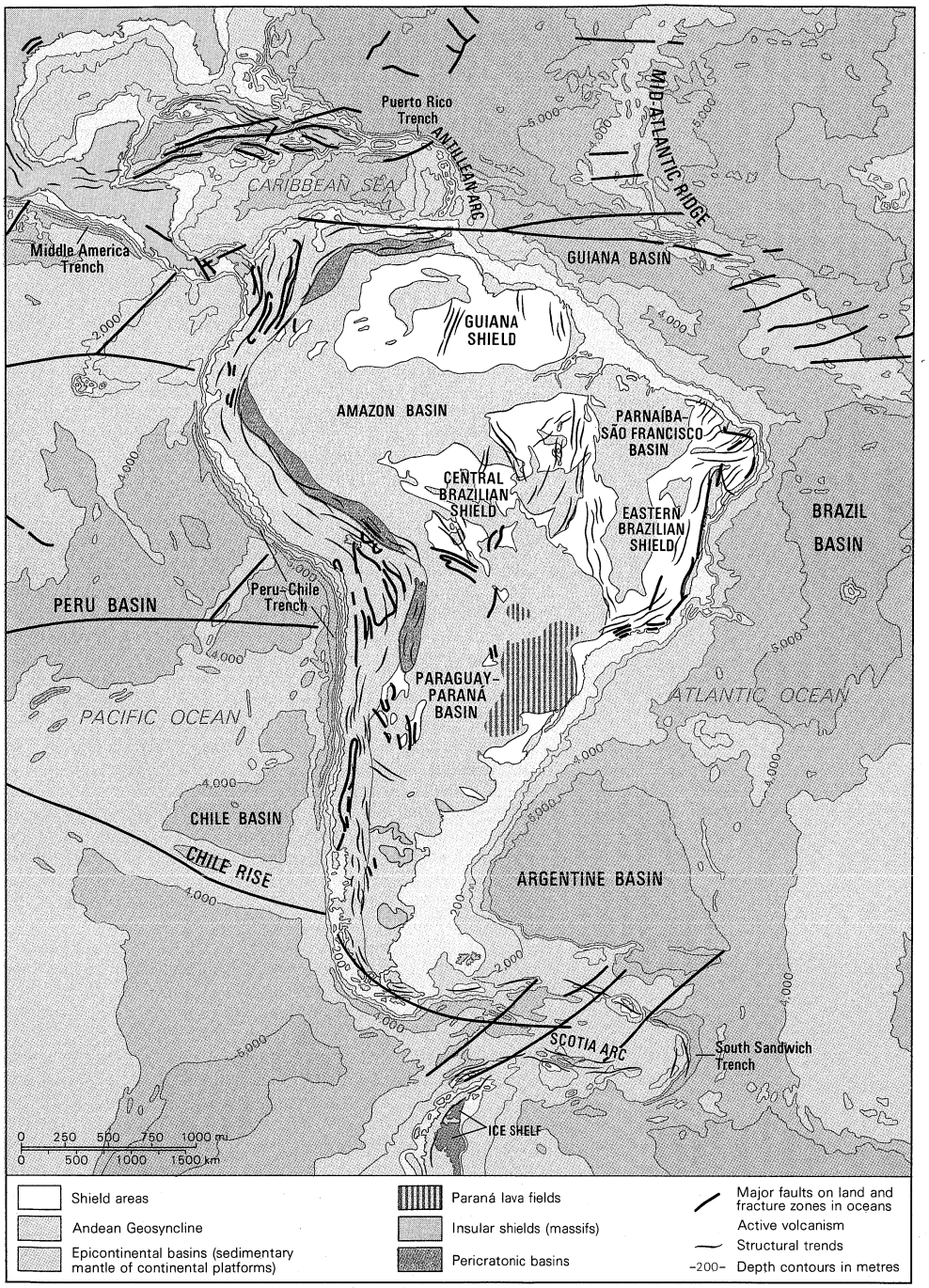
The Amazon Basin is about 540,000 square miles in area and consists mostly of marine sediments from 280,000,000 to 570,000,000 years old, which underlie continental deposits. The Parnaiba–São Francisco Basin, in northeastern Brazil, is filled with sediments almost 10,000 feet deep; the oldest of these are marine limestone deposits, covered by sandstone, conglomerate deposits from rivers, and continental sediments which accumulated during the Mesozoic Era (from 225,000,000 to 65,000,000 years ago), together with some marine deposits in the north. The Paraná Basin, more than 500,000 square miles in area, is filled with sediments about 6,500 feet thick and with up to 5,000 feet of basaltic lava. Continental sediments of sand and silt, from 190,000,000 to 295,000,000 years old, are covered with lava deposits which extend over an area of more than 460,000 square miles—the largest lava field on Earth.

The insular shields, which lie to the east of the Andes, are Precambrian massifs (mountainous masses) situated in Argentina; they include the Pampine Sierra and the Patagonian and Deseado massifs and may extend along the continental shelf as far as the Falkland Islands.

The belt of pericratonic basins lies along the western edge of the Brazilian shields and has a tendency to subsidence. It is long and narrow and runs from north of the Orinoco through Acre, Beni, and Chaco to south of the Río de la Plata. From the Precambrian Era to the beginning of the Mesozoic Era (about 225,000,000 years ago), seas invaded this slowly sinking region, depositing sediments, which slowly merge with continental strata on the east. The belt as a whole constitutes a transitional region before the Andean Geosyncline is reached.

The Andean Geosyncline

The Andean Geosyncline occupies a zone stretching from the island of Trinidad on the northern coast to Tierra del Fuego in the extreme south. The original geosyncline contained sediments from the western highlands and from a continental mass that has now subsided under the Pacific, as well as of marine deposits that accumulated from the time of the Late Precambrian Era onwards. Rapid and persistent subsidence was counterbalanced by the occurrence of extensive sedimentation. Movements of the Earth's crust began in the Paleozoic Era and were accompanied by intense volcanic activity, but the main activity began in the Early Mesozoic Era, reaching a maximum during the Middle and Upper Tertiary Period (from about 26,000,000 to 2,500,000 years ago). Movements continued up to the Pleistocene Epoch (from 2,500,000 to 10,000 years ago). Very intense volcanic activity occurred from the Jurassic Period (from 190,000,000 to 136,000,000 years ago) to the Tertiary Period (from 65,000,000 to 2,500,000 years ago). Plutons (bodies of rock formed beneath the Earth's crust by solidification from a molten or partly molten state), running from Colombia in the north to Cape Horn in the



Structural features of South America.

south, form the bases of many of the cordilleras (groups of mountain ranges) and even the summits of important massifs (mountainous masses).

GEOLOGICAL DEVELOPMENT

The Precambrian Era is divided into early and late periods. The Serra do Mar and the "sugar loaf" formations in the Baía de Guanabara (Bay of Guanabara), both in southern Brazil, both belong to the early period; the so-called Minas series (a long belt of metamorphosed sediments located in the State of Minas Gerais, Brazil), on the other hand, which contains considerable quantities of valuable minerals, belongs to the late period.

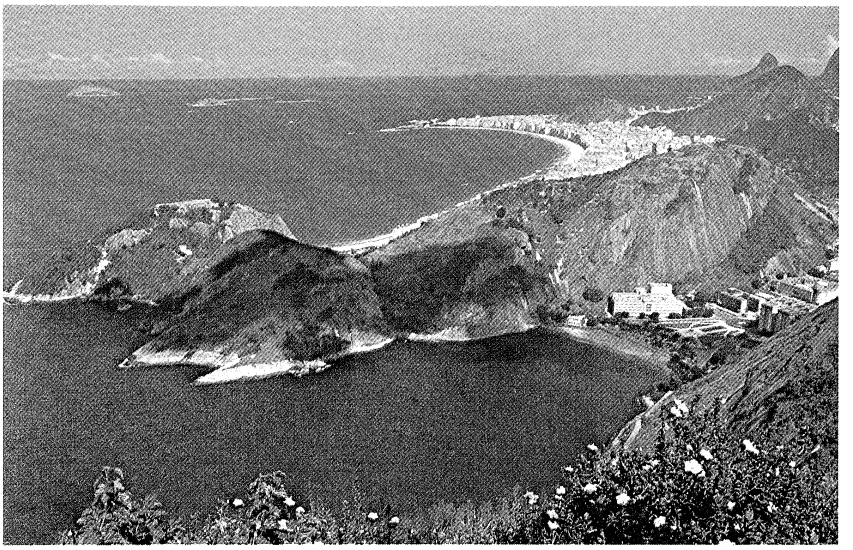
During the Cambrian Period (from 570,000,000 to 500,000,000 years ago) marine deposits accumulated in the Andean syncline and in the Amazon region of Brazil. Continental deposits of glacial or fluvial origin were laid down in various other districts, including northern Paraguay and the Brazilian Mato Grosso.

During the Ordovician Period (from 500,000,000 to 430,000,000 years ago) only the Brazilian shields were not beneath the sea, which covered the Gran Chaco region, the Paraguay Basin, and the whole of Patagonia. Thick beds of shale (laminated sediments consisting mostly of clay) and sandstone were consequently scattered from Huánuco, Peru, to Rioja, Argentina.

Those parts of the geosyncline that extended from Tierra del Fuego to central Peru and from Quito in Ecuador to the Orinoco Basin were uplifted during the Silurian Period (from 430,000,000 to 395,000,000 years ago). Between these two ranges the sea invaded the Amazon Basin, and even the Parnaiba–São Francisco Basin. A geosyncline extended from Bolivia to the mouth of the Río de la Plata and to northeastern Argentina. During the late Silurian Period the sea regressed from almost all the land, except the Andean Geosyncline.

During the Devonian Period (from 395,000,000 to 345,000,000 years ago) the sea again invaded the whole


NORTH AMERICA
CARIBBEAN SEA
ATLANTIC OCEAN
PACIFIC OCEAN
SOUTH AMERICA
BRAZIL
VENEZUELA
COLOMBIA
ECUADOR
PERU
BOLIVIA
PARAGUAY
CHILE
ARGENTINA
URUGUAY
GUYANA
SURINAME
FRENCH GUIANA
PLANALTO DO MATO GROSSO
BRAZILIAN HIGHLANDS
PAMPAS
PATAGONIA
FALKLAND ISLANDS
SOUTH GEORGIA
SOUTH SHETLAND ISLANDS
SOUTH ORKNEY ISLANDS
ANTARCTICA
Drake Passage
Scotia Sea
Tropic of Cancer
Tropic of Capricorn
Equator
Caracas
Georgetown
Paramaribo
Cayenne
Bogotá
Quito
Lima
La Paz
Asunción
Montevideo
Buenos Aires
Santiago
Brasília
Rio de Janeiro
São Paulo
Belo Horizonte
Salvador
Fortaleza
Manaus
Belém
Stanley
Amazon
Paraguay
Paraná
Uruguay
Orinoco
Magdalena
Negro
Madeira
Tocantins
São Francisco
Chaco
Gran Bermejo
Pilcomayo
Andes
Aconcagua
Mercedes
Rosario
Córdoba
Santa Fe
Bahía Blanca
Golfo San Matías
Golfo San Jorge
Golfo de Penas
Strait of Magellan
Punta Arenas
Isla Grande de Tierra del Fuego
Isla de Chiloé
Archipiélago de los Chonos
Isla Wellington
Isla de Pascua
Isla de Malpelo
Galapagos Islands
Isla San Ambrosio
Isla San Félix
Isla Juan Fernández
Isla de Marajó
Serra do Roncador
Serra Geral de Goiás
Serra do Espinhaço
Serra dos Parecis
Pantanal
Lagoa dos Patos
Lagoa Mirim
Lago Titicaca
Lago de Poopó
Lago Argentino
Lago Buenos Aires
Bahía Grande
Golfo de Penas
Península Valdés
Península Mejillones
Península Taitao
Península Brunswick
Península Antonio Varas
Península Muñoz Gamero
Península Hardy
Península Byron
Península Forster
Península Coronation
Península Mitre
Península Trinidad
Península Tres Montes
Península Graham
Península Palmer
Península Antarctica
Península Duse
Península Ellsworth
Península Weddell
Península Jason
Península Joinville
Península Danco
Península Louis Philippe
Península Trinity
Península Hope
Península Tabarin
Península Gerlache
Península Brabant
Península Arrowsmith
Península Hurd
Península Potter
Península Byers
Península Livingston
Península Hero
Península Fildes
Península Barton
Península Keller
Península Ardley
Península Maxwell
Península Nelson
Península Dufayel
Península Admiralty
Península Ezcurra



Exposed Precambrian rock along the coast at Rio de Janeiro, Brazil. The "sugar-loaves" in Baía de Guanabara are among the oldest rock formations of the continent.
Carl Purcell

Andean Geosyncline and later joined Amazonia and the Parnaiba Basin to Patagonia. During the later part of the period it regressed once more when mountain-building movements occurred in the Andean region. Much sedimentation occurred during this period, with sands being deposited in the Andean Geosyncline, sandstones in Amazonia, and conglomerates and sandstones in the Parnaiba Basin.

The Carboniferous Period

During the Carboniferous Period (from 345,000,000 to 280,000,000 years ago) the sea was at first restricted to Chile, western Argentina, central Colombia, and part of the Parnaiba Basin but later invaded a large area from west of Lake Titicaca to Venezuela, depositing deep layers of sandstone in the Amazon Basin and limestone and shales in Peru and northern Colombia. Glaciation occurred south of a line from Lake Titicaca to São Paulo, as a result of which glacial sediments and moraines (deposits of boulders, gravels, sands, and clays left by glaciers) were laid down in the Paraná Basin.

During the earlier part of the Permian Period (from 280,000,000 to 225,000,000 years ago) a shallow sea invaded Bolivia; later in this period, however, the sea withdrew from the entire continent. Red beds (red coloured sedimentary rocks) and continental sandstones were deposited on the border of a geosyncline which extended from Peru to northern Argentina. In Peru, volcanic activity resulted in the depositing of volcanic detritus that had been explosively ejected from volcanic vents. The Andean region of the continent as a whole was subjected to movements of the Earth's crust that permitted basaltic or andesitic magma (mobile rock material of volcanic origin) to flow out.

South America was free from invasion by seas during the Triassic Period (from 225,000,000 to 190,000,000 years ago); red beds were deposited from Colombia in the north to Patagonia in the south, sometimes attaining thicknesses of almost 10,000 feet. In the later part of the period, the Paraná Basin became a huge desert; sandstones deposited at this time are at least partly of windborne origin and now cover more than 500,000 square miles.

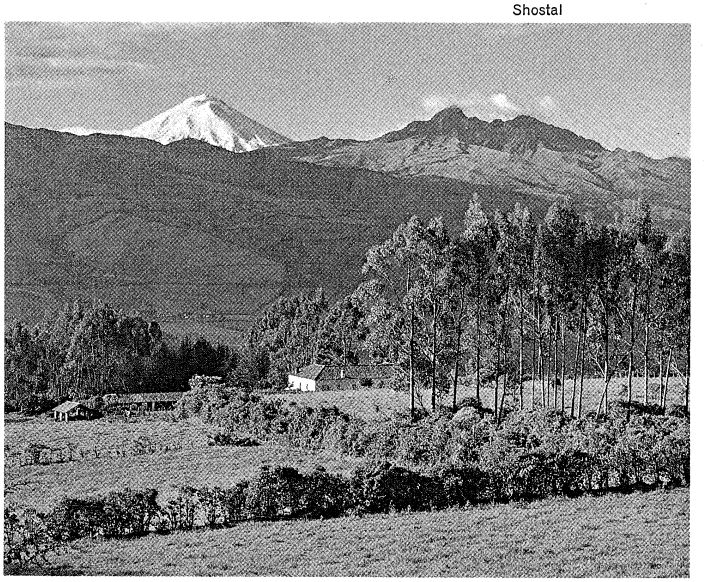
During the Jurassic Period (from 190,000,000 to 136,000,000 years ago) the continent was generally not covered by the sea, although it periodically invaded some districts in the Andean region, depositing limestone beds containing ammonites (one of a large extinct group of mollusks) in many places from Colombia to Patagonia; the Chilean coastal plain (associated with the Andean Geosyncline because of the downward warping of the western section of the Chilean Andes) also remained maritime. There was much volcanic activity on the east side of the geosyncline; in Peru and western Chile it resulted in the depositing of lava and tuffs (rocks formed of compacted volcanic fragments) from about 30,000 to 60,000 feet thick, forming one of the largest volcanic masses on Earth. Volcanic activity continued to the end of the Mesozoic Era, when it reached its greatest extent, notably in the Paraná Basin where lava flows about 5,000 feet thick occurred, as well as along a lengthy belt from Venezuela to the lower Amazon and to the Parnaiba Basin. At the end of the Jurassic Period the rise of the Andes began.

Beginning of the Andes

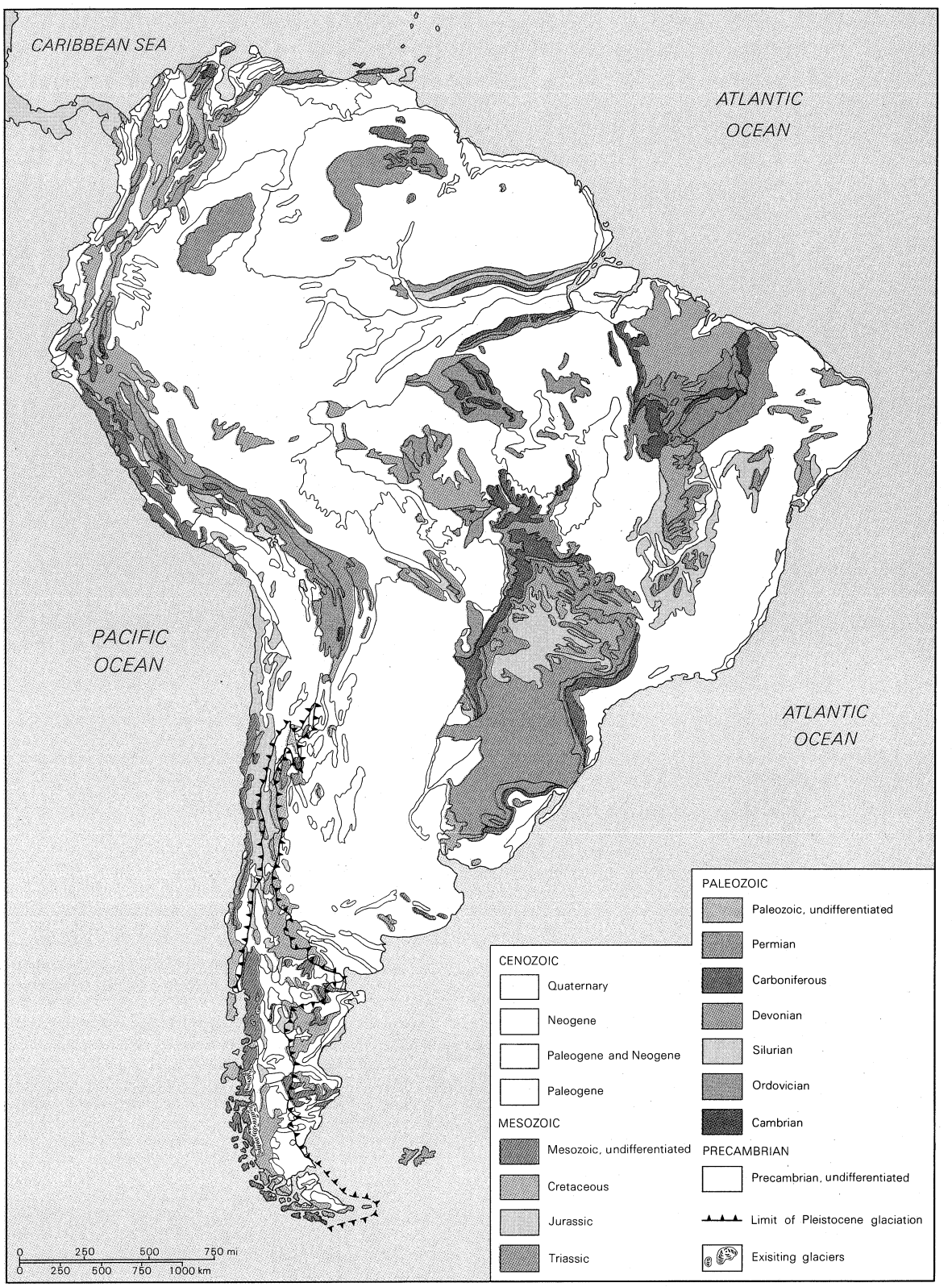
During the Cretaceous Period (from 136,000,000 to 65,000,000 years ago) seas invaded the larger part of the continent, with the Precambrian shields remaining as islands. The northern part of the Andean Geosyncline was still sinking rapidly, as indicated by the deposits of limestone about 46,000 feet thick that are to be found in Colombia. Continental sedimentation was active in the Amazonian, Parnaiba, and São Francisco basins, while layers of calcareous sand were deposited over the Brazilian Shield.

During the final part of the Cretaceous Period, the rise of the Andes reached its first maximum of activity with the uplifting of the Cordillera Central of Colombia, the Cordillera of Ecuador, the Cordillera Occidental of Peru, and the main ranges of Chile and Argentina. Mountain-building activity was strong along the length of the western Andes, and many intrusive rocks coalesced to form large and complex batholiths (very large masses of

Shostal



Cotopaxi, an active volcano in the Andes of Ecuador.

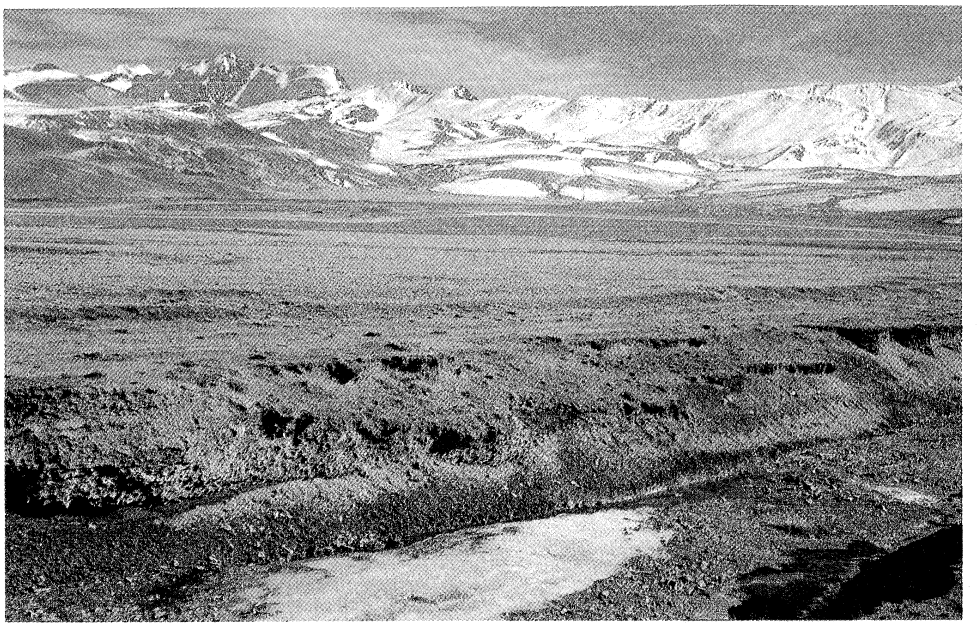


Geological structure of South America.

intrusive igneous rock), which are visible from Peru to Chile. Volcanic activity continued, and erosion resulted in continental sedimentation occurring in Peru, Bolivia, and Argentina (including the depositing of sandstones with abundant dinosaur remains in Patagonia). The sea regressed from the Amazon Basin at this time.

During the Eocene Epoch (from 54,000,000 to 38,000,000 years ago) of the Tertiary Period (from 65,000,000 to 2,500,000 years ago), great changes occurred. The sea regressed almost everywhere, except in the northwest and in the centre of Patagonia between the Deseado massif and the cordilleras.

During the Oligocene Epoch (from 38,000,000 to 26,000,000 years ago) the sea returned, invading eastern Peru. The main event during the Tertiary Period was, however, the persistent uplifting and folding of the Andes, which assumed their present form in the Miocene Epoch (from 26,000,000 to 7,000,000 years ago). From this time onward the range formed a huge barrier and the gate between the Amazon region to the Pacific via Ecua-



The Cordillera Oriental of the Andes, near La Cumbre, Bolivia.
Shostal

dor was finally closed. The Amazon Basin at first formed a huge freshwater lake but later drained into the Atlantic. At the same time the uplifting of the Andes determined mountain-building movements on their eastern border; nappes (laterally displaced, faulted, and overturned folds), sometimes more than 3,000 feet thick and several hundreds of miles long, were projected eastward against the Brazilian plateaus. Basins became filled with sediments eroded from the surrounding mountains, and much volcanic activity occurred.

Developments in the Pleistocene Epoch

During the Pleistocene Epoch (from 2,500,000 to 10,000 years ago) sedimentation continued and the general uplifting of Patagonia occurred. The Pleistocene Epoch was mainly marked by important glaciation. Ice caps covered the Andes from Tierra del Fuego to Ecuador as many as four times, reaching the Atlantic in the south where moraines are to be found today in Patagonia; glacial terraces covered the plateaus at these times. Much loess (a loamy deposit, formed by the wind) deposition took place, particularly in the Argentine Pampas (grassy plains), which are almost entirely covered by a thin layer of continental deposits of the Quaternary Period (the last 2,500,000 years). Deposits from rivers or lakes covered more than 38,000 square miles in the Pantanal (the marshy plains in the Paraguay River Basin). Volcanic activity continued, especially in the Andes. Thick deposits of volcanic ash were laid down in the inter-Andean trough extending from Chile to Ecuador. Climatic changes following the last glaciation produced laterization (the formation of reddish residual soil, leached of silica and containing concentrations of iron and aluminum hydroxide) and large-scale soil modifications in Brazil and the Guianas (Guyana, French Guiana, and Surinam). Gigantic lakes formed in the Andean region as the ice melted included Lake Ballivan, which covered the area now occupied by Lake Titicaca, as well as others in Argentina and Paraguay.

II. Physical geography

RELIEF

An overview. South America has two major mountain systems of unequal mass. The young cordilleras of the Andes in the west extend down the entire continent from north to south, bordering the Pacific. On the eastern side are the ancient Guiana and Brazilian highlands, which are much less high and slope gently to the west; further south are the Patagonian plateaus. Lowlands—the Orinoco, the Amazon, and the Paraguay basins and the Pampas—divide the highlands from one another. The relief of the whole continent shows an imbalance; a drop of rain that falls only 100 miles east of the Pacific will flow down to the Atlantic, 2,500 miles away.

Physiographic provinces. *The Andes.* The Andes mountain ranges (*q.v.*), about 5,500 miles long and second only to the Himalayas in average height, constitute a formidable and continuous wall with many summits more than 22,000 feet high. Near the southwestern frontier of Colombia and through Ecuador they form a single mountain mass, from which three distinct massifs—the Cordilleras Oriental, Central, and Occidental—branch out. The valley of the Río Magdalena, between the Oriental and the Central ranges, and the valley of the Río Cauca, between the Central and the Occidental ranges, are in fact huge rifts. An overall view of the Andes in Colombia shows, within relatively short distances, a succession of hot lowlands and high ranges with a cold climate.

The Colombian Andes

In Peru, the Andes form two parallel cordilleras, one along the Pacific and the other descending eastward toward the valleys of the Ucayali and Urubamba rivers. The western range is the highest, particularly in central Peru where it splits into two chains; its impressive snow-clad peaks include Huascarán (22,204 feet or 6,768 metres). Between the two cordilleras extends the Altiplano—a vast complex of high plateaus between 12,000 and 15,000 feet high and sometimes as much as 125 miles wide. This Altiplano forms a maze of valleys, hills, and vast plains without equivalent except in Tibet. Water accumulates in closed basins to form marshes and lakes, one of which is Lake Titicaca. This huge massif has been dissected by rivers such as the Marañón, Huallaga, Apurimac, and Urubamba, all of which have cut spectacular gorges down the eastern slopes.

In the south, the Andes form a single but very complex chain with many peaks; one of which—Aconcagua—at 22,834 feet, or 6,960 metres, is the highest mountain in the Western Hemisphere. In Patagonia, a part of the cordillera runs beneath the sea, forming innumerable islands with steep slopes. The Andes have also been deeply carved by glaciers, which still occupy about 1,900 square miles, constituting an ice cap with long terminal tongues running into lakes or into the sea.

The Andes are studded with volcanoes that form part of the so-called "ring of fire"—the vast circle of volcanoes that border the Pacific. Earthquakes are frequent; almost every city has at one time been totally demolished, even along the coastal plains, where clear signs of recent vertical movement are visible.

The eastern plateaus. To the east, the Guiana Highlands and Brazilian plateaus consist of very old crystal-



Physiographic regions of South America.
Adapted from *Odyssey World Atlas*, © General Drafting Co. Inc.

line rocks, most of them greatly worn down. The Guiana Highlands are a monotonous expanse, about 800 feet high, forming a series of small hills separated by marshy depressions and surmounted by dome-shaped granitic inselbergs (steep-sided residual hills), some 2,000 feet high. The southern edge rises abruptly to a mountain chain, the Tumucumaque, and to some high tablelands, such as Mt. Roraima (9,094 feet, or 2,772 metres) and Mt. Duida (7,859 feet, or 2,396 metres).

The Brazilian Plateau

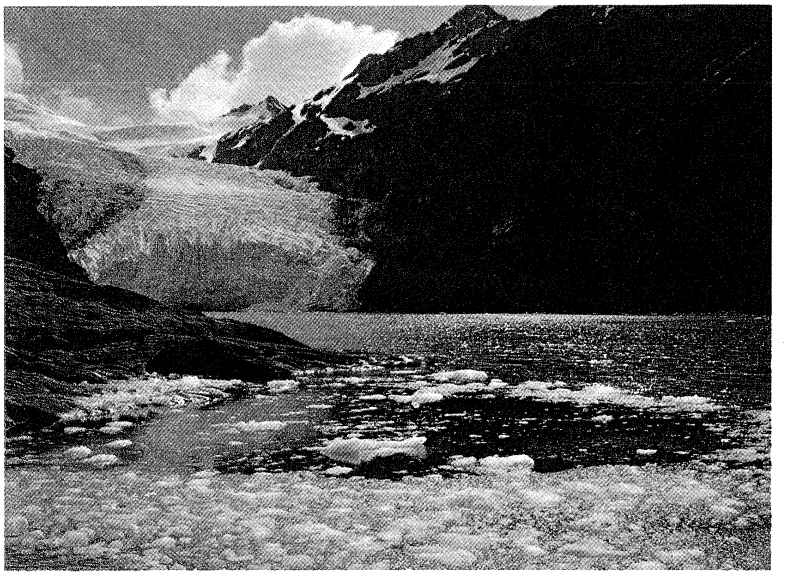
Covering an area of about 580,000 square miles, the Brazilian Plateau (also called the Brazilian Highlands) rises to about 3,000 feet, and is crowned by short ranges of hills; the São Francisco River has cut deeply into it. In the north the plateau slopes gently down to the sea, but in the east it drops abruptly down, sometimes as much as 2,600 feet. At its southern edge the Serra do Mar has summits of more than 9,000 feet—such as the Pico da Bandeira (9,482 feet, or 2,890 metres), the highest point of Brazil. The sea has partly invaded the original coastal ranges; here rocky masses, such as the Pão de Açucar, or Sugar Loaf (1,279 feet, or 390 metres), and the Pico do Corcovado (2,309 feet, or 704 metres), have been isolated by the sinking of surrounding land.

Eastern Patagonia constitutes a series of gigantic tablelands covered with rounded pebbles and crumbling sandstones. Volcanic eruptions have also spread layers of basaltic lava flows and have dotted the sedimentary plateaus with volcanic cones.

The lowlands. The Orinoco Basin, also called the Llanos, extends between the coastal ranges of the Venezuelan Andes and the Guiana Highlands and is covered with alluvia brought down by the Andean torrents.

The Amazonian depression, the greatest river basin in the world, forms an enormous drainage area of 2,722,000 square miles, bounded by the Andes to the west, the Guiana Highlands to the north, and the Brazilian plateaus to the south. The ancient platform of primary rock that underlies it is covered with layers of alluvial sand and clay, so that it forms an immense plain of low undulations, the general incline being extremely slight. Iquitos, in Peru, is at an altitude of only 384 feet (117 metres), while Manaus, in the heart of the basin, has an altitude of 144 feet (44 metres).

The basin of the Paraguay River, between the Bolivian highlands in the west and the Brazilian plateaus in the east, consists of a series of vast alluvial plains drained by a complex network of rivers interspersed with marshes. These plains, called the Pantanal, are only a few hundred feet above sea level and are subject to annual floods; they form an immense swamp during the rainy season.



Tidewater glacier at the end of a fjord in the Tierra del Fuego region of southern Chile.
Reflejo—EB Inc.

The Pampas, covering almost 300,000 square miles, consist of an immense accumulation of loose sediments brought down from the Andes. These deposits, 1,000 feet deep at Buenos Aires and much deeper at some other places, have completely buried the ancient features of the land. The landscape is monotonous and seems perfectly level, though in fact it rises imperceptibly toward the west—rising from an elevation of 66 feet near Buenos Aires to 1,640 feet at Mendoza. Some hills, such as the Córdoba and San Luis ranges, stand out from the plains like islands.

CLIMATE

As South America extends over a vast range of latitudes, it is hardly surprising that it has a great variety of climates. Although it is the only southern continent to reach such a high latitude, it nevertheless has its broadest extent in the equatorial zone, so that tropical conditions prevail over half of the continent.

Air circulation patterns

Factors influencing climate. The circulation of air over the continent is influenced by winds related to the great anticyclones (circulation of winds around central regions of high atmospheric pressure) of the Atlantic and Pacific oceans. Trade winds blow from the northeast in the Northern Hemisphere and from the southeast in the Southern Hemisphere. Another southern high-pressure region is located off the west coast, so that strong westerly winds blow toward the continent. Two low-pressure centres—one situated over the south polar region, the other extending from northeastern Brazil to the Gran Chaco region of Argentina—affect wind patterns inland.

Landforms largely influence specific circulation patterns. Above all, the Andes prevent winds from the South Pacific anticyclone from reaching the interior of the continent.

Ocean currents have a strong influence, particularly the so-called Peru (Humboldt) Current, which in effect is a series of upwellings along the steep Pacific coast, which result in cold water being carried to the latitude of southern Ecuador, where the flow turns westward. Average surface temperatures vary from about 58° to 64° F (about 14° to 18° C)—about 18° F (10° C) lower than is usual in such latitudes. The east coast is washed by the cold Falkland Current in the south and by the warm Brazil Current in the north.

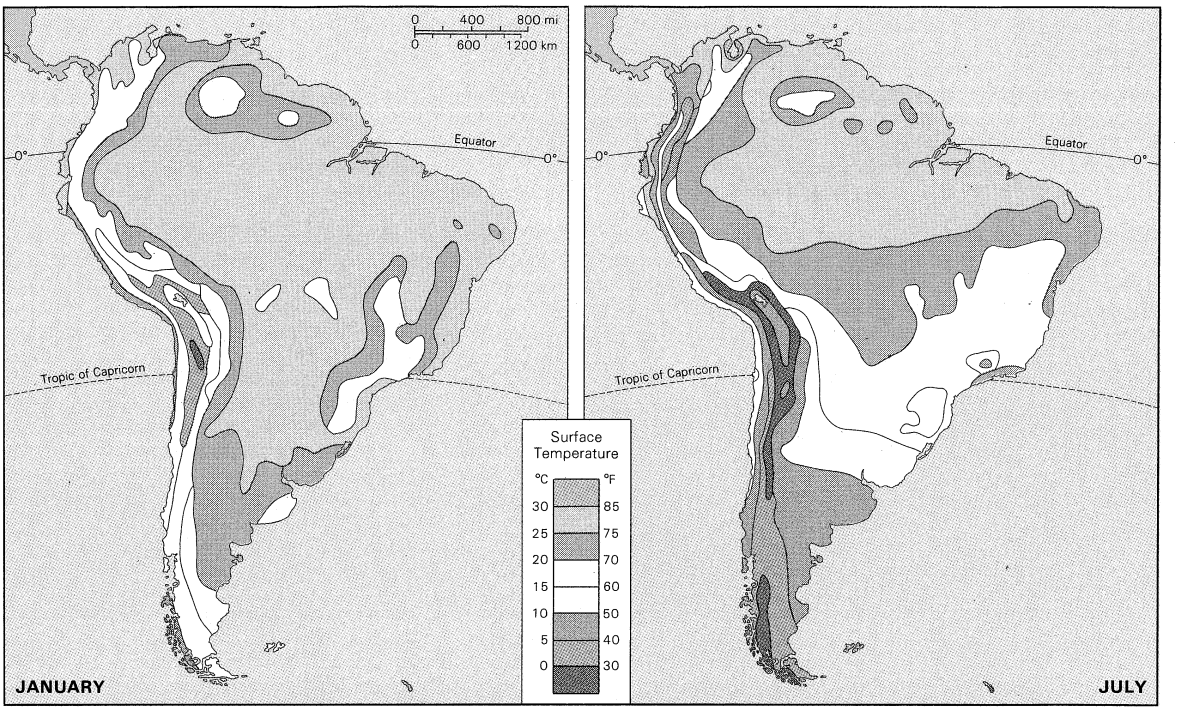
Finally, altitude plays an important part. Wide areas located within the tropical zone were uplifted by the rise of the Andes.

Climatic regions. South America can be divided into several well-marked climatic regions, as determined by the factors already mentioned. Four major types of climate occur—tropical, temperate, arid, and cold.

Among the tropical climates, the tropical rain forest type occurs on the Pacific coast of Colombia, in the Amazon Basin, and on part of the coast of Brazil. The average temperature is around 86° F (30° C), with monthly and annual variations of less than about 5° F (about 3° C); heavy rainfall, evenly distributed all the year round, averages about 98 inches in Pará, about 108 inches in Iquitos, 71 inches in Manaus, and 354 inches in the Chocó region of Colombia—one of the most watered zones in the world, where it rains more than 300 days a year. In the Amazon region, rains do not fall regularly all over the basin. The southern part receives rain especially during the southern summer (from October to April), while the northern part has its rainy season during the northern summer from May to September. The "dry" season is neither lengthy nor noticeable; humidity is always high.

The second type of tropical climate—the savanna (grassy parkland) type—is characterized by high temperatures (with monthly minimum temperatures above 64° F, or 18° C) but receives less precipitation and experiences a definite dry season. It is found around the rain-forest belt, in the Orinoco Basin, on the Brazilian Highlands, and in part of western Ecuador. Temperatures are still high and annual variations small, but average daily temperature extremes are greater, ranging from 42° to 63° F (6° to 17° C). There is a prolonged dry season.

Temperate climates, characterized by lower winter temperatures, are found south of the Tropic of Capricorn (in Paraguay, parts of Bolivia, Brazil, Argentina, and Chile) and in the Andes, where, in their rainfall pattern, they resemble tropical climates. On the Atlantic side, temperatures in the warmest month average 77° F (25° C), but cold-month averages vary from 63° F (17° C) in the north (Asunción, Paraguay) to 50° F (10° C) in Buenos Aires; rainfall is greater than one and a half inches each month in the east but decreases to the west. In central Chile, between latitudes 32° and 38° S, the features of the climate are similar to those of the Mediterranean, with mild winters and winter rains; summers are, however, cooler (69° F, or 21° C, in Santiago, Chile, in January—19° F, or 5° C, cooler than in the Mediterranean). In southern Chile, winter temperatures are lower, but not as low as the latitudes would seem to indicate. The islands and channels have a relatively uniform climate throughout the year, and winters are much less severe than in Labrador, which is at a comparable latitude north of the Equator. The presence of glaciers is the result of cold and cloudy summers during which ice does not melt. Rainfall is very high (102 inches in Valdivia, Chile, and probably twice this figure on the western slopes of the mountains), while the coast is one of the most heavily watered regions in South America. A short distance inland, rainfall de-



Average temperatures for January and July for South America.

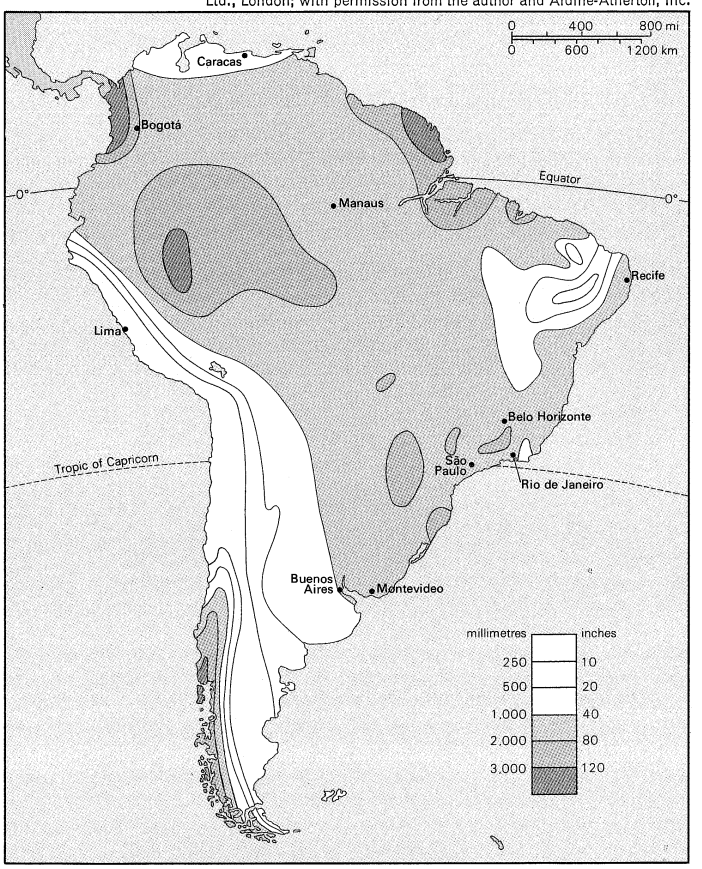
Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.

creases considerably (20 inches at Ushuaia, Argentina, which is protected from the westerly winds). Thus, in Patagonia there are surprisingly greater differences in climate from west to east than from north to south. Winds are very violent.

Arid climates are found in four areas. Patagonia east of the Andes is in the rain-shadow zone, and rainfall is low (about four inches in San Juan, Argentina). The annual range in temperature is more than 36° F (20° C)—the highest in South America—and is caused by warm summers and cold winters. Another arid zone is found in a narrow coastal strip along the Pacific coast between latitudes 5° and 31° S. The cold seas (the Peru, or Humboldt, Current) and the proximity of the high Andes produce an inversion of normal atmospheric temperatures, as air in contact with the water cools more rapidly than the upper strata of air. The result is a cloud layer about 1,200 feet thick, lying at altitudes varying between about 1,000 and 3,000 feet, that prevents the warming of the air near the ground. Temperatures are consequently low: Lima has an average annual temperature of 64° F (18° C), ranging from about 72° F (22° C) in February to about 59° F (15° C) in August. The coast of Peru is thus the cloudiest desert in the world, with no sunshine for at least six months of the year. It almost never rains, except under abnormal circumstances, but condensation of fog (called *garúa* by the Peruvians) provides some humidity. Another desert extends from northeastern Colombia to Venezuela, covering a zone where rains are scarce and droughts prolonged.

Arid zones

Adapted from Norton S. Ginsburg (ed.), *Aldine University Atlas* (1970), Aldine Publishing Co., Chicago; copyright © 1970 by George Philip & Son Ltd., London; with permission from the author and Aldine-Atherton, Inc.



Average annual precipitation in South America.

Finally, an arid zone occurs in northeastern Brazil, between the Parnaiba and São Francisco rivers. The interior highlands act as a wedge separating the sea winds from the northeast and those from the southeast, all of which carry their moisture beyond the region. Average annual rainfall is less than four inches, and the dry season may last as long as seven months. The worst feature of the climate is the irregularity of the rainfall, as a result of which severe droughts plague the region.

Cold climates, with average annual temperatures of less than 50° F (10° C), occur in the southernmost parts of Argentina and Chile and in the high Andes above about 11,500 feet. Mean temperatures are relatively low, but daily variations are wide. There is a marked difference in humidity between the northern and southern parts of the upper Andean zone. In Colombia and Ecuador the climate at such altitudes is cool and damp. Temperatures, always low, may on the average vary daily from 54° F (12° C) during the daytime to 28° F (—2° C) at night. Rainfall is high and well distributed throughout a good part of the year. Clouds and mist are dense, and sun shines only for short periods. From central Peru to Bolivia and Chile, temperatures are still lower. Near Lake

Titicaca, the average annual temperature is only about 34° or 36° F (1° or 2° C); November is the warmest month, with a temperature of 41° F (15° C), while the coldest month, July, has an average of 28° F (−2° C). Daily variations are considerable; the daily maximum of 68° F (20° C) drops to a nocturnal minimum of 5° F (−15° C). Annual rainfall varies from 24 to 56 inches but is concentrated during the southern summer, from November to April. The dry season is long and is characteristically accompanied by drought. Winds are constant and often violent, accelerating the coldness and the dryness of the climate, which produces a harsh and hostile environment. As in any mountainous region, climate varies largely according to local conditions.

DRAINAGE

River systems. Drainage has been deeply affected by the dissymmetry of the continent. The largest basins are found on the east side, and the main rivers run to the Atlantic Ocean.

The Amazon River

By far the largest river is the Amazon River (*q.v.*), which, with a length of 4,000 miles, is second only to the Nile; the volume of water carried, however, surpasses that of all other rivers, comprising a fifth of all the flowing water of the world. About 6,350,000 cubic feet of water per second is emptied into the Atlantic by the Amazon—ten times more than the Mississippi. The Amazon drains over three tenths of South America and has more than 1,000 tributaries, seven of which are more than 1,000 miles long. The source of the Amazon (although not the most distant of its headwaters from the sea) is in a small lake (Lauricocha) at an altitude of 15,875 feet (4,840 metres) in the Cordillera Huayhuash, near the Cordillera Blanca, Peru. It is named the Marañón in its upper course; after being joined by several rivers, it escapes from the Andes through narrow canyons. It is from the source of one of these rivers, the Ucayali, that the Amazon's length is measured. Below Iquitos, Peru, the river—often called Río Solimões—is joined by the Río Napo, which flows out of Ecuador; further on, near Manaus, it is joined by the mighty Río Negro, which drains a large part of northern Brazil. The Amazon, then at full strength, winds through the low plains to pass between the Guiana Highlands and Brazilian Highlands before emptying into the Atlantic.

As rains do not fall regularly year round all over the basin, the river has two annual floods in its upper course, which is subject to the alternate influence of the tributaries that descend from the Peruvian Andes (where rains fall from October to January) and the tributaries on its left bank, which descend from Ecuador (where rains fall from March to July). This pattern of alternation disappears further downstream, the two seasons of high flow gradually merging into a single one, which progresses slowly downstream in a gigantic wave from November to June. The height of the floods is lower further downstream. At Tefé, the difference between the high and low water mark is 50 feet and, at Obidos, Brazil, only 25 feet. The width of the river likewise varies; exceptional floods may extend more than 60 miles beyond the main river bed. Because of the low gradient of the basin, tides are perceptible as far as Obidos, 600 miles from the Atlantic coast.

The Paraguay, Paraná, and Uruguay system

The second important drainage system, variously estimated to cover from 1,600,000 to 1,700,000 square miles, is formed by the Paraguay, Paraná, and Uruguay rivers, which empty into a vast expanse of the water, the Río de la Plata (*q.v.*)—actually not a river but a gulf. About 2,800,000 cubic feet of water per second flow from the common mouth of these rivers. The Paraguay River (*q.v.*), with a length of 1,584 miles, rises in the Bolivian hills; it is a river of the plains, flowing across a wide stretch of marshes (the Pantanal) in its middle course; its lower course is, however, drier. The Paraná River (*q.v.*), with a length of 2,485 miles, flows mainly among high plateaus.

The Orinoco (*q.v.*), 1,337 miles long, is the third-largest drainage system, covering about 365,000 square miles. It flows west and then north; plunging down a series of steep slopes, it then flows along an almost flat basin. Near the ocean, it divides into a series of distributaries that form the Delta Amacuro, which has an area of about 5,000 square miles. The rainy season from April to October causes annual floods, which at Ciudad Bolívar, Venezuela, reach a height of 60 feet in July. By means of the Casiquiare Canal, it communicates with the Amazon system through the Río Negro.

The São Francisco River (*q.v.*), with a total length of 1,800 miles, rises in the State of Minas Gerais, Brazil, and flows northward for 1,000 miles before curving eastward to the Atlantic. The river has been an important artery of communication since colonial times and is now the location of large hydroelectric projects in operation since 1955, at Cachoeira da Paulo Afonso (Paulo Afonso Falls), and at Três Marias, north of Belo Horizonte.

Among other important rivers are the Río Magdalena in Colombia (navigable in two sections separated by rapids at Honda), and the Essequibo, Maroni, and Oyapock rivers in the Guianas.

Drainage to the Pacific is different because of the proximity of the Andes from the Pacific coast and the scarcity of rains from southern Ecuador to northern Chile. The rivers are consequently very short, and few convey any large quantity of water. The only important rivers are the Río Guayas in Ecuador, on which the port of Guayaquil is located, and the Río Santa in Peru. Some torrents had great importance since early times, when good water management permitted ancient civilizations to develop. In central Chile the valleys of Río Aconcagua and Río Bío-Bío are very fertile.

The Andean lakes

Lake systems. Most of the lakes are mountain lakes located in the Andes or along their foothills. Because of the complex topography, water has accumulated in closed basins. Among permanent lakes, the largest is Lake Titicaca (*q.v.*), which lies at an altitude of 12,500 feet (3,810 metres) between Peru and Bolivia; the lake measures 110 by 35 miles and was much wider in the past. Lago de Junín in central Peru; Lake Sarococha, also in Peru, between Puno and Arequipa; and Lago de Poopó in Bolivia also rank among the larger Andean lakes. These create uniform physical conditions throughout the year in terms of temperature and percentage of dissolved gas; up to an altitude of 16,000 feet they never freeze. In consequence, the climate of their shores is temperate.

Another type of lake is to be found in Patagonia where, in the wake of melting glaciers, lakes have formed, such as Lago Buenos Aires, Lago Argentino, and Lago Nahuel Huapí. Their eastern parts, which stretch to the end of the Argentinian plateau, generally have gently sloping banks bordered by low mountains, while their western parts form a series of narrow arms, like fjords, lying between steep slopes.

Marshes and swamps. Marshes are to be found in depressions in many parts of the continent. One of the widest marshy areas is the Pantanal, in the middle valley of the Paraguay River; it is subject to flooding in December, reaching its highest watermark in June, when it becomes an immense swamp; similar habitats are to be found in the Orinoco Basin.

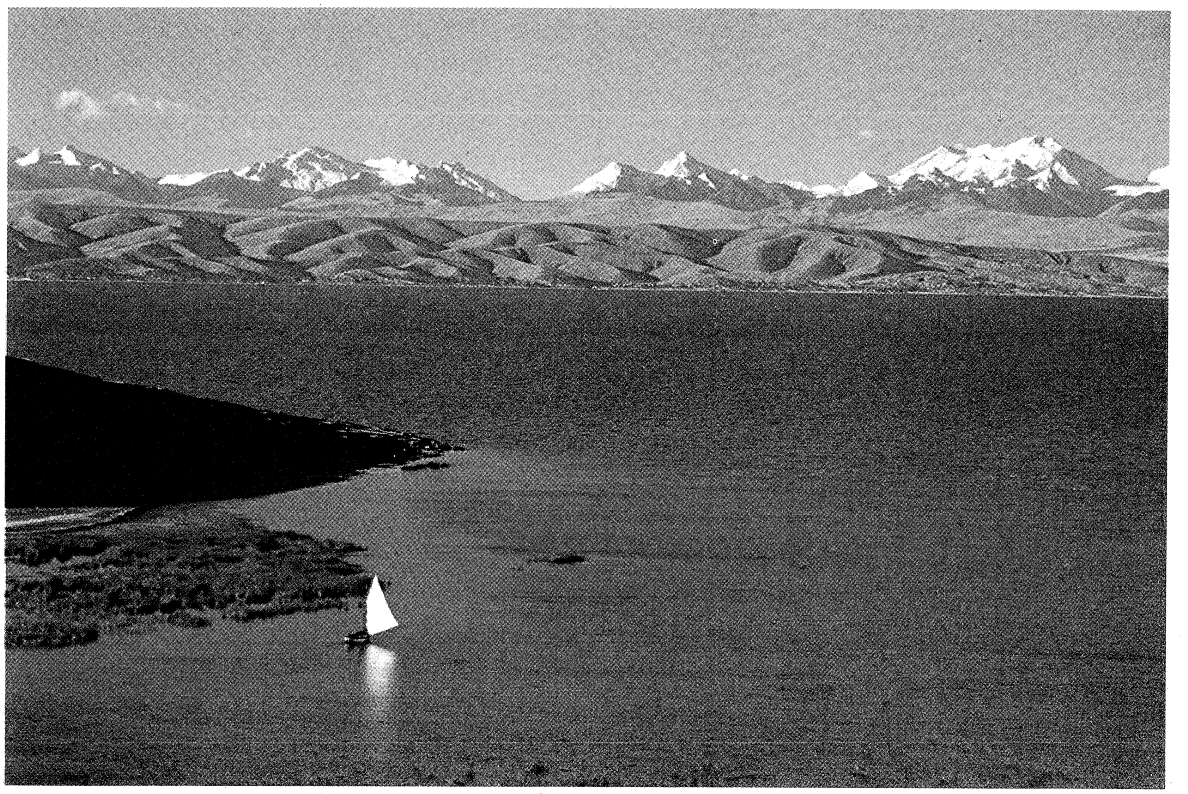
Swamps of another type are found in rain forests, mostly in the Amazon region and in northwestern Colombia. The ground is inundated or very marshy throughout the year or else only at the time of the annual flood.

Finally, the mouths of the Orinoco and of the Amazon rivers form wide marshy deltas, while mangrove swamps of various types are to be found from southern Ecuador to Santos in Brazil.

SOILS

Soil regions. As a consequence of its geological history, geomorphology, and climate, no less than 26 soil regions, many of which are split into several subdivisions, can be distinguished in South America. They consist of three major groupings, which correspond to the continent's three major geographical components—the lowlands, the uplands, and the Andes.

Low natural fertility is an outstanding feature of South American soils. About 50 percent of the continent's soils



Lake Titicaca, the world's highest large lake, located between Peru and Bolivia. The Cordillera Real of the Andes rise along the northeastern shore of the lake.
Jacques Jangoux

are made up either of unconsolidated and nutrient-poor sediments (such as china clays and quartz sands) in the river basins, latosols (red residual soils, also called laterite, leached of silica and containing concentrations of iron and aluminum hydroxides), red-yellow podzols (highly bleached soils that are low in iron and lime), and regosols (azonal soils consisting chiefly of imperfectly consolidated material and having no specific morphology). About 20 percent of the continent is covered by desert soils of various types in which agriculture is hazardous because of lack of water. Other regions, representing about 10 percent of the total area, suffer from poor drainage, the soils being gleys (layers of soil in which the material is blueish grey, generally sticky, and often structureless, due to excessive moisture), groundwater laterites, grumosols (soils formed of granules), and planosols (a type of soils found in humid climates in which soluble salts and minerals are leached out of the upper layers and are cemented or compacted at a lower level). In the Andes, slopes are often steep, and lithosols (soils with no clear morphology, consisting of imperfectly weathered masses of rock fragments) abound, accounting for about 10 percent of the continent's surface. In the inter-Andean valleys and on some of the foothills, nevertheless, eutrophic soils (deposited by lakes, and containing much nutrient matter, but often shallow and subject to seasonal oxygen deficiency) can be found.

The fertile soil regions

Fertile soils extend over only about 10 percent of the surface of South America. The most important of these are brunizems (any of a zonal group of deep dark prairie soils, developed from loess; *i.e.*, unstratified loam deposited by the wind), chestnut soils, ferruginous tropical soils, roxo latosols, and reddish-brown lateritic soils. On the low coastal ranges and in the foothills of the western Andes and on the nearby plains and terraces of Colombia and Ecuador, the soils consist mainly of red-yellow latosols, podzols, and alluvial soils. Soils in southern Brazil and Uruguay are not uniform and consist of brunizems, reddish prairie soils, and planosols. The Argentinian Pampas, the largest fertile area on the continent, is uniformly covered with the so-called pampean loess, which is calcareous, rich in minerals, and partly of volcanic origin. Other rich soils are found in the uplands of northeastern and central Brazil; they consist of sandy regosols in the north and roxo latosols in the south.

The agricultural development of South America reflects exactly the distribution of soils according to their fertility. It is mostly confined to the eastern temperate plains, in which is concentrated the production of wheat and corn and the grazing of cattle; to the subtropical and temperate parts of the Andes, from Colombia to Chile, where grazing takes place and a variety of crops are cultivated; and to eastern and southeastern Brazil, where coffee, cacao, and sugarcane are grown, while the interior plateaus are devoted to cattle grazing. The Amazon Basin is still a region of elementary agriculture, though it produces cacao and rubber.

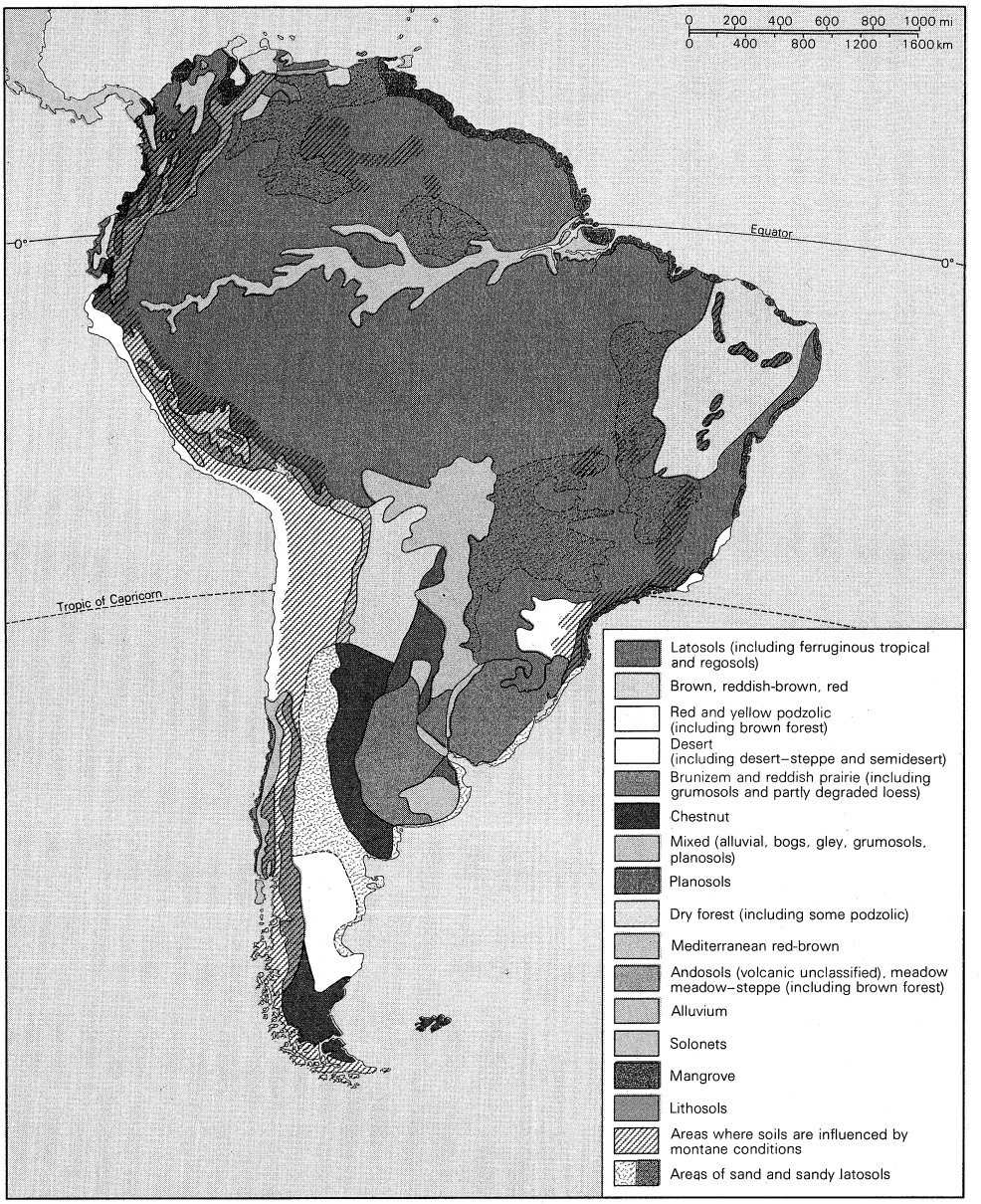
Soil problems. Soil erosion has ravaged a large part of the continent. According to some estimates, in several countries up to 50 percent or more of the presently arable land has been severely affected or ruined by bad land use. In the Andes, land that once produced up to 25 bushels of wheat to the acre is now abandoned; mountain forests are still cleared for cattle grazing and cultivation, which ruins the soil of the region for years after. Soil damage has been slight in the areas in which the topography is relatively smooth. Campaigns for soil conservation or restoration are now in progress in most countries.

III. Vegetation and animal life

South America, with its distinct vegetation and animal life, is the very heart of the neotropical region (that part of the Americas that extends from the Tropic of Cancer southwards). With Australia it is, because of its isolation from the rest of the world during the Tertiary period, the landmass with the strongest biological originality.

Vegetation and animal life are a mixture of elements of various relationship. Ancient groups of plants and animals including mollusks, chilopods, some fishes, reptiles, and amphibians show affinities with the animal life of Africa, Australia, and New Zealand. More recent elements migrated from or through North America during the late Tertiary Period and the Pleistocene Epoch; most of these are vertebrates.

The pattern of distribution within the continent is much diversified because of the variety of climatic and ecologi-



Soils of South America.
Adapted from R. Ganssen and F. Hadrich (eds.), *Atlas zur Bodenkunde*

cal zones. The northern tropical regions are the richest, while the southern part and the Andean highlands are much impoverished, despite some differentiation.

VEGETATION

The origin of South American vegetation is much diversified, as is shown by the relationship of its components. The proportion of endemic plants is very high, even at the family level. Among angiosperms (plants having seeds enclosed in an ovary) no less than 25 families and 3,500 genera are endemic in the neotropical region. Others are related to African plants or belong to southern plant groups also distributed in Southern Africa and in Australasia.

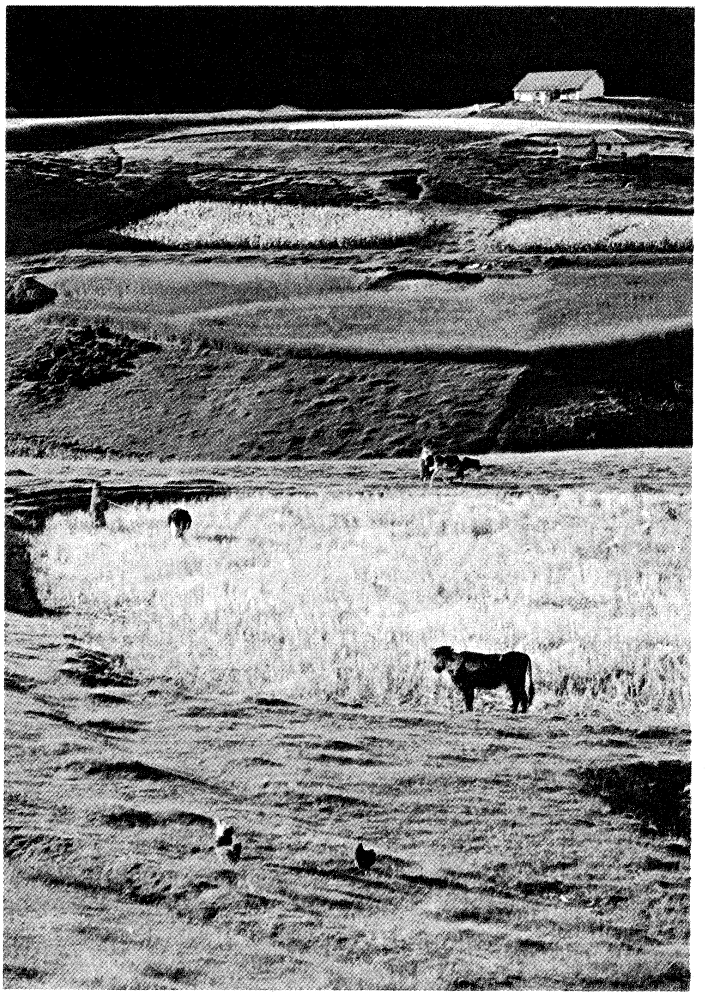
Vegetation is by no means uniform all over the continent; its distribution is determined by climatic, geographic, and soil differences. The main geographic plant zones are as follows:

Tropical and subtropical rain forests. Rain forest covers the largest part of the Amazon region, most of the Guianas, southern and eastern Venezuela, the Atlantic slopes of the Brazilian Plateau, and the Pacific coast of Colombia and northern Ecuador. The huge Amazon region is the largest and probably the oldest forest area in the world; it also ascends the slopes of the Andes until it merges with subtropical and temperate rain forest. On its southern border it merges with the woodlands of Mato Grosso, with galleries of its trees extending along the rivers. Consisting of enormous trees, some exceeding a height of 300 feet, it is composed of an almost incredible number of species growing side by side in the greatest confusion and arranged in different strata. In the region of Manaus, 1,652 plants belonging to 107 species in 37 different families were found on about 1,900 square feet. Amazonian trees number about 2,500 species.

The Amazon rain forest

The forests of the *igapós* (swamps, where the ground is inundated or very marshy throughout the year) cover the lowlands. Characteristic trees are, among others, jacareúbas (*Calophyllum brasiliense*), which is a tall tree with hard reddish-brown wood used for heavy construction, araparis (*Macrolobium acaciaefolium*), abiuranas (*Lucuma* species); piranheiras (*Piranhea trifoliata*) and louros-do-igapo (*Nectandra amazonum*). Undergrowth is dense.

The *várzea* regions are those that are inundated only at the annual flood. Trees are higher and quite diversified; they include oeiranas (*Alchornea castaneifolia*), a euphorbia (having a milky juice) of the chestnut family, and imbaúbas (*Cecropia* species), a rapid-growing tree of the nettle family with a light wood. Palms and hevea (a kind



Farming in the Andean valleys near Ibarra, northern Ecuador.
Reflejo—EB Inc.

of rubber plant) grow in these forests. The forests of the nonswampy areas are rich in hardwoods, of which acapu (a tree with dark-chocolate-brown wood), pau-amarelo (*Euxylophora paraensis*), pau-santo (*Zollernia paraensis*), massaranduba (a Brazilian tree with light-reddish-brown wood that yields a milky juice), jarana (a tree with hard, heavy, durable wood), and matamatá (a tree with hard, heavy, compact wood, used for pilings) are the best known. Caucho (a wild rubber tree) and the castanheiro (*Bertholletia excelsa*) are characteristic of these forests where spiny palms cover the ground.

Epiphytes (nonparasitic plants that grow on other plants, deriving moisture and nutrients from rain and air) are very numerous, mostly Bromeliaceae (a species having leaves that are often spiny, and spiny flowers), orchids, and ferns. Lianas abound, particularly in drier forests.

Tropical deciduous forests. These forests, dominated by trees of moderate height, notably of leguminous species, are found widely dispersed through northern South America, where the climate is characterized by a prolonged dry season, notably in Venezuela, Colombia, and in the Brazilian Plateau.

Caatinga. Caatinga (white forest) is stunted and somewhat sparse. Thorny vegetation and trees that are leafless for long periods, and able to resist drought, are characteristic of the dry interior of northeastern Brazil, particularly in the basin of the São Francisco River. Dominant species are leguminous trees, particularly catingueiras (*Caesalpinia*), juremas (*Mimosa*), and joazeiros (*Zizyphus joaseiro*), Euphorbiaceae (a family of herbs, shrubs, or trees with usually milky and often poisonous juice), crotons (herbs and shrubs of the spurge family), and Bombacaceae (a family of tropical trees with palmate leaves and large, dry, or fleshy fruit). Undergrowth consists of thickets, bromeliads (plants with basal, often spiny leaves), and innumerable cacti, among which is the xiquexique (*Cereus gounellei*), the complicated intertwinings of which cover the soil.

South Brazilian forests. These parklike forests, sometimes very dense but interspersed with savanna, occupy vast expanses from the border of the Amazonian rain forest to the marshes of the Upper Paraguay River. The typical landscape is a grassland strewn with smaller trees. In effect, it includes a mosaic of associations, from hygrophilous (living or growing in moist places) to xerophilous (adapted to life in the presence of minimal amounts of water) forests and even desert.

Notable is the Paraná pine forest (*Araucaria angustifolia*), which covers a vast area between the Río Paraná and the Atlantic Ocean, stretching from Curitiba, Brazil, to northern Argentina. These trees dominate a dense forest of numerous species including hardwoods, Podocarpus, and the yerba maté (*Ilex paraguariensis*), or South American holly.

Xerophytic associations. Thickets of small trees and shrubs, often very thorny, among which prosopis (a genus of tropical or subtropical branching shrubs or trees), acacia, and mimosa predominate, cover regions that experience alternate dry and relatively wet seasons; these regions particularly include coastal Venezuela, northeastern Colombia, southwestern Ecuador, and northern Peru.

In Peru this association merges with the desert, which extends along the coast to northern Chile, having a width of from 50 to 100 miles. Only a few shrubs and some terrestrial (as distinct from epiphytic) bromeliads grow in this area, which becomes greener only in the Andean foothills, where cacti and other xerophytic plants are to be found, and along the valleys of rivers flowing down from the Andes. In some areas, called *lomas* by the Peruvians, winter mists bring some humidity, and a specialized type of vegetation, consisting of annual and bulbous plants, grows for a short period.

The temperate rain forests

Sub-Antarctic beech forests. Temperate rain forests—similar to those found in British Columbia and in the northwestern United States—grow in southern Chile at low and moderate altitudes, thanks to very high rainfall. The most typical trees belong to the genus *Nothofagus* (timber trees found in the cooler parts of the Southern Hemisphere), the northern species of which are evergreen, whereas the southern species are deciduous. Various conifers, notably the larch and araucarias, mingle with the leafy trees. A dense undergrowth of shrubs, lianas, bamboos, ferns, mosses, and epiphytes grow in the northern districts but disappear toward latitude 49° S. In southern Chilean Patagonia, forests consist of twisted, creeping trees merging into a kind of heath.

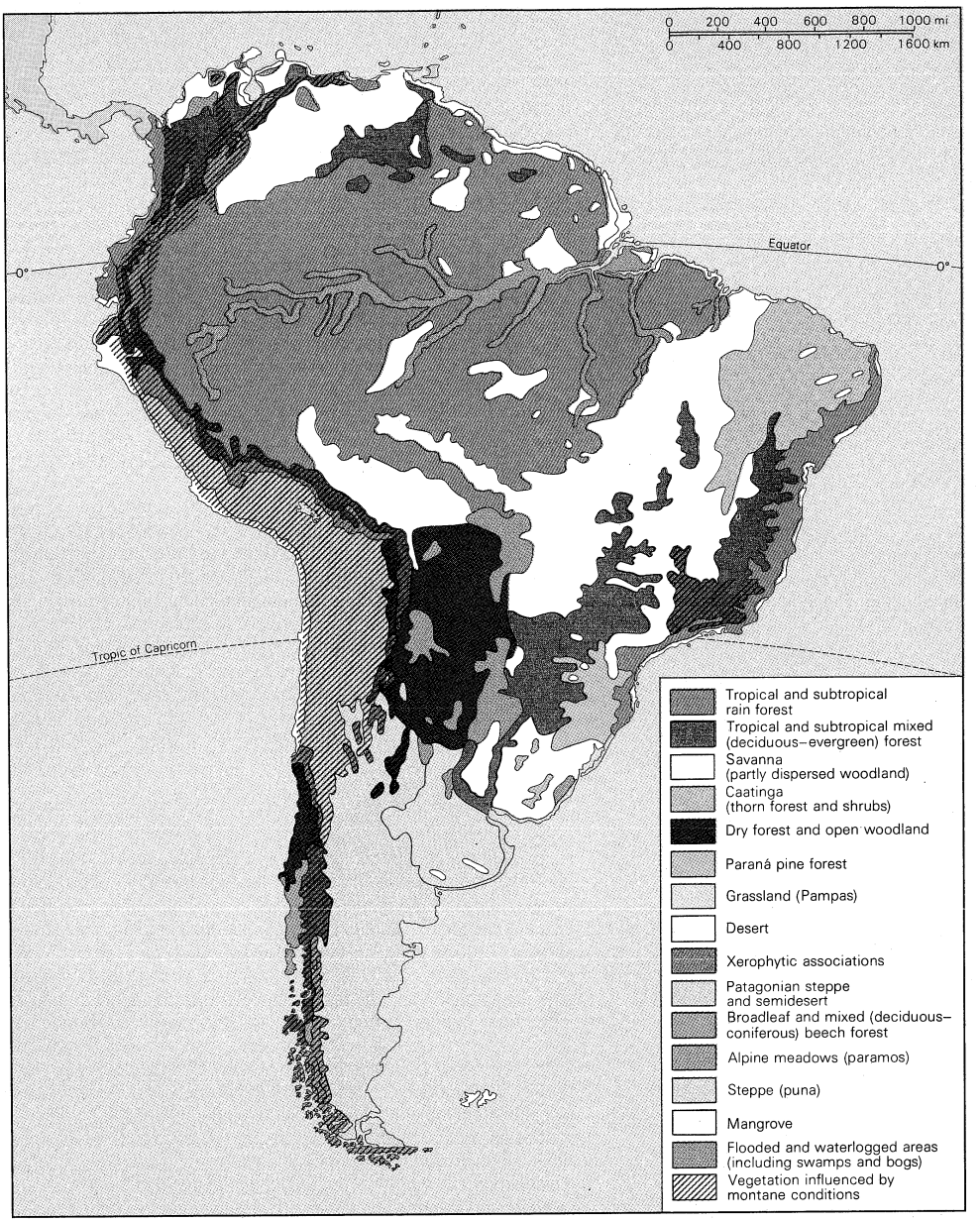
Savannas. True savannas are found mostly in Venezuela where the plains known as the Llanos (*q.v.*) of the Orinoco Basin are covered with tufts of grasses and sedges. Trees and some palm trees are fairly abundant, especially along the rivers.

Pampas. The flat plains called Pampas, which constitute the greater part of eastern Argentina, are covered with grasses. The Pampas were originally covered with trees, but man has acclimatized various species, including pines, eucalyptus, oaks, and poplars.

To the south the Pampas merge with the Patagonian steppe, where grasses are mixed with rosette and cushion plants; the vegetation becomes much poorer as one proceeds south.

Mountain vegetation. In the high Andes, a temperate zone extends from the upper limit of the subtropical rain forest, around 6,000 feet, up to the timberline, which occurs at an altitude of between 10,500 and 12,500 feet. This zone is very humid on the Amazonian side, where both the epiphytes and the undergrowth are dense.

The upper zones have a peculiar vegetation that reaches the snow line. In the wet northern Andes in Colombia and Ecuador, Alpine meadows called *páramos* consist of grasses and of herbaceous plants, often with bright flowers; these are surmounted by taller plants, especially good-sized plants called *Espeletia,* or frailejones, which



Vegetation zones of South America.

are 15 to 20 feet high and are crowned by enormous bouquets of long, lanceolate, hairy leaves.

In the south the *páramos* merge with the *pajonal de puna*—a typical steppe vegetation consisting of rough grasses, between which grow a variety of herbaceous, cushion, and rosette plants, shrubs, and cacti. Vegetation extends to the limit of permanent snow but becomes very scarce at high altitudes where the soil is often barren.

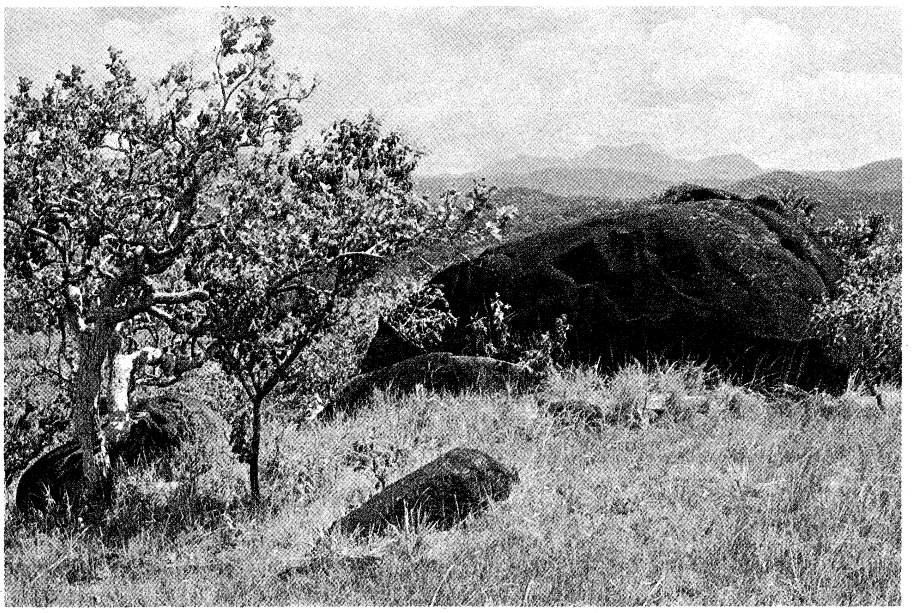
Destruction of the forest cover

Man has transformed the original vegetation cover to a large extent throughout South America, particularly in forested areas. The forests of eastern Brazil were ravaged in the process of clearing the ground for sugarcane plantations as well as to grow other crops. The forests of *Araucaria* (a genus of tall trees of the pine family) in the southern states of Brazil are rapidly vanishing. The slopes of the Andes are also severely deforested. New roads have been cut through the Amazon region, which is no longer the enormous inviolate mass it once was. In Patagonia the practice of burning to convert forest into pasture increases from year to year. Man has also destroyed grasslands through overgrazing, notably in the Venezuelan Llanos and in the high Andes. The process of the destruction of habitats is accelerating alarmingly throughout the continent, where previously primeval conditions had remained stable for centuries.

ANIMAL LIFE

South American animal life is particularly rich and well diversified as a result of the wide range of habitats. A number of animal groups well distributed over the rest of the world are nevertheless missing, because of South America's lengthy isolation. Many animals belong to exclusive groups, and, even at the family level the percentage of endemic forms is very high. Speciation has reached a much higher degree when compared to other parts of the world—a phenomenon that makes attempts to establish systematic classification difficult in South America. Neotropical animal life nevertheless evidently consists of several well-defined elements.

Fish and bird life. Freshwater fishes are very numerous, numbering about 2,700 species, though deriving from only a few ancestral groups. Amazonian fishes may approach 1,500 species in number. Characins (800 species), which are tropical fishes with deep scaly bodies and strong jaws; gymnotidae, South American syprinoid fishes that include the electric eel; catfishes; cyprinodonts, a large family of small scaly-headed soft-finned fishes; and cichlids (a family chiefly consisting of fishes resembling sunfish) are the dominant groups. Amphibians, well represented by caecilians (small wormlike, burrowing amphibians), salamanders, toads, and various frogs, in-



Savanna vegetation in the Guiana Highlands of eastern Venezuela.
Jacques Jangoux

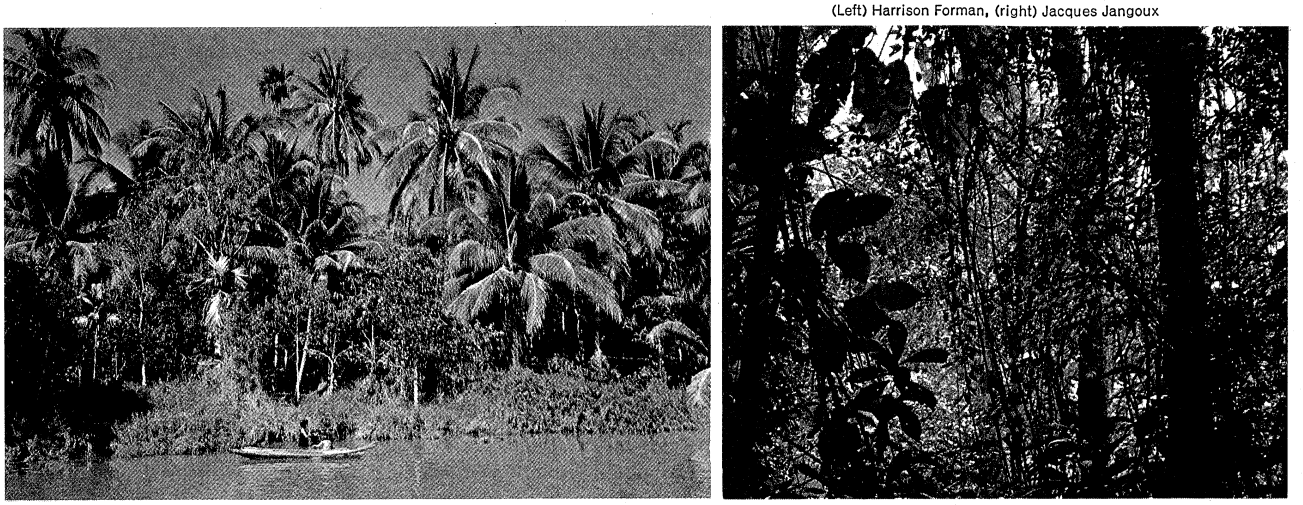
cluding clawed frogs, the most aquatic of all; the tree frogs, arboreal amphibians, are particularly abundant through the Amazon Basin, and are very different from their African and Asian counterparts, although the frog faunas of Australia and South America are often surprisingly alike. Reptiles include a great variety of turtles and tortoises, crocodiles, caimans, (endemic crocodilians), geckos, many iguanas, teiids (a family of mostly tropical American lizards), amphisbaenids (a genus of harmless, limbless lizards) and many snakes, including boas, colubrids (a very large family of nonvenomous snakes), coral snakes, and vipers. Birds are represented by 89 families and about 2,700 species—a much higher figure than in Africa or Asia, which justifies the application of the name bird continent to South America. Twenty-five families are endemic to the neotropical region, including rheas (large, tall, flightless birds), curassows (large arboreal birds distantly related to the domestic fowl), hoatzins (a brownish crested bird, having claws on the digits of the wing when young), oil-birds, motmots (bright-coloured birds related to kingfishers), jacamars (small, bright-coloured, clamatorial birds), toucans, manakins, cotingas (related to manakins), and many passerine birds. Hummingbirds are differentiated in the neotropical region, where parrots, pigeons, cuckoos, tyrants (a kind of flycatcher), woodhewers, and orioles are among the dominant groups. Remarkably, the nonpasserine birds are more numerous in proportion to passerine birds than in other parts of the world.

Mammals. Mammals include types that immigrated to the continent before its complete isolation in the early Tertiary Period. Among these are marsupials (pouched animals), sloths, anteaters, and armadillos; other kinds, now extinct, persisted until the Pleistocene Epoch. More recent elements include hystricomorph rodents (a suborder including porcupines and guinea pigs) and monkeys among the oldest arrivals, and tapirs, deer, carnivores, and bats among the newer arrivals.

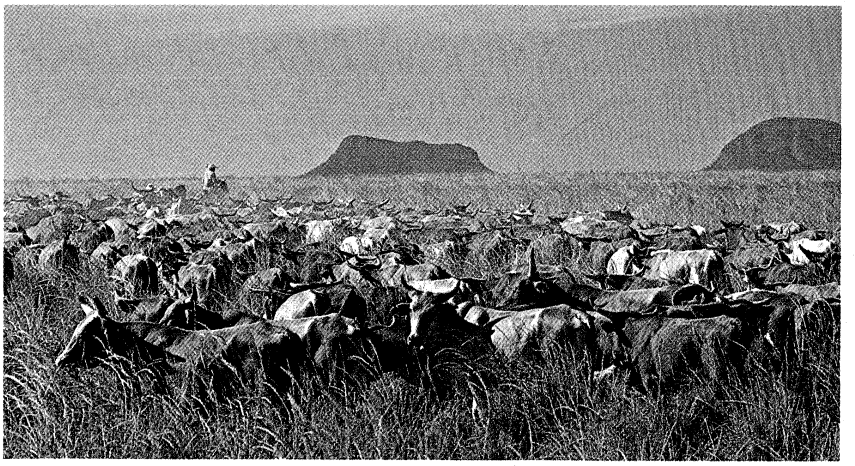
Ecological communities. Animals are distributed according to the pattern of vegetation zones and several well-defined communities can be distinguished.

The Amazonian and Guianan forests. The most diverse community is found in the Amazonian and Guianan forests where abundance of water and trees makes life easy. Rivers are the realm of large numbers of invertebrates and fishes, such as pacu (*Metynnis*), which is a big brownish flat fish, the meat of which is highly appreciated, coumarou (*Curimato*), which is a vegetarian fish without teeth resembling the marine mullet; electric eel (*Electrophorus electricus*); arapaimas, or pirarucu, which attain a length of 15 feet and a weight of 200 pounds; and piranhas provided with teeth so sharp that they can cut

(Left) Harrison Forman, (right) Jacques Jangoux



Amazon and Guianan rain forests.
(Left) Traffic on the Amazon River in Brazil. (Right) Dense tropical vegetation in the southern Guiana Highlands region of Venezuela.



Range cattle grazing in shoulder-high grasses that grow in the Pampas of eastern Argentina.
E. Hartmann—Magnum

pieces of flesh like a razor, as well as a wealth of small fishes, many of which are vividly coloured. The manatees (a chiefly tropical, aquatic, herbivorous animal with a broad tail) and the inia, a primitive dolphin-like cetacea (an order of completely aquatic mammals), frequent the large rivers. Arraus (*Podocnemis expansa*), a water turtle, is still extraordinarily numerous. Impressive populations of crocodiles and caimans inhabit the main waterways.

Arboreal forms of the Amazonian forests

Amazonian forests constitute an environment to which most animals responded by becoming arboreal. Tree frogs can move across the surface of the leaves thanks to adhesive pads; lizards have very elongated fingers; monkeys, sarigues (a close relative of the opossum), potos (carnivorous mammals, arboreal and nocturnal, with rounded heads), and kinkajous (a kind of nocturnal, arboreal, carnivorous mammal) have prehensile tails. Birds are numerous and, because of the enormous amount and variety of available foods, well diversified. Antbirds, tyrants, cotingas, tangaras (brilliantly coloured birds related to the finches), hummingbirds, toucans, woodpeckers, barbets (loud-voiced tropical birds closely related to honey guides), parrots, and tinamous (quail-like terrestrial birds) are the dominant groups. Many of them never leave the forest canopy where they display their brilliant colours, which contrast with the modest plumage of those birds living in the undergrowth. Mammals are represented by a number of terrestrial or half-aquatic species, such as very small deer; some large rodents (including the agouti, paca, and capybara); tapirs; and carnivores (including the jaguar). Monkeys range from pygmy marmosets to larger douroucoulis (small, round-headed, stocky-bodied, bushy-tailed monkeys), woolly monkeys, spider monkeys, and howler monkeys. Bats are numerous, including fruit bats. Sloths feed on the leaves of certain trees. Predators are represented by carnivorous mammals, a series of snakes—including anacondas and boas—and big raptors (birds of prey), such as the harpy eagle, the most powerful bird of prey in the world. Insects, including a great variety of butterflies and ants, are innumerable.

The Brazilian Plateaus. The Brazilian Plateaus have an impoverished animal life, from which species strictly adapted to the dense forest are excluded. The plains of Uruguay and the Gran Chaco have a varied animal life that includes some particular species, such as the maned wolf. The marshes are inhabited by a wealth of waterfowl, as well as by a lungfish (*Lepidosiren paradoxa*) related to its African and Australian counterparts.

Animal life of the Pampas

The Argentinian Pampas. The Pampas of Argentina are inhabited by a very limited number of animals such as rheas, a series of smaller birds, including the popular ovenbird (*Furnarius rufus*), the name of which comes from its globe-shaped nest made of mud, and mammals such as the mara (*Dolichotis patagona*), a long-legged, long-eared rodent; the plains vizcachas (*Lagostomus*), a burrowing rodent related to the chinchilla; guanacos (*Lama guanacoe*), a South American mammal related to the camel but resembling a deer; and Pampas deer (*Blastoceros campestris*). The restricted number of larger herbivorous mammals is remarkable and illustrates the scarcity of recent mammalian types in the neotropical region.

The southern Chilean forests. The forests of southern Chile are inhabited by a specialized animal life, with a high percentage of endemic species. Parakeets and hummingbirds are found as far south as Tierra del Fuego. A marsupial, the rincolesta of Chiloé (*Rhyncholestes raphanurus*), is one of the most primitive mammals still in existence.

The high Andes. The high Andes have been settled by a very impoverished animal life, which had to adapt to the harsh and cold environment, to the scanty vegetation, and to low oxygen pressure. The wealth of lakes has attracted numerous aquatic birds, including flamingoes, which nest up to 16,000 feet in northern Chile, and amphibians such as the giant toads of Lake Titicaca, which spend their entire life in water. Mammals are represented by guanacos, vicunas (wild ruminants related to the domesticated llama), deer, and numerous rodents, including vizcachas, chinchillas, and guinea pigs. Predators number foxes, pumas, and many birds of prey, including the condor, the giant of living birds, with a wingspan of more than ten feet.

The arid west coast. The arid coast of Peru and northern Chile is inhabited by a terrestrial animal life impoverished when compared to its rich offshore marine life. The zones washed by the Peru Current, cold and rich in salts, are swarming with life, from plankton to fishes, including an anchovy (*Engraulis ringens*); these small forms of life provide food for higher levels of the marine community, represented, for example, by sea lions and birds, many of which are endemic to the area. Bird life includes a penguin, many species of gulls and terns, shearwaters, petrels, cormorants, pelicans, and boobies (a kind of gannet). Three kinds of these birds—the guanay (*Phalacrocorax bougainvillii*), the variegated booby (*Sula variegata*), and the brown pelican (*Pelecanus occidentalis*)—nest by millions on small islands off the coast, where their droppings accumulate to form guano, a highly prized fertilizer.

Protection for animal life. Overhunting and the destruction of habitats have wrought great harm to South American animals. Until recent times, South America had suffered less than the rest of the world. But since the mid-20th century the situation has changed very rapidly; many kinds of animals are on the decline; some are even threatened with extinction. Destruction of the Amazonian forest, which in the 1970s had just begun, could be disastrous for its animal life.

To remedy the situation, a number of reserves and national parks have been established through the continent.

National parks

Argentina was indisputably the pioneer in wildlife protection in South America. Nahuel Huapi National Park, the original nucleus of which was established in 1903, preserves a part of the Andes, featuring glaciers, beech forests, and many animals. Two national parks, one on the Argentinian, the other on the Brazilian side, protect the Iguaçu Falls (*q.v.*) and the surrounding rain forest. Some 17 parks and reserves, including the Itatiaia National Park, are maintained in Brazil. In Venezuela the Henri Pittier or Rancho Grande National Park, in the coastal cordillera, is famous for its subtropical forest, bird life (530 species), and migrations of birds and butterflies. Other reserves and parks have been recently established in Ecuador, Peru, and Chile in order to preserve representative areas of the habitats of the Andean slopes and plateaus.

IV. Natural resources

MINERAL RESOURCES

South America has a great variety of mineral resources. Though Peru, Bolivia, and Chile enjoy a long mining tradition, until the early 1970s the reserves of the continent were only very roughly estimated, particularly where minerals exist in widely scattered beds or under dense vegetal cover. Geological investigations are still limited to some regions, although new estimates can be made as knowledge advances.

Mineral fuels. South America is relatively poor in coal, which is found in Colombia (12,500,000,000 tons of estimated reserves), and in lesser quantities in Brazil, Argentina, Peru, Venezuela, and Chile (where lignite and peat are also available). Petroleum is present in great quantities in sedimentary terrains around Lake Maracaibo in Venezuela, in the coastal plains of Ecuador and Peru, along the eastern Andes of Bolivia, and in southern Patagonia. The most important reserves are located in Venezuela, the fifth-largest oil producer in the world, with nearly 180,000,000 tons of reserves, representing over 54 percent of the reserves of the continent, and 2 percent of the world's total. Other reserves are found in Argentina, Colombia, Peru, Brazil, Chile, and Bolivia. Natural gas is found in Argentina (estimated reserves are 7,275,000,000,000 cubic feet) and Chile. Uranium is found in significant quantities in the form of various ores in Argentina, Brazil, and Chile, as well as in Colombia.

Oil reserves

Shostal



Oil derricks on Lake Maracaibo, Venezuela, centre of the petroleum industry in Latin America.

Metallic deposits. Iron ore is very abundant; South American reserves represent 12 percent of the world total. The most important beds are found in Brazil and Venezuela, representing a fabulous wealth for these countries. In Brazil, the states of Minas Gerais and Mato Grosso contain lodes of magnetite and hematite with 50 to 65 percent of iron, estimated to contain 13,000,000,000,000 tons of ore, representing approximately 83 percent of the Latin American reserves. In Venezuela, at the base of the Guiana Highlands, the Sierra de Imataca, particularly Cerro Bolívar and El Pao mountains, have reserves of ore containing 65 percent of iron, estimated at over 1,600,000,000 tons. Other beds of this type are to be found in Mutún, Bolivia and Chile. Oolitic iron ore (*i.e.*, consisting of small round grains cemented together) is located in Argentina (Sierra Grande), and in Colombia (Paz del Río). Other iron ore deposits are found in Peru (Marcona) and Chile (along a belt from Taltal, Antofagasta, to Ovalle, Coquimbo). Lateritic deposits of ferrous hydroxides are widespread, mainly in Colombia, Brazil, and Argentina.

Among ferroalloys, manganese is found in sedimentary forms in Brazil—mostly in Amapá, Bahia, Minas Gerais, and Mato Grosso—where reserves are estimated at 60,000,000 tons of ore containing 38 to 50 percent of manganese; in Bolivia (estimated reserves 20,000,000 tons); and in much lesser quantities in Chile, Ecuador, and Uruguay. South America is remarkably deficient in nickel, chromite, and cobalt, though small quantities of all are found in Peru, Brazil, and Argentina. In the Free World, Chile is second to the United States in the size of its reserves of molybdenum (a metallic element used to harden steel).

Nonferrous metals

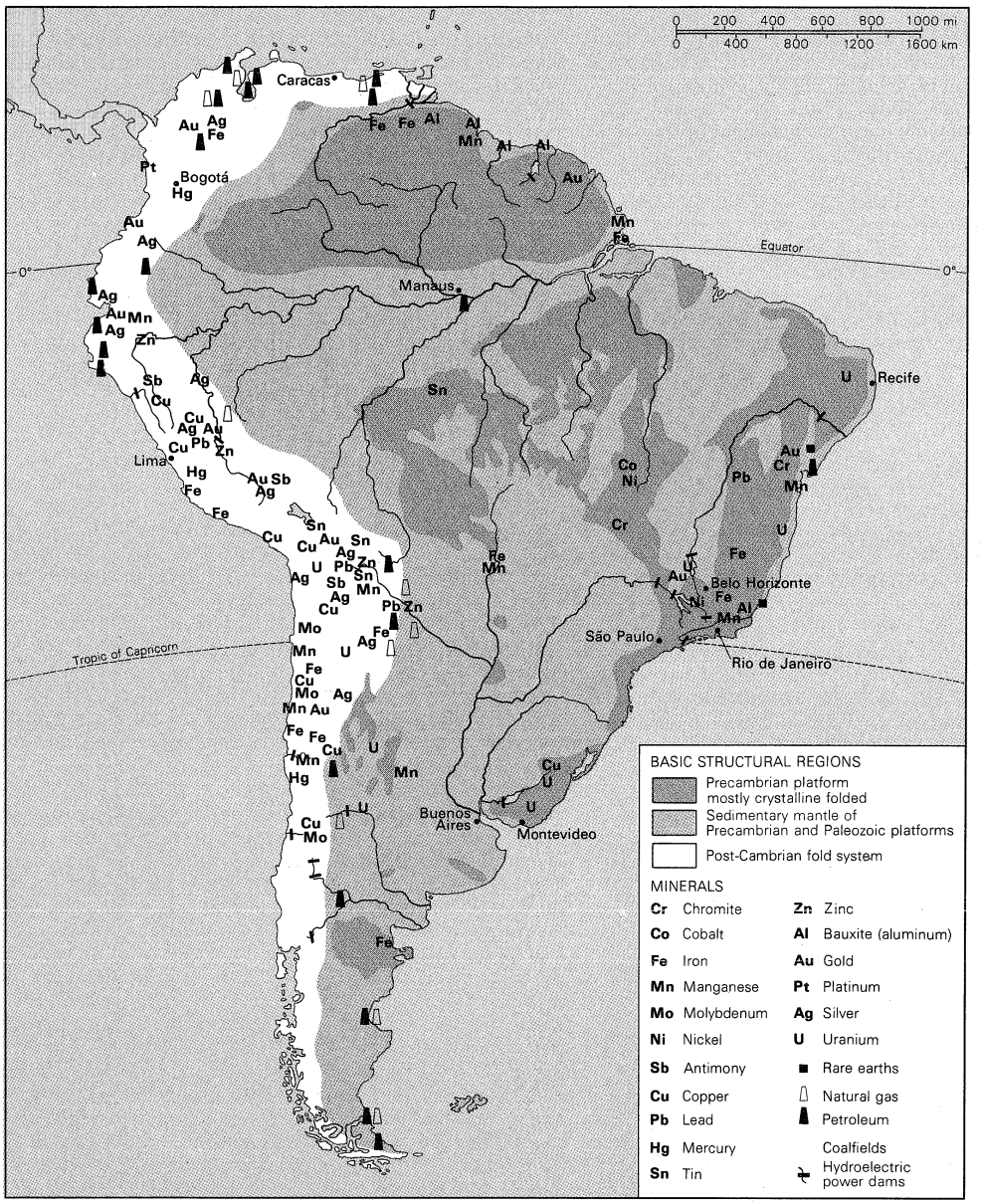
Nonferrous metals are particularly abundant in South America, where copper reserves represent 27 percent of the world's known reserves, and 97 percent of the South American reserves are found in Chile and Peru. In Chile, the Chuquicamata deposits contain the largest amount of copper known in the world, with reserves estimated to be at least 1,000,000,000 tons of ore containing 2.5 percent of copper; El Teniente also has reserves of about 1,000,000,000 tons. In Peru the most important deposits are found in the Central Andes (at Yauricocha, Morococha, Casapalca, Cerro de Pasco, and Huarón) as well as in the southern Andes, where the mine of Toquepala, opened in 1960, was the most important in the early 1970s, with reserves of 500,000,000 tons. Copper is also found in Bolivia and in Brazil (at Caraíba, in the State of Bahia).

Tin, in the form of sedimentary deposits and in veins, is found in the eastern Andes of Bolivia in a narrow belt about 500 miles long by 95 miles wide, mainly running from Oruro to Potosí. Reserves are estimated at 500,000 tons of metallic tin. South America contributes 15 percent of the world's tin production. Lead and zinc, frequently associated in nature, are dispersed among many countries but are mostly found in Peru, as well as in Argentina, Brazil, and Bolivia. Aluminum, in the form of bauxite, is exploited mostly in Surinam, where reserves are estimated at 200,000,000 tons of ore, containing 32 to 50 percent of aluminum; about 200,000,000 tons more are found elsewhere, mostly in Guyana, French Guiana, Brazil, and Venezuela. Mercury is found in small quantities (1.4 percent of the world production) in Chile, Colombia, and Peru; antimony is found in Bolivia and in lesser quantities in Peru.

Among precious metals, gold now represents only 1.4 percent of world production, although it is found all over the continent in varying quantities. In Brazil the states of Minas Gerais, Goiás, and Mato Grosso were famous in colonial times for their gold, but production diminished as a result of the exhaustion of the lodes; in the early 1970s it came mainly from Colombia, Brazil, Peru, Bolivia, and Chile. Silver is found above all in Peru, where production is more than 1,000 tons a year, but is also found to a lesser extent in Bolivia, Chile, Argentina, Colombia, and Ecuador. Colombia has the largest reserve of platinum and allied metals.

Precious stones

Many parts of South America, mainly Brazil, also were famous for their gems. The ancient bedrocks are rich in precious stones, which are found in veins of pegmatite (a coarse granite) or are washed out by streams. Dia-



Basic structural regions and principal mineral and hydroelectric sites of South America.

monds are spread over a good part of the Brazilian Plateau, notably in the region of Diamantina. Some are suitable only for industrial use, but others are of considerable value as gems. Brazil, however, contributes less than 2 percent of world production.

Other precious or semiprecious stones abound in the same region, notably topazes, tourmalines, beryls, aquamarines, chrysoberyls, garnets, and sapphires, as well as quartz of sufficiently high grade to be used in electronics (the only source apart from those in Communist countries). Colombia is famous for its emeralds, found in the Muzo mines.

Various mineral elements used in industry—such as beryllium, colombium, tantalum, thorium, lithium, rare earths, and mica—are found in Brazil and Argentina.

Nonmetallic deposits. South America has neither a significant output nor important reserves of the principal fertilizers—phosphates and potash. In the Atacama Desert (*q.v.*) in northern Chile, there are important deposits of nitrates in the form of caliche, a grey rocklike substance with a high content of sodium chloride, sodium nitrate, and iodine salts. These deposits were formerly of importance, but synthetic fertilizers have now superseded them; they are, however, still used as source of iodine. Many various saline deposits found through the continent await investigation as potential sources of salts.

BIOLOGICAL RESOURCES

Biological resources are abundant and much diversified, although not evenly distributed. Southeastern Brazil, Uruguay, the Argentinian Pampa, and Central Chile are suitable for wide-scale agriculture.

Botanical resources. Plant diversity is enormous; many South American plants are useful to man. In pre-Columbian times, Indian cultures domesticated numerous plants, some of which now play a major role in the world's food supplies.

Trees and forests. Though fewer than 200 species constitute the bulk of the neotropical forests, their mixed character is a major obstacle to large scale exploitation of timber. Many species are used as cabinet woods, including the highly prized mahogany (*Swietenia macrophylla*), from Venezuela, Brazil, Peru, and Bolivia, and several leguminous species such as rosewood (*Dalbergia nigra*); other species are exploited as general utility woods and are mainly used domestically, often as fuel. Balsa (*Ochroma* species) is the lightest and fastest growing wood in the world. Other species produce tanning materials (such as the bark of the quebracho tree), various drugs, or edible products.

Forests cover about 2,200,000,000 acres, representing 51 percent of the continent. Brazil is the most densely forested country; its total forested area is almost as great

as that of all the rest of South America put together, and it possesses more than two-fifths of the continent's accessible forests.

Food plants. Maize, or Indian corn (*Zea mays*), probably native to tropical America, is still cultivated to a large extent throughout the continent. Beans (including several species of the genus *Phaseolus*), widely cultivated by small scale methods, form an important food item in most countries. Cassava, or yucca (*Manihot esculenta*), and sweet potato (*Ipomoea batatas*) are indigenous to the New World, from which they have been introduced to the other warm regions. Potato (*Solanum tuberosum*) originates from the high Andes, where related wild species are still found. Some other plants have been domesticated in this particular environment, such as the quinoa (*Chenopodium quinoa*) and the canahua (*C. canihua*), both giving very small seeds used as cereals, and tuberoses such as ulluco (*Ullucus tuberosus*) and oca (*Oxalis tuberosa*). Squashes and pumpkins (*Cucurbita* species) have been cultivated since pre-Columbian times and have spread all over the world, as has the tomato (*Lycopersicon esculentum*), indigenous to the west coast. The peanut (*Arachis hypogaea*) originated in the Paraná River Basin, where the wild species is now apparently extinct. Cashews (*Anacardium occidentale*), now cultivated in most tropical countries, and Brazil nuts (*Bertholletia excelsa*), harvested from wild trees in the Amazon region, are widely used as delicacies but are also eaten locally raw or cooked. Cacao (*Theobroma cacao*), native to the Amazon region, was prized by the ancient Indian civilizations and is cultivated in South America, particularly in the State of Bahia, Brazil. Avocado (*Persea americana*) is native to the same region. Pineapple (*Ananas comosus*), probably indigenous to southern Brazil and the Paraná River Basin, was cultivated in pre-Columbian times throughout tropical South America and the West Indies. Papaya (*Carica papaya*) and guava (*Psidium guajava*) are both native to tropical America.

Man also introduced a great number of plants to the continent. Sugarcane (*Saccharum officinarum*) has been cultivated in all humid tropical countries since early colonial times. Bananas have become an important local food item and are exported in enormous quantities. Mango, oranges, lemons, and grapefruits are widespread. Coconut is very common; its origin is, however, much disputed.

Cereals

Among cereals, rice, introduced from Asia, is grown in Colombia, Peru, Brazil, and Uruguay but not in sufficient quantities. Wheat, raised locally in Colombia and widespread in Argentina, as well as other cereals, has been introduced to the southern part of the continent.

The two most important native spices are allspice (*Pimenta officinalis*) and red or chili pepper (*Capsicum annuum*).

Forage. Grazing is extensively practiced on all grassland areas throughout the continent. Apart from the Llanos of the Orinoco, the main grazing regions extend through eastern and southern Brazil, as well as across Uruguay and Argentine, where various breeds of cattle are raised. Stock raising also flourishes in the temperate zones of the Andes, where cattle are raised, while the upper, drier zone is mainly devoted to raising sheep, alpacas, and llamas. Foraging is also widespread in savanna areas, and even in regions such as the caatinga, where xerophytic plants grow.

Beverages and stimulants. Coffee (*Coffea arabica* and related species), imported from the Old World, is now extensively cultivated, particularly in Colombia and Brazil. The most celebrated native plant is the yerba maté, or maté (*Ilex paraguariensis*), indigenous to the Paraná Basin and still exploited in its wild state; its leaves are used to prepare a drink that is very popular in some parts of South America. Tobacco (*Nicotiana tabacum*), already smoked by Indians in pre-Columbian times, is cultivated in many countries, but mainly in Brazil and Colombia.

Pharmaceutical plants. A great number of South American plants provide valuable drugs, among which are quinine (obtained from the bark of several trees of the genus *Cinchona*, indigenous on the eastern slopes of the Andes) and the cocaine (obtained from the leaves of the coca tree, *Erythroxylon coca*, found in the eastern Andes from Peru to Bolivia). Among a great variety of aboriginal remedies still in use, some may contain valuable substances.

Oil plants. Besides peanuts and some introduced plants, South America has a great number of oil producing plants, such as the babassu palm, a native of Brazil, the nuts of which are rich in oil and are used for soapmaking. The production of essential oils could be greatly developed.

Vegetable waxes are produced mainly by the carnauba palm (*Copernicia cerifera*) of Brazil, the leaves of which are covered with a waxy secretion. Vegetable ivory is yielded by the hard seeds of the tagua palm (*Phytelephas macrocarpa*) found in northern South America, notably in Ecuador.

Fibre plants. Cotton is partly indigenous to South America and has been used since prehistoric times; large quantities are produced in Brazil, Argentina, Peru, and Ecuador. Kapok and plants producing stem or leaf fibres, such as sisal, are extensively grown. The iraca (*Carludovica palmata*), a plant looking like a stemless palm, is cultivated in Colombia and Ecuador for its fibres; extracted from the young leaves, they are woven into "Panama" hats.

Latex-yielding plants. Several plants furnish latex, from which rubber is extracted. Para rubber, or seringa (*Hevea brasiliensis*), and related species, a native of the Amazon region, was known by the Indians. Ceará and manihot rubbers, found in the driest part of Brazil, could be used if subject to suitable agricultural management. Panama rubber, or caucho (*Castilla elastica* and related species), native to the Amazon region, has so far yielded poor results.

Balata (*Manilkara bidentata*) yields a nonelastic rubber; chicle is a gum extracted from the latex of sapodilla (*Achras zapota*) and is used in the preparation of chewing gum.

Animal resources. *Mammals.* Animal resources are less diversified than botanical resources. Llamas and alpacas were domesticated in the high Andes in Inca times and probably earlier; both derive from the guanaco, the former being kept as a beast of burden and the latter for its wool. Guinea pigs were domesticated in pre-Inca times in the same area and raised for meat. The keeping of native domesticated animals did not extend beyond the geographic limits in which their wild ancestors were found, with the exception of the Muscovy duck (*Cairina moschata*). Among animals introduced to the continent were cattle brought by Columbus on his second voyage. Numerous in Brazil, Uruguay, Argentina, Venezuela, and Colombia, cattle today are distributed all over the continent where suitable grazing occurs. Horses are widespread and play an essential role in the economy of most countries. Sheep are raised successfully in the Argentinian region of Patagonia and in the Andean highlands. Pigs and smaller domestic animals are present everywhere.

Domesticated animals

Game is plentiful in most habitats, though mammals are few and specialized. Deer are represented by several species; red deer have been acclimatized in Argentina. Specialized rodents (such as agoutis, pacas, capybaras, and maras), waterfowl, and a large series of terrestrial birds (such as tinamous, guans, and curassows) provide excellent game. Many Indians, mainly in the rain forest, live upon birds and monkeys. Many animals are collected alive and exported as pets or for zoos, the heavy demand being justified by their strangeness and beauty.

Several neotropical animals provide world-famous fur or wool. Chinchilla (*Chinchilla laniger*) were native to the high Andes of Peru, Bolivia, Chile, and northern Argentina, where thousands lived in flourishing colonies. They were hunted for their delicate gray fur to the point of extinction in their wild state, except possibly in some districts of Bolivia and Argentina. Vicuñas provide the finest wool on Earth and have been hunted since the time of the Incas; they have dwindled alarmingly in the upper Andes and are still poached despite laws and a ban placed

upon trade in their fur. The giant otter (*Pteronura brasiliensis*) of the Amazon; several spotted cats, including the jaguar and the ocelot; and rodents, such as the nutria, also provide highly prized furs, but overexploitation of many species has depleted the stocks.

Fish and birds. Freshwater fish, abundant in most of the rivers, have been exploited on a large scale since the earliest time, notably in the Amazon region and in Guiana. Trout have been introduced into Andean lakes and torrents—sometimes, unfortunately, to the detriment of the endemic species.

Marine fisheries are very prosperous, since South America is surrounded to a large extent by highly productive fishing areas. In 1970, South American fishermen collected nearly 15,000,000 tons of fish, representing about 21 percent of the total world catch. Peru is the principal world producer, with a catch exceeding that of Japan.

South America is famous for colonies of seabirds along the Peruvian and Chilean coasts, resulting in the accumulation of dung (guano), a highly prized fertilizer of which about 350,000 tons are annually produced. Pinnipeds (a suborder of aquatic carnivorous animals, including seals and walruses) are exploited for their oil, particularly in Uruguay. Fur seals are distributed along the southern coasts of South America, but their numbers have been considerably reduced by hunting. (J.P.D.)

V. Human resources

ETHNIC ORIGINS AND MIGRATIONS

Four main components have contributed to the present population of South America—the American Indians (Amerindians), who were the pre-Columbian inhabitants; the Iberians (Spanish and Portuguese who conquered and dominated the continent until the beginning of the 19th century); the Africans, imported as slaves by the colonizers; and, finally, the postindependence immigrants from overseas, mostly Europeans.

Indians. During the epoch of the discovery and the conquest (that is to say, in the early 16th century), Indian societies existed at three different cultural levels, a fact that to a large extent determined the composition of the population during the colonial period and afterward.

Andean civilization

Andean society, a civilization of archaic type, comparable to Egyptian, Mesopotamian, and other Old World pre-Greek civilizations, existed on the Pacific coast with its focal centre in Peru. It is estimated that agriculture (based on maize) had been developed in this area from at least 2500 years BC, while the beginnings of the formative period of Andean civilization date to the 1st millennium BC. At that time the agricultural village was the basic community; different arts and crafts had also been developed, and, toward the end of the period, religious–ceremonial buildings began to appear.

The culture of the Classic Andean period (approximately from AD 1 to 1000) reached a higher refinement in the architectural arts; palaces and elaborate multiroom buildings were constructed, metallurgy (the smelting of copper, gold alloys, etc.) was developed, and agricultural techniques, such as terracing and irrigation, were evolved. Urbanization, which is to say the emergence of inhabited cities, as distinct from religious–ceremonial centres, also came into existence during the Classic period. With a relatively high level of political and military organization, these civilizations established several kingdoms. Neither the Classic nor the post-Classic Andean civilizations, however, had a written language. The post-Classic phase, moreover, was marked by further extension and centralization. The Inca Empire of Peru, for example, expanded from its homeland in Cuzco both to the north (to the area occupied by present-day Colombia, with its relatively advanced Chibcha cultures), and to the south (present-day Chile, with its primitive Araucanian tribes). The Incas began their conquests around AD 1200, but the expansion of their empire was greatly accelerated after 1400, until the process was interrupted by the Spaniards in 1536.

The remaining South American societies—among which lower and intermediate levels could at the same time be distinguished—had not evolved beyond the earlier phases of social development. The lower level occurred in the Guanaco area, inhabited by nomadic hunting tribes. It was located on the present territory of Argentina and Uruguay and included the extreme south (Tierra del Fuego and Cape Horn). The intermediate level occurred in the Manioca area, which covered the Amazon Basin to the Atlantic coast, occupying part of the present territory of Brazil. Although the intermediate level was mainly represented by hunting people, they also practiced agriculture, using the slash-and-burn technique (*i.e.*, cleaning ground for temporary cultivation by cutting and burning the vegetation). Andean civilization and the peoples dominated or influenced by it included perhaps one-half of South America's indigenous population and also covered the most densely settled areas of the continent. Elsewhere, the population was sparse, and quite large areas remained deserted.

The contemporary Indian population

The number of Indians at the time of the conquest is uncertain: estimates vary from 8,000,000 to 100,000,000 for North, South, and Central America combined (for the Incas, from 3,000,000 to 32,000,000). Some authorities assign to South America 6,800,000, of which a little less than half belonged to the Inca Empire or were under Inca influence. Other more recent estimates put South America's population at about 14,000,000 and seem more realistic.

Equally controversial is the origin of South America's Indians. Most anthropologists believe that they arrived in America from Asia, in successive waves, from about 25,000 BC onward—probably across the Bering Strait, which separates the extremities of northeast Asia and northwest North America. But other alternate origins and routes are also considered, along with the possibility of a local evolution. With a great deal of variation, the Mongoloid type seems to prevail among the population.

Iberians. Until the end of the era of their domination, only the Spanish and Portuguese were admitted to their South American colonies. The rigid exclusion of all other foreigners had but few exceptions, though a small number of other Europeans settled as a result of illegal or tolerated immigration. The Iberian ethnic background itself was, however, highly diversified. Apart from the Romans themselves, many pre-Roman peoples—Visigoths, Jews, Arabs, Berbers, and Moors—contributed to form both the Spanish and Portuguese population. Most of the Spaniards came from Castile and the southern regions. Very little is known about the principal regions from which the Portuguese came. It is estimated that the total number of *licencias* (authorizations to emigrate) granted by Spain was about 150,000 for the whole colonial period, which lasted from the 16th to the 19th centuries. It is possible that the number of illegal immigrants amounted to as much again. Of these, no more than 40 percent of the emigrants went to South America.

Africans. A few African servants accompanying their Spanish or Portuguese masters were the first slaves to enter the continent. Larger scale importation of slaves from Africa developed two or three decades later: the slave trade was authorized by Spain for the first time in 1518, though here again reliable quantitative information is lacking. An idea of their demographic contribution may be given by some estimates—for example, 4,000,000 for Brazil, and 3,000,000 for all Spanish America, of which only a minority went to the Andean region and an even smaller number to the southern cone, consisting of what is now Argentina, Uruguay, and Chile. Paradoxically, the slave trade was supported by the defenders of the Indians. It was also considered that African slaves were more efficient than American Indians, particularly for working on tropical plantations. Most of the slaves were from West Africa, including Angola. The slave trade ceased in the earlier years of the 19th century.

Postindependence overseas immigrants. With the attainment of independence by South American countries in the early 19th century, the legal exclusion of foreigners came to an end. Mass immigration, however, did not begin until late in the second half of the 19th century, acquiring momentum in the last three decades of the century and continuing until 1930, when it decreased

abruptly. Some 11,000,000 to 12,000,000 people arrived in South America; the great majority of these went to Argentina (more than 50 percent) and Brazil (about 37 percent). Although many later left, the demographic and sociocultural impact of this influx was tremendous in Argentina and (to a lesser extent) in southern Brazil. Immigration to other countries was numerically insignificant (although socioculturally meaningful), except in Uruguay, where because the pre-existing population was not numerous, the percentage of the foreign born was high—18 percent in 1908, and even higher in the 19th century. In Argentina this proportion reached nearly one-third of the total, and stayed at that level for many years. In both cases the contribution of post-independence immigration was proportionally much higher than in the United States at the peak of mass immigration.

Sources of immigration

The great majority of the immigrants were Europeans —Italians (forming nearly one-half of the immigrants in Argentina, one-third of those in Brazil, and probably the majority of immigrants in Uruguay), the Spaniards (one-third in Argentina), and the Portuguese (30 percent in Brazil). Other small but socially relevant immigrant streams arrived from central and eastern Europe. This source of immigration became more important after the turn of the century and especially during the 1930s and 1940s, when it included more middle class and educated people, among whom were many Jews and other refugees. After World War II another smaller wave of immigration arrived from Europe, directed mostly to Venezuela and Argentina.

POPULATION AND ECOLOGICAL DISTRIBUTION

The present population. The present population of South America is the result of four centuries of race mixture among these four components—Indians, Iberians, Africans, and overseas immigrants—and their mixed descent. Race mixture began when the first Iberians touched South America. Their own previous traditions and their basic values and attitudes, coupled with other characteristics of their conquest and colonization, facilitated intermixing not only with the Indians but in general among all the various racial and ethnic groups even although its intensity, extent, and frequency varied both among different groups and at different epochs.

Legal marriage between Iberians and Indians was tolerated, often permitted, and even, in some special circumstances, promoted. It was possible—and in certain epochs easy—to recognize mestizo (generally, mixed European and American Indian) children, though often a mestizo was considered automatically illegitimate. With regard to mixing between Europeans and Africans and between Africans and Indians and their offspring, such social permissivity did not exist—a fact that nevertheless failed to prevent generalized race mixture. This prolonged process created a great variety of physical types, resulting in the emergence of a complex terminology to describe them.

The more important types are the mestizo (called caboclo in Brazil); mulatto (European–African), zambo (African–Indian), and cholo (mestizo–Indian). When the postindependence European immigration occurred, other national groups contributed to race mixture. As a result, Argentina and Uruguay completely modified their ethnic composition.

The three ecological regions

Ecological distribution. For these ethnic groups, three regions of ecological distribution may be distinguished —Indo-America, Mestizo-America, and Euro-America.

Indo-America consists of the highlands from north to south on the western side of the continent. It corresponds roughly to the territory of Venezuela, Colombia, Ecuador, Peru, and Bolivia, where—together with a minority of white or "whitish" groups (mostly of Spanish origin)—mestizo and Indian groups are predominant. Mestizo-America is situated in the lowlands of tropical and subtropical South America. Both on the eastern and western coasts, the African influence is strongest; the area is then characterized by European groups (mostly Portuguese in Brazil), many of them with various proportions of black ancestry, as well as by Africans and mulattos. Euro-America, or the temperate zone, consists of southern Brazil, Argentina, Uruguay, and central Chile. There the Europeans are overwhelmingly predominant, and there is little African or Indian influence (with the exception of Chile, where the original Indian contribution was much more significant, although it is not any more visible from the cultural or even the somatic point of view). Geographical, historical, cultural, and socio-economic factors account for this ecological distribution.

Indo-America. It must be remembered that the Indo-American zone was the site of the higher civilization in South America and was the most densely settled at the time of the conquest. There the social and economic organization of the Spanish empire could be founded on the relatively advanced institutions of the pre-Columbian cultures, and, at the same time, Indian labour could be easily exploited. Although the original Indian population suffered what has been called a demographic disaster (during the first century of Spanish domination the Indian population of this area decreased by at least 30 to 60 percent), a substantial Indian population survived, and race mixture was intense; the Indians subsequently began their demographic recovery at about the end of the 17th century. The new nations in this area did not receive substantial European immigration in the 19th and 20th centuries, so that the European component in the population comes mostly from the colonial period.

Mestizo-America. In the lowlands of Venezuela and Colombia, however, and much more especially in Brazil, African slave labour was important. There the sparse indigenous tribes were much less well adapted to work efficiently; they also resisted enslavement and were decimated by disease and warfare. Indians were therefore replaced by imported Africans, mostly for the purpose of cultivating plantation crops. European–Indian mixture became less frequent, and the African racial and cultural influence became significant, particularly in Brazil where the Negroid (*i.e.*, African and mulatto) population constitutes perhaps 30 to 40 percent of the total. The other two South American countries with a substantial proportion of blacks (or with black components) are Colombia (no more than 30 percent black and mulatto) and Venezuela, where the black or Negroid proportion of the population is probably smaller.

Euro-America. In the temperate zone, the indigenous tribes, already scarce, were destroyed or reduced to an insignificant size, and the mestizo population was mostly overwhelmed by the avalanche of European immigration.

Another factor in the ecological distribution of ethnic groups in South America is the degree of urbanization. People of European origins, who were white or "whitish," tended to be more urban than rural and to live in larger cities rather than smaller ones. This pattern, however, was being blurred in the middle and later years of the 20th century by internal migrations. In effect, from World War II onward, these internal population movements, as well as some inter-American migrations (mostly between contiguous states and mostly from the less to the more developed states) represented a new contribution to the South American melting pot.

Racial criteria

Sociocultural factors. As a general comment on the foregoing discussion, it must be remembered that racial criteria in all of Latin America are sociocultural (that is, generally based on language, clothing, and behaviour patterns) and not genealogical (that is, based on descent, as in the United States, South Africa, and elsewhere). In Latin America, "passing" may be no more than a question of buying a European dress and moving elsewhere. This does not mean lack of prejudice, discrimination, and exploitation—which indeed characterize the whole history of South America—but it involves significant differences for the life chances of the individuals. "Race" classification becomes in any case very subjective, being based also on the particular social status of the person. (The higher the social class, the "whiter" the person is considered to be.)

Linguistic patterns. The linguistic diversity and multiplicity of South America is said to be unmatched anywhere in the world. Thousands of languages and dialects

have been cataloged, including all those existing since the conquest. Classification systems vary a great deal—from more than 100 "linguistic families" and many unrelated languages at one extreme to an extremely simplified scheme including four main families and a few subfamilies at the other.

There is also considerable disagreement on the composition of these "stocks" and how many languages should be classified. Most are now extinct, either because the tribes who spoke them have disappeared or because of acculturation into a European language or, in some instances, into another indigenous tongue. Relations between the South American languages and other New and Old World languages are still highly controversial. There are limited evidences of ties between some southern Colombian languages and one large North American linguistic family. Hypotheses have been advanced about connections between South American Indian languages of the Pacific coast and the Malayo-Polynesian languages, but linguistic and cultural evidences fail to support them.

The surviving Indian languages

The survival of Indian languages follows the same pattern as the ecological distribution of the population and is equally determined by the level of civilization. Quechua and Aymara (both languages of the Andean highlands, the former being spoken originally by the Cuzco Incas) are the most diffused languages; they are used by an estimated 61 percent of the population in Bolivia, 40 percent in Peru, and a sizable proportion in Colombia and Ecuador (where both Quechua and Jíbaro are spoken). Another important tongue is Guaraní, which, together with Spanish, is diffused most of all in Paraguay, where it is also spoken by the non-Indian population. Contrasting with this, the languages of the lower level Indian cultures have in general disappeared or remain only as scanty remnants. African dialects have been wiped out, although they did exercise some influence on Portuguese and Spanish—as also was the case with both existing and extinct Indian languages. Such countries as Argentina and Chile, therefore, have only a small segment of the population still speaking an indigenous language. A higher proportion speak such languages in Brazil, and perhaps 10 percent do so in Venezuela.

Sociological changes. All or most legal discrimination against the Indians and other ethnic sectors of the population was nominally abolished at the time of independence, or during the 19th century. The real conditions of the Indians (and to a certain extent of the Africans after the abolition of slavery in Brazil in 1888), however, remained the same or became worse, since while on one hand liberal legislation tended to eliminate all communal property and the legal existence of the Indian communities, the various forms of exploitation continued unchanged on the other. These de facto conditions were also reinforced by 19th-century biological doctrines, which held Indians and Africans to be inferior races.

In the 20th century, however, a complete change in intellectual attitudes resulted in a series of public and private initiatives toward ameliorating the conditions of these groups. Illustrations of this change in attitude are to be seen in the creation by Latin American governments of the Instituto Indigenista Americano (1940); in the (formal) restoration of the "legality" of the Indian communities, in Peru (1919); and in the many projects of agrarian reform, as well as in other legal and administrative innovations discussed and sometimes approved in different countries, even though these are seldom seriously enforced. Institutions such as the Serviço de Proteção ao Indio (Protective Service for the Indians) in Brazil, as well as missions, centres, and the like were established with the purpose of providing educational, sanitary, and other social services to the Indians. Special areas were established as reservations, particularly for tribal Indians, although these received little attention of any kind from the governments. At present, two types of situation may be distinguished—that of the Indian tribes and that of the Indian communities.

Although the Indian tribes should nominally be under the "protective" regulations existing in several Latin American countries, they must nevertheless be considered as totally or mostly marginal to the life of the national societies. Statistically, they represent an insignificant proportion of the total population (between 30,000 and 100,000 where they exist at all). They can be found in the Gran Chaco and Amazon regions and also in the selva (tropical rain forest) and mountain regions of Peru and Bolivia. Some Indian tribes are also located in Chile, Colombia, and Venezuela. These groups are disappearing —in most cases either through physical extinction, caused by poor living conditions, illness, or violent death, or through absorption into some sector of the national society.

The status of Indian tribes

The Indian communities (as distinct from tribal Indians) are related to the pre-conquest communities belonging to the higher cultures but have been deeply modified by colonial domination and Hispanic culture. They can be found mostly in Peru but also exist in smaller proportions in the other areas influenced or formerly dominated by the Inca, such as parts of Bolivia, Colombia, or even Venezuela. These communities—some of which in Peru were recognized by the state—are nominally under the formal national bureaucracy but sometimes maintain their traditional Spanish–Indian authorities, such as alcaldes (mayors) or *regidores* (council members). Those that maintain more strongly their identity and traditional nature have been characterized as "corporated" (*i.e.*, incorporated) communities (*i.e.*, in the sociological, not the legal, sense of the term). Often, however, the communities are (or were) de facto tributaries of the haciendas—the large landed estates that until recently exercised quasi-feudal dominion over the Indians. This situation, however, has been in the process of rapid transformation, particularly since the Bolivian revolution (1952) and since the social movements and political changes that have taken place in Peru, particularly in the 1960s.

VI. Political geography

HISTORIC EVOLUTION

The nature of European influences. The nature of European contact and domination and its lasting impact on post-colonial South America was determined on the one hand by the economic, political, cultural, and ideological forces orienting the Spanish and Portuguese empires and on the other by the level of civilization characterizing the indigenous population. Although economic and political factors were certainly essential, the proclaimed intentions of civilizing and converting the Indians cannot be dismissed merely as a superficial justification. The tension that resulted between ideology and behaviour also expressed the Iberian cultural heritage, in which two contradictory tendencies coexisted—a legal perfectionism, oriented by Roman universalism and Christian faith, and an extreme individualism, greatly reinforced by the normless situation of the conquest of a remote and immense New World.

Role of the Spanish crown

The Spanish crown recognized very early the "free" human nature of the Indians and rejected their enslavement. Within the conception of royal authority, as a Spanish writer, Salvador de Madariaga, put it, both "Spaniards and natives were vassals of the king and equally entitled to the King's protection." After the terrible violence and destruction brought by the first conquistadores and throughout the subsequent period of colonial rule, efforts were made by the metropolis to control the greed, rapine, and cruelty of colonial officers, as well as of other powerful groups—European, Creole (born in South America but of European ancestry), or mestizo. Formal laws and good intentions were to a great extent utopian but were nonetheless not entirely inoperative. To the extent that they had roots in the social consciousness, they introduced an ambivalence that was also characteristic of the Iberians' actual behaviour in the Indies.

The problem posed by the civilized Indians was very different from those arising from the situation in primitive areas. Regarding the former the task of dominating, exploiting, and administering was easier—the Spanish replaced the pre-existing central authorities (where they existed, as in the Inca Empire) and tended to maintain a

kind of indirect rule. It was, however, a large population, with a relatively high density and a rich cultural heritage: simple absorption was impossible. Racial and cultural hybridization generated a new synthesis in which the Indian component was not less important than the Western. In the primitive areas—and this was the situation in all Brazil and most of South America—the conquest often required prolonged struggle (an extreme case being Chile), while cultural contact and domination involved the physical destruction of the indigenous element, resulting in its making a much less important cultural contribution to the new societies. A different situation characterized the African component, which was also absorbed but which exercised a deeper influence on the national culture, particularly in Brazil.

This basic sociocultural framework, with all its inherent ambivalences and contradictions, emerged during the colonial era and exercised a deep and lasting impact on the political, economic, and social structure of the new South American nations. Only after many decades was this impact blurred by the action of other external influences, change occurring from within, and a growing sense of national identity.

Forces shaping political geography

The political geography of the region emerged to a great extent as a result of the administrative and judiciary divisions established by Spain and Portugal. The three Spanish viceroyalties of Lima (1542), Nueva Granada (1717), and Río de la Plata (1776), the captaincies general, and the *audiencias* (units that were partly administrative and partly judicial) shaped the political boundaries, while the isolation in which the Spanish metropolis maintained each unit was at least one factor resulting in the fragmentation of Spanish America into 10 (and later into 20) independent states (geography and foreign interests were probably other important causes). Though Brazil maintained its unity, its component states, as well as their pronounced regionalism, were at least in part a heritage of the Portuguese administration and the laxity of its central control. Although the political organization of the independent nations was molded after the French and American models, the political culture, values, attitudes, and behaviour—particularly during the 19th century—were to a considerable extent the reflection of Iberian politics.

The cultural heritage. Though intellectual life, and to a lesser extent education, were strongly influenced by France in particular and Europe in general (later also by the United States), much of the Spanish and Portuguese intellectual heritage was transmitted, together with the language, to the new nations. Spanish remained the official language of all the independent states except Brazil (with Portuguese), Guyana (English and some Hindi, Urdu, and Creole, as well as surviving Indian dialects), French Guiana (French and a Creole patois), and Surinam (Dutch, Surinamese, and English, as well as some Spanish and French).

Both Spanish and Portuguese have been influenced by local conditions. Spoken Spanish presents some differences between the various South American countries and also exhibits differences from the Spanish spoken in Spain. The hypothesis that the Andalusian dialect exercised the greatest influence seems controversial. At least one worker has distinguished four linguistic varieties of Spanish in South America and two main types of pronunciation. Indian influences (and, in Argentina and Uruguay, the influence of European immigration) had a certain impact. The differences between spoken Portuguese in Brazil and in Portugal itself are perhaps greater than those between Spanish as spoken in South America and as spoken in Spain.

Economic and social factors. In the economy a similar continuity is discernible. Colonial South America was a producer of raw materials—precious metals and tropical crops—and a market for European (mostly Spanish) manufactures. Local manufacturing was discouraged, and the rest of the economy—based mostly on subsistence agriculture and on some artisanal production for local lower class consumption—contributed to the surplus that helped to support the local colonial administration and the higher stratum of colonial society, as well as the Spanish (or, in Brazil, the Portuguese) royal treasury. After independence this economic framework continued to exist in the form of the primary export economy that was predominant in the region until the 1930s and that still remains an essential part of it today.

The same process that may be observed with regard to social organization and stratification was closely related, both during the colonial period and afterward, to the system of land tenure and to the pattern of labour relations. The great landed estate, the hacienda (called fazenda in Brazil), was one of the cornerstones of colonial society. The hacienda became an autonomous self-sufficient institution not only in economic terms but also politically and socially. It was a self-sufficient productive unit, with the hacendado (the great landowner) generally exercising absolute authority—protective, oppressive, or both—over all the groups living on the hacienda; these included his own family, as well as numerous servants, serfs, and slaves who provided labour for the agricultural, artisanal, and other purposes.

Institution of the hacienda

The role of the hacienda was reinforced during the 17th century by the weakening of the colonial administration centred in the cities. Increasingly, the countryside—that is to say the great landowners—acquired importance. For many decades after independence, the role of the hacienda was further reinforced by the breakdown of state organization and authority. The increasing concentration of landed property that emerged during the colonial period typically occurred against the will of the crown, as expressed in legislative measures, and sometimes in actual attempts to enforce them. Labour relations established in colonial times also reflected the contradictory attitudes of Iberian rule. The *encomienda* (which meant "to entrust" a number of Indians to a Spaniard) was theoretically established as an institution for the protection of the indigenous population and for its conversion to the Christian faith. In fact, however, it became a means of reducing the Indians to slavery or serfdom. This was an important factor in the abolition of the *encomienda* and similar institutions, but the same semifeudal relationships were reproduced in the hacienda and continued under various partially modernized or transformed forms through the 19th century and, in certain areas, during the first half of the 20th century until social change in rural areas began to gain momentum.

In colonial times three main estates (social groups) were legally recognized—the Spaniards, the Indians, and the *castas* (castes), the latter being composed of those of mixed blood (mestizo, mulatto, etc.) and of African slaves. In fact, however, the hierarchization was different—the European-born Spaniards, who had a monopoly of most government offices and of authorized foreign trade, as well as other advantages, formed the highest stratum; below them were the Creoles (white or "whitish") followed by the mestizos; in the lower strata were the mulattoes, zambos, free Africans, slaves, and—finally—the Indians.

The colonial hierarchy and its effects

This inherited social structure resulted in a failure to create a viable modern political system in the new nations, while the central economic role played by raw material exports not only reinforced the *latifundium* (landed estate system), as well as economic dependence and vulnerability, and provided an obstacle to industrialization but also determined forms of social relationships that in turn tended to perpetuate the archaic social structure.

20TH-CENTURY MODIFICATIONS

Deeper changes that were nonetheless taking place became visible and then accelerated in the 20th century. The economic expansion created by a growing demand for primary exports, the new services required as a result of this expansion, and the growing centralizing and coordinating role of the state all helped to transform South American urban society, particularly in the larger cities, and led to the emergence of a new urban middle class and to the beginnings of a modern urban proletariat and of a modern bourgeoisie. These transformations reacted in

turn on the political system, leading to an enlargement of participation. The breakdown in international trade created by two world wars and the Great Depression generated conditions favouring industrialization. At the same time urban concentration increased enormously, while the growth of the mass media and of transportation networks broke down the isolation of the countryside as well as of the large marginal areas that usually formed the greater part of the national territory. The timing, speed, and character of these transformations varied a great deal among the South American nations. In general, Euro-American countries initiated their transformation first, followed by the largest nations in Mestizo-America, and by some of the Indo-American states. Stagnation and involution (*i.e.*, deterioration of already attained levels of modernization and development) did occur. But mass mobilization and social change are the main characteristics of contemporary South America.



Countries of South America.

The emergence of new social sectors—first the urban middle classes and then the urban proletariat—deeply modified the cultural, social, and political scene. Nationalism, in different ways and degrees, characterized the outlook of the middle classes and of the populistic regimes that replaced the oligarchic regimes, which had been characterized by "limited" democracy, or (more often) undisguised authoritarianism. Military intervention continued, but its meaning changed. In the past it had represented an expression of factional conflict within the ruling elites. Increasingly, after the 1930s, it became instead a component of the "participation crises" generated by the entry of an increasingly larger proportion of the population into politics. Economic, political, and cultural "dependence" on foreign hegemonic nations—mainly the United States—became a central issue. A growing sense of national identity generated a revaluation of the Indian and mestizo heritage. Even where this element was virtually lacking—as in Euro-America—efforts were nevertheless made to re-create a national-popular tradition. Economic development and the integration of national society became universally accepted, but deep conflicts were created by the contrasting ideologies and interests of the various social sectors that were actively participating in the political arena. New regimes were established, varying from those oriented toward Marxism on the one hand to the modernizing authoritarianism of military rule on the other. The old colonial heritage was clearly disappearing. (G.Ge.)

VII. Resource development

Perhaps the most comprehensive way of summarizing the level of development of any country or region is by reference to its gross national product (GNP). This figure approximately indicates the capacity of a given country or region to produce (or to buy from abroad) the many goods and services used for the satisfaction of the needs and wants of its population. The level of production, in turn, is dependent upon the amount of available resources (quantity and quality of the labour force, land, machines, technical knowledge, etc.) and upon the efficiency with which these resources are utilized.

Average per capita income

At the beginning of the 1970s the overall gross annual regional product of South America was over $100,000,000,000 U.S. (adjusted by purchasing power), which yielded an average of $556 U.S. per head. These figures, however, naturally concealed marked disparities between the various countries. Thus, for example, in 1971 Argentina and Venezuela each had an average per capita product of more than $1,000 U.S., while Brazil, Colombia, Ecuador, Paraguay, and Bolivia all had averages of $300 U.S. or less.

Another important aspect of the development of South America is that the rate of economic growth of the region as a whole and of most of the countries which comprise it has not kept pace with that of the more dynamic economies. This is a still more serious problem in view of the population increase, which has caused South America to lag further and further behind most other regions of the world except Africa and East and Southeast Asia. Thus, only three countries—Bolivia, Chile, and Brazil—have achieved a per capita average annual rate of growth above 2.5 percent in the past decade; of the remaining countries, only Argentina and Venezuela have approached such a rate.

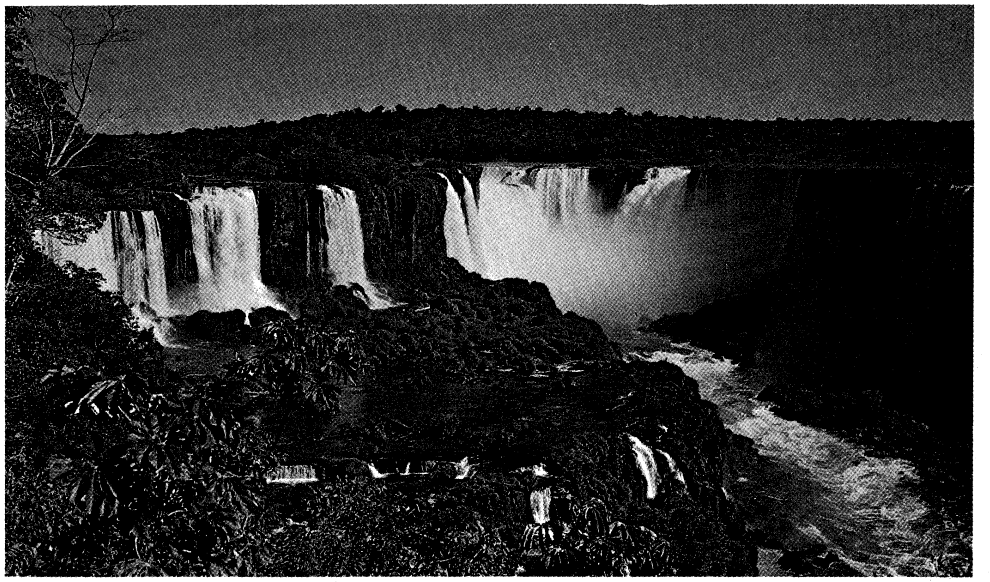
The inability of the South American economic mechanism to make the most of its available resources is seen as disquieting, particularly in view of the fact that social unrest and conflict in the region have been aggravated. Furthermore, there is a growing belief that the long-standing obstacles to economic development are of a structural nature and that some changes must be made in basic institutions to ensure a more definite and stable economic growth.

The economic life of a country can be classified in several sectors to obtain a general picture of the productive structure and of the distribution and utilization of productive resources, as the following Table illustrates.

Table 1: Percentage of Total Gross Domestic Product Contributed by Each Sector in the South American Countries in 1970*†

	agriculture	industry	mining	commerce and finance	electricity, gas, and water	transportation and communication
Argentina	15	41	2	19	2	7
Bolivia	23	18	14	11	2	8
Brazil	19	26	1	...	3	9
Colombia	30	23	3	...	1	7
Chile	10	30	10	18	2	10
Ecuador‡	31	22	2	14	2	4
Guyana	17	21	20	...	...	6
Paraguay	34	22	§	18	1	4
Peru	18	26	7	19	1	...
Uruguay	21	25	‖	...	...	...
Venezuela	7	15	20	14	3	4

*Excludes contribution to gross domestic product of miscellaneous services and government sectors. †Preliminary figures. ‡1969 figures. §Contribution of less than 0.5 percent. ‖Mining contribution included in industry figure.
Source: United Nations, Economic Commission for Latin America: *Economic Survey of Latin America*, 1970 edition.



Iguaçu Falls, where the Río Iguaçu drops about 270 feet over a series of approximately 275 separate waterfalls.
Manley—Shostal

INDUSTRIES

Stages of industrialization

The first stage of industrialization, which might be called the pre-factory stage, has been left behind by most South American countries. The second stage, commonly known as that of the traditional industries (textiles, food, beverages, tobacco, sugar), is still present in many of them, although its importance varies from country to country as a result of special conditions or historical circumstances. The final stage, the development of basic industries and the production of simple machinery, was being reached in the early 1970s by Argentina, Brazil, and Venezuela.

The process of import substitution that began in some countries even before the Great Depression and that later accelerated during World War II gave birth to many industries producing intermediate and capital goods, which in some instances still survive only because of protection from foreign competition afforded by means of subsidies, tariffs, and trade restrictions. In recent years, however, import substitution has shown signs of weakening in the traditional industries sector, as well as in the production of consumer goods.

Manufactured products are sold at high prices not only because of trade protection but also because of the small scale of production, the use of inadequate technology, and defects in the economic infrastructure (structure of basic services), mainly in the form of insufficient transport and energy.

The slow growth of the industrial sector and the adoption of production techniques requiring a high ratio of capital to labour resulted in a very limited contribution of industrialization to the creation of new sources of employment, especially in view of the rapid growth of the labour force.

The gross domestic product (GDP) originating in the industrial sector is approximately 25 percent of the gross national product. Of this total, about 20 percent represents manufacturing and the rest civil construction. About 16 percent of the labour force is occupied in the manufacturing sector.

In the two most industrialized countries—Argentina and Brazil—the output of foodstuffs, beverages, and tobacco accounts for more than a quarter of total manufacturing output; the production of metallurgy and mechanical industries represents another quarter; while chemicals and petroleum refining contribute about 15 percent, and textiles, footwear, and apparel contribute another 15 percent. In Chile, Colombia, Peru, Uruguay, and Venezuela, the last two groups of products comprise about 45 percent of manufacturing output; metallurgy and chemicals, meanwhile, represent only about 30 percent.

In 1970, South America produced 22,000,000 tons of cement, 6,403,000 tons of pig iron, 9,013,000 tons of steel ingots, rolled steel, plates and sheets, and 635,000 motor vehicles. Furthermore, most countries produced their own household appliances.

Unlike some manufacturing industries that have become sophisticated, the construction industry in South America is still in an elementary stage. Construction techniques use much labour, quality is low, and costs are quite high. Despite the existence of a huge housing shortage, building is proceeding at a very slow pace.

ELECTRIC POWER AND IRRIGATION

The total generation of electric energy in 1970 amounted to 107,600,000,000 kilowatt hours (kWh), of which 53 percent was generated by hydroelectric plants and 47 percent by thermoelectric plants. South America's first atomic energy plant is in Argentina and was scheduled for completion in about 1973.

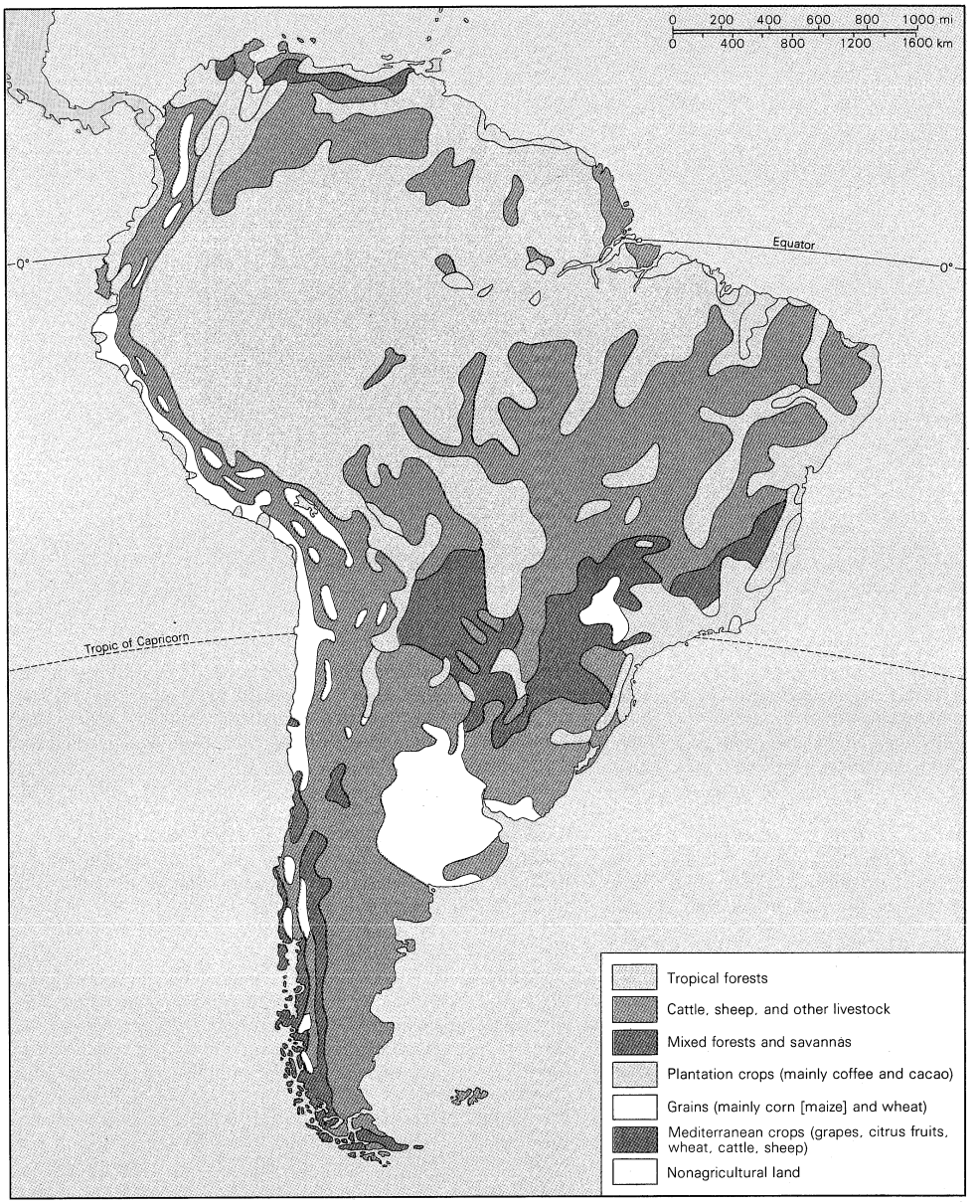
Electrical energy output

Per capita electric energy output is very high in Venezuela (1,215 kWh), Argentina (704 kWh), Chile (806 kWh), and Uruguay (739 kWh). The average per capita output for the whole region, however, is about 552 kWh, compared with a world figure of almost 1,350 kWh. Industrial consumption accounts for almost 40 percent, on the average, but is higher in the four countries mentioned, either because of the requirements of petroleum production and mining or because of the progress of industrial development. Household consumption takes another 40 percent, and the rest represents either transmission losses or other uses.

Private installed capacity, which consists mainly of thermoelectric plants operated at high cost, still represents a large 20 percent of total installed capacity. The trend, however, has been toward increases in hydroelectric power—in part because of the increasing importance of the public power services as distinct from private power generation.

Argentina, Brazil, Uruguay, and Venezuela have ambitious electrical installation programs that will take advantage of the region's abundant hydroelectric resources. This will undoubtedly require extensive use of interconnecting systems that have already been effectively developed in such countries as Brazil, Chile, and Uruguay. In Brazil and Venezuela, national interconnection plans have been hampered by the existence of systems operating at different frequencies. It has been estimated that, in 1967, South America had an installed hydroelectric capacity equal to only 2 percent of its total potential or usable capacity.

Most of the hydroelectric projects are also used for



Agricultural regions of South America.

irrigation purposes, but there are many instances in which the first purpose has been achieved, but the construction of distribution canals for irrigation is still awaited.

About 6 percent of arable land and of land under permanent crops is irrigated. Notwithstanding the fact that the largest agricultural countries account for most of South America's 13,210,000 irrigated acres (5,345,000 hectares), Chile, Peru, and Ecuador show the highest proportion of irrigation in relation to cultivable land. In Peru, irrigation activities are confined to the narrow strip of coastal plain. In Chile, 25 percent of arable land uses irrigation water; this land consists mostly of farms of less than 25 acres in size. Because of the relative scarcity of farmland in these countries, it may be said that large-scale irrigation is practically a condition of national existence.

Argentina accounts for almost 30 percent of the land under irrigation in South America; this area is devoted mainly to the growing of fruit, especially grapes.

AGRICULTURE AND LAND USE

Even though agriculture is a large sector in South America's economy, several countries have to import agricultural commodities that could be supplied by expanding local or regional production. In 1971 the largest crop in terms of the area harvested (41,000,000 acres) and in output (25,200,000 tons) was corn (maize); the second largest was wheat (17,900,000 acres and 8,600,000 tons, respectively), followed by rice with 14,000,000 acres and 8,300,000 tons. Cotton, potatoes, barley, sweet potatoes, rye, millet, and sorghum are also important. There are 203,000,000 head of cattle (18 percent of the world total), 17,000,000 horses (26 percent of the world total), 83,000,000 pigs, and 123,000,000 sheep.

Land productivity is, on the whole, far below that prevailing in other regions. The increases in production are chiefly attributable to extensions of the cultivated area. Output per acre in 1970 was equal to 0.65 ton for corn, 0.49 ton for wheat, and 0.77 ton for rice.

Rural unemployment

In spite of the large shifts of population toward the cities in recent years, a large proportion of the widespread unemployment prevailing is still found in the rural areas. About 39 percent of the economically active population works in the agricultural sector, as compared with 52 percent for the world as a whole. The lack of educational skills among farm workers, the widespread existence of unemployment conditions, and the limited use of modern methods and management techniques result in a very low average product per worker; in 1963 to 1965, this amounted to $588 U.S., as compared with $1,769 U.S. per worker in the nonfarm sector.

South America extends over an area of 4,400,000,000 acres (1,780,500,000 hectares), of which only 5 percent is arable land or land under permanent crops, 23 percent is occupied by meadows and pastures, 51 percent is covered by forests, and 21 percent has other uses. At the beginning of the 1960s there were 6,780,000 farm holdings (more than 60 percent operated by owners), with an average size of 200 acres per holding. It is a general conviction throughout the region that land utilization is inefficient. Inflation control, land taxes based on productive capacity, agrarian reforms that consist of more than changes in the system of land tenure, as well as other policies, are all felt to be helpful in improving the efficiency with which land is used. The great shortage of farm capital (as well as its high cost) is, however, one of the most serious difficulties facing agricultural development in South America.

In short, the one factor preventing a rapid growth in agricultural output is an unfavourable relation between product prices and input costs. Capital, fertilizer, modern machines, and other new inputs command very high prices, while prices received by farmers tend to be low.

VIII. Commerce

Table 1 (above) shows that the percentage of total product contributed by commerce and finance in South American countries averages out at about 18 percent. This is composed mainly of the value of the services of all the people engaged in distribution, wholesale, retail sale, advertising, and financing activities and also includes the rental value of the buildings used for this purpose, as well as interest on the capital invested. Approximately 9 percent of the labour force is engaged in this sector.

Banks and financial institutions are usually large enterprises with branches in many cities and towns; a high proportion of these are government owned. Wholesale and particularly retail business enterprises, on the other hand, are mostly individual concerns and in many cases are family-type shops. Department stores or chain stores are seldom found in South America. Full-fledged supermarkets were a somewhat recent phenomena in the early 1970s.

Characteristics of commerce in South America

In contrast with some highly sophisticated industries in the manufacturing sector, distributive and retail firms use rather elementary methods and are old-fashioned in organization. This is probably due to the scarcity of managerial skills and to the relatively small share of credit that this sector receives in comparison with industry and agriculture. Furthermore, the development of commerce depends heavily on an efficient system of communications and transportation, which is still lacking.

TRANSPORTATION

In an area the size of South America, an efficient system of transportation is necessary for the expansion of national markets and for the development of the hinterland. Unlike North America, South America has no integrated transportation system. There are many separate road and rail systems of varying extent and type. In recent years, however, internal and international air routes have begun to form a reasonably well-integrated and complete transportation system.

Waterways. Traffic on the inland waterways faces many drawbacks—such as dry seasons (and sometimes dry years), the direction of water flow, and difficult rapids. There are two systems of international importance—the Paraná–Uruguay basin (which includes territory in four countries) and the Amazon Basin (six countries), each having several thousand miles of navigable waterways. Furthermore, there are three other minor systems—the Magdalena in Colombia, the Orinoco in Venezuela, and the São Francisco in Brazil. The remaining rivers are unsuitable for navigation. In general, the volume of traffic on the waterways is small, and the prospects for increasing it are limited by the above-mentioned causes. The major products transported by waterway are timber and tropical fruits.

Maritime transport. Sea transport can be divided into three categories. The first of these is represented by the service linking the different ports along a country's coastline, as in Chile, Brazil, and Argentina, which have established regular cargo and passenger services of this type. Second are the services running between the various South American countries. Third are regular services running between countries in South America and other countries of the world.

In 1969 South America's inland waterway fleet was composed of 25 tankers (21 of them belonging to Argentina) and 55 other vessels (37 owned by Argentina). The coastal fleet amounted to more than 100 tankers (including gas tankers), of which 40 were Brazilian and about 30 Argentinian, and about 260 other vessels. In 1971 the oceangoing fleet had 136 tankers, 21 passenger vessels, and numerous cargo vessels, tramp steamers, refrigerator ships, livestock carriers, and bulk carriers. The tonnage of this merchant fleet was 5,742,000 deadweight tons, representing only about 1.6 percent of the world total.

Late in the 1960s, South American ships transported 30 percent of the volume of imports and 10 percent of the volume of exports, representing an average of 13 percent of the total volume of foreign trade. A determined effort is being made by several countries to develop and enlarge national merchant marines, partly in order to arrest trends that in the past have resulted in their trade being carried by ships from outside the region and partly as an instrument to promote regional integration and to improve the national balance of payments.

Problems of rail transportation

Railways. The position of the railways in South America is unsatisfactory. Rail transport is plagued by operational problems and by a persistent loss of importance as a result of competition from road and air transport. Proportionally, railways handle only a small amount of goods because most bulky items moved are for export and are produced close to the coast. Almost all lines are single track, making traffic very slow and discouraging passenger traffic. Many countries have two or more gauges, hindering the possibility of an efficiently integrated rail system. Railways administrations, however, were making continuous efforts in the early 1970s to improve services and reduce budget deficits.

In 1970 South America had over 8,300 locomotives (steam, diesel, and electric), close to 175,000 freight wagons, and 10,000 carriages. This rolling stock was used over a track length of about 58,000 miles, of which about 29,000 belonged to Argentina and about 12,000 to Brazil. Except for some minor industrial lines, almost all railroads are government owned.

Roads. The length of roads used in South America is impressive. In many countries, however, only a small percentage is actually paved, while in other instances they are not wide enough for two vehicles to pass conveniently. In 1971 there were about 1,047,000 miles, of which only about 84,000 miles were paved; the rest were improved roads or simply dirt roads.

In the early 1970s the improvement of existing roads and the construction of new ones was proceeding rapidly—perhaps because some countries, such as Argentina and Brazil, manufacture motor vehicles in large quantities, while other countries are interested in expanding their automobile output. Furthermore, the inefficient service of the railroads has opened the way for the development of truck transportation, which has taken an increasingly large share of the volume of goods carried by land. In 1971 there were 7,580,000 motor vehicles registered in South America.

In developing international highways in the early 1970s, particular attention was being paid to road-integration projects in which the Inter-American Development Bank and the World Bank were heavily engaged. In the construction of the Brasília to Lima route, important progress has been made. Argentina is investing large amounts to improve the integration of highways between its capital of Buenos Aires and Bolivia, from Mendoza to Santiago (Chile), and in the northeast region bordering Paraguay and Brazil. At present, Brazil has the longest network of roads belonging to the Pan-American Highway system; the system itself extends all over Latin America, and in the early 1970s there were 14,000 miles

(23,000 kilometres) of paved road, 6,800 miles of all-weather road, 1,050 miles of dry-weather road, and 4,350 miles of impassable road.

Air transport. Air transportation developed very rapidly after World War II, mainly because it is well suited to the area's requirements and avoided the forbidding geographical obstacles that hamper surface facilities. This was particularly true with respect to passenger traffic but applied less to the handling of bulky freight.

Each country tends to have its own system of internal air services, operated generally by government-owned but sometimes by heavily subsidized private companies. This circumstance creates certain gaps in the otherwise reasonably well-integrated air transport system that has developed over the continent.

International air links

All of the South American capitals and most of the large cities are linked by direct air transport services with the major traffic centres of the United States and Europe. In 1967 a scheduled freight transport service, using jet aircraft, began operations between North and South America. In 1969 the South American airlines recorded an extraordinarily large increase of more than 20 percent in total traffic (passengers, freight, and mail), entirely attributable to the growth of international traffic. Domestic traffic grew by only 2 percent. These developments resulted in a further decline of the share that South America held in total world air transport in previous years. In the early 1970s, this was less than 4 percent.

INTERNAL TRADE

The relative importance of intra-regional trade continues to be small, in spite of its rapid growth in recent years. Intra-regional trade accounted for about 11 percent of total exports in 1969 to 1970, as compared with intra-regional trade in western Europe, where it represented 67 percent of exports in 1970. In South America there exists a firm conviction that intensification of intra-regional trade is a necessary condition for rapid economic growth first, because it would help to reduce the region's excessive dependence on a few foreign markets; second, because it would diversify exports; and third, because it would help to solve balance-of-payments problems.

South American trade with the rest of Latin America is centred in a few countries. Argentina (12 percent), Brazil (19 percent), and Venezuela (18 percent) account for 49 percent of exports, while Argentina (11 percent), Brazil (18 percent), Venezuela (11 percent), and Chile (6 percent), absorb 47 percent of imports. During the early 1970s, however, the degree of concentration declined, mainly because of the rapid increase in trade among the Central American countries promoted by the Central American Common Market.

The Latin American Free Trade Association

All the independent South American countries except Guyana, as well as Mexico, in the early 1970s belonged to the Latin American Free Trade Association (LAFTA), which—despite formidable obstacles—had forged ahead. LAFTA efforts were directed toward designing a common trade policy for member countries, toward the gradual and automatic reduction of import duties and other restrictions on imports from the rest of the world, and toward arriving at agreements to compensate trade payments between member countries, as well as at making reciprocal credit arrangements between central banks. In addition, Bolivia, Chile, Colombia, Ecuador, and Peru have formed the Andean Subregional Group for the purpose of reaching agreement on common trade problems, such as an external tariff, reductions in tariffs applicable to subregional products, and coordinating their policies toward foreign investment.

In the early 1970s more than 11,000 products on which some concession of one kind or another had been granted entered the intra-area trade. Trade among LAFTA members consists mainly of basic commodities (78 percent) and semi-manufactured and manufactured goods (21 percent). In the first group the most important are foodstuffs, beverages and tobacco, crude materials, and nonferrous metals. In the second group the main items are chemicals, machinery, and transport equipment.

All these efforts toward integration are partially responsible for the rapid growth of the area's internal trade and for a large contribution toward a greater balance in trade among these countries. During the 1960s intra-area surpluses and deficits have declined considerably.

EXTERNAL TRADE

The external sector represents a key element in South America's economic growth. Essential imports, particularly capital and basic intermediate goods, are needed to accelerate the industrial process. Under conditions prevailing in the early 1970s, exports and net external financing were not providing sufficient income to pay for those imports.

Despite a satisfactory increase in trade recorded by the early 1970s, South America's share of world trade remained small and showed a slightly declining tendency, primarily because the growth in trade between major world industrialized countries had been quite impressive. The share of South America was about 4 percent of world imports and 4.5 percent of world exports. South America's share in total United States imports, approximately 7 percent, was also declining.

In 1968 South America's major exports, in value terms, were foodstuffs (50 percent), fuels (20 percent), and crude materials (17 percent). Within the first group the most important are sugar, bananas, cocoa, coffee, tobacco, beef, maize, and wheat; in the second group, petroleum and petroleum products; and in the third group, linseed oil, cotton, cattle hides, quebracho, fish meal, wool, copper, tin, iron ore, lead, and zinc. More than 30 percent of these exports went to the United States, about 30 percent to western Europe, and 12 percent to Latin America.

About 70 percent of South America's imports in 1968 consisted of machinery, vehicles and parts, chemicals and pharmaceuticals, paper and paperboard, textile products, and other manufactures; 40 percent of total imports were from the United States, 30 percent from western Europe, and 14 percent from Latin America.

In general, South America's foreign trade sector has shown little diversification of exports, a heavy dependence on imports for domestic supplies of industrial goods, and a marked imbalance in trade with the industrialized countries.

(H.F.A.)

IX. Demographic patterns

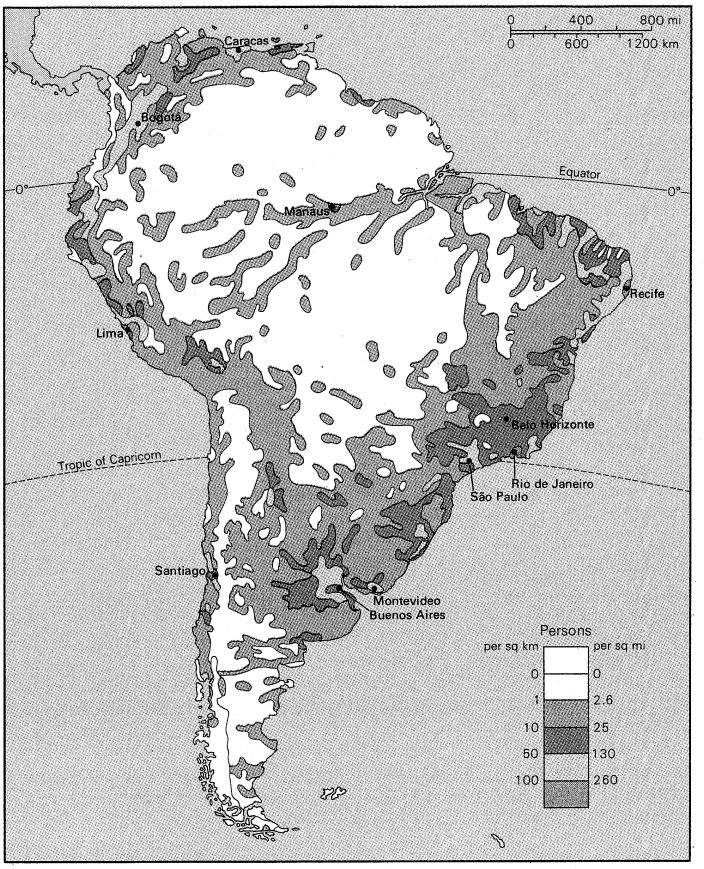
Censuses have been taken in South America since the colonial period and have been held with some regularity since the 19th century. Usually, in most of this period, their reliability was low. Since 1950, census data and other official statistics have become considerably more reliable. Failure to record deaths and births—one of the major problems—has decreased, although questionable statistics may still frequently be encountered. Such deficiencies, however, are partly compensated by better census information.

CURRENT TRENDS

The growth rate. An estimated 190,000,000 people lived in South America in 1970. In the 50 years between about 1920 and 1970, the population had increased by 129,000,000, and by the end of the 1970s another 50,000,000 will have been added. The region has one of the highest growth rates in the world—3 percent for tropical South America (1963–70). Foreign immigration is not the cause: its contribution has been demographically insignificant, except during the period from 1860 to 1930 for Argentina, Uruguay, and (to a much lesser extent) Brazil. Natural growth is the only significant factor in this population explosion.

The population explosion

The region is now at the peak of a phase of demographic transition, a modernizing process that includes three main stages. The first of these is the traditional stage, in which the population balance is maintained by both high birth and death rates; second is the transitional stage, in which decreasing mortality combined with stable high fertility (or else with a delayed or slow decline, or both) generate a high growth; finally is a modern stage, in which low mortality and low fertility produce a small demographic increase. Areas developed earliest experi-



Population density of South America.

enced the same process, with two differences—they experienced first a lower fertility in the traditional stage and, second, a slower or delayed decline, or both, in mortality.

It must nevertheless be remembered that some 60,000,000 Europeans emigrated overseas. Natality in Europe's pre-industrial phase never exceeded (usually) 30 to 35 per thousand; in South America it is 40 to 50, with three exceptions—Argentina and Uruguay (with 22 to 23), which have been at the modern stage since the 1930s, and Chile, which has experienced declining fertility since the 1920s (but whose natality is still at 39 per thousand). Mortality decline in Europe often took 100 or 150 years. In 1900 the South American average life expectation at birth was 25 to 30 years (not including Argentina and Uruguay). In 1970 it was estimated at a little more than 60 years. The great change occurred in all countries (except the two "first comers," Argentina and Uruguay) between 1920 and 1940 and still continues. Mortality declined spectacularly in 40 or 50 years—or even in 20 (as in Chile).

The fertility rate. The population explosion is the result of a traditional type of fertility rate (the highest ever known), combined with a quasi-modern mortality rate. The reason for this discrepancy in comparison with experience elsewhere in the Western world lies in the fact that while mortality can be reduced through the relatively fast diffusion of sanitary and medical innovations, a spontaneous fertility decline requires complex social transformations (in social stratification, and in urbanization, mobility, family, and life-style) correlated with psychological changes. These demand much time. It is also possible that cultural values peculiar to the Latin and Indian societies produce additional obstacles to spontaneous limitation of fertility. Government-sponsored policies to induce family planning are meeting two ideological obstacles: the attitude of the Roman Catholic Church and that of the political parties toward birth control. Though resistance has declined in the church, the politicians still remain indifferent or hostile to birth control. The fact that the region is underpopulated disguises the enormous strain imposed on the economy by too rapid a rate of demographic growth.

Obstacles to family planning

Spontaneous changes are nonetheless taking place. As in Europe, birth control is initiated by the urban population and by the middle classes. In South American cities the average number of live births per married woman becomes lower with increased education. In Bogotá (Colombia), for instance, women with no education had an average of 7.9 live births, while those with secondary education had 4.2, and those with some college education had 3.8. The same pattern was observed in eight capital cities studied (1963 to 1964). Rural and urban areas also differed in fertility: in Brazil the observed average was 4.9 live births in urban places and 7.3 in rural areas; the same pattern was observed in other countries. Because of the large rural population and the limited size of the educated group, the overall impact of the differentials is not yet reflected in national fertility rates. The process, if left to its own spontaneous evolution, may require several decades to take place, while mortality is meanwhile expected to decline further in various countries. Hence the alarming projections for population increases mentioned above. But if fertility is not curbed or economic development is not accelerated, the operation of Malthusian laws (enunciated by the 18th- and 19th-century English demographer Thomas R. Malthus, who stated that when population outstrips its means of subsistence further growth will be checked by disease, famine, or war) would seem to ensure that the mortality decline would be interrupted or reversed. In this case the expected growth would be reduced.

EFFECTS OF RAPID POPULATION INCREASE

Rapid population increase has important demographic and social effects. Two examples are illuminating.

At the peak of population growth, the proportion of children tends to be very high, while in the third stage it is low. In South America the population under 15 years is 42 to 48 percent of the total (except, as noted, in Argentina, Chile, and Uruguay). As a consequence, the working-age group is greatly reduced—55 percent in Latin America, as against 60 percent in France or England. The dependency ratio (the number of young and old dependents per 100 persons of working age) increases. The Latin American average was quite high in 1970—85 percent—with its minimum in Argentina and Uruguay (57 percent) and the rest oscillating between 80 and nearly 100 percent. This high ratio means a heavier burden for the working group, while the economy is not yet able to raise the productivity needed to compensate for it.

Population growth in the cities

Another crucial consequence is the so-called urban explosion. Already, near the turn of the century, in about 1890, Argentina and Uruguay were among the most urbanized countries in the world, but their urban growth was due to mass foreign immigration. The dramatic in-

Table 2: Percentage of the Population Living in Cities with 20,000 Inhabitants or More*

	1890	1920	1930	1940	1950	1960	1970
Argentina	23	37	38	41	51	55	63†
Uruguay	30	30	35	38	43	56	...
Chile	14	28	32	37	42	53	...
Venezuela	8	11	13	19	33	45	60
Colombia	7	7	10	14	21	31	...
Brazil	9	13	14	16	21	29	36
Peru	8	6	10	13	18	26	...
Bolivia	4	9‡	10	20	20	20	21
Ecuador	12	10	13	16	18	25	31
Paraguay	4	...	14	14	17	18§	19
French Guiana	0	0	0	0	0	0	55
Guyana	...	...	20	20	20	26	23
Surinam	...	39	32	31	40	34	33

*Data or estimates not corresponding to the year indicated are given in the column nearest actual date. †25,000 or more. ‡1910. §Estimate from CLAPSECO, *Situación Social de América Latina* (1969).
Source: *Growth of the World's urban and rural population, 1920-2000*, United Nations, New York, 1969. Various official country sources. Date for 1890 from Adna Weber, *The Growth of Cities in the Nineteenth Century* (1963).

crease in urban concentration evident in the early 1970s began approximately in the 1930s. In all of Latin America, urban centres with more than 10,000 inhabitants increased from 25 percent (in 1950) to an estimated 50 percent in 1970.

South America is one of the most urbanized regions in the world, following the industrially advanced areas. Although the rate of growth in larger cities has somehow decreased since 1950, the urban population still is, and will be in future, mostly concentrated in the bigger cities: the majority, or a sizable proportion, of the urban (and in some cases of the total) population lives in a single urban centre. This situation prevails in Uruguay, Argentina, Chile, Paraguay, and Peru.

Demographic growth has a direct as well as an indirect influence on the urban population explosion. Contrary to widespread opinions in some South American republics, the principal component of urban growth is natural increase, not internal migration. This is the case in Venezuela, Chile, and other countries. Internal migration, nevertheless, remains important. Here the demographic influence is indirect and operates in conjunction with other factors—such as unfavourable conditions in the countryside, including the concentration of land ownership in a few hands and the rejection by the rural population of the resulting extremely low standard of living.

High urbanization is coupled with a very low population density in South America as a whole with variations among individual countries, in the early 1970s, from a low 9 per square mile (Guyana) or 12 (Bolivia) at one extreme, to 53 (Ecuador) or 43 (Uruguay) at the other. Large nations such as Argentina and Brazil stand respectively at 22 and 30 per square mile.

The averages disguise the existence of sharp internal inequalities in density. High density is found in South America only in urbanized areas—that is, in large cities or metropolitan areas and their immediate hinterland. More and more, South America is coming to resemble a vast desert ocean, dotted with a few scattered densely populated islands. Though geographical and physical environmental factors play a role, high concentration is largely the consequence of uneven internal development caused by international and domestic factors. The present trend is toward an even higher imbalance, although in several countries the state is intervening in an attempt to achieve a more rational use of the national territory. (Ed.)

BIBLIOGRAPHY

General works: PIERRE BIROT, *Les Régions naturelles du globe* (1970), contains two chapters on the Andes and on the old shields that constitute a modern approach to the structure of South America; JEAN DORST, *South America and Central America* (1967), a natural history of the continent, region-by-region; E.J. FITTKAU *et al.* (eds.), *Biogeography and Ecology in South America*, 2 vol. (1968–69), a series of valuable contributions on various aspects of natural history and relations between man and environment.

Geological history: HEINRICH GERTH, *Der Geologische Bau der Südamerikanische Kordillere* (1955), a controversial book from which our modern views on the geology of the Andes originated; and *Geologie Südamerikas*, 2 vol. (1932–35), a comprehensive book on the geology of the continent, now somewhat outdated; WILLIAM F. JENKS (ed.), *Handbook of South American Geology* (1956), a country-by-country survey of the geological history of the continent and an analysis of the main formations; HANS STILLE, *Einführung in den Bau Amerikas* (1940), a classic book on the geology and structure of the Western Hemisphere; GEORGE W. STOSE (ed.), *Geologic Map of South America* (1950); E.F. SUSZCZYNSKI, "La Géologie et la tectonique de la Plateforme amazonienne," *Geol. Rdsch.*, 59:1232–53 (1970), a modern view on the structure of the oldest parts of South America; L.G. WEEKS, "Paleogeography of South America," *Bull. Am. Ass. Petrol. Geol.*, 31:1194–1241 (1947).

Physical geography: "The Soil Resources of Latin America," *World Soil Resources Rept. 23* (1965), a survey of soils and land use; HENRY W. BATES, *A Naturalist on the River Amazon*, 2 vol. (1863), a classic book on the discovery of Amazonia by one of its first modern explorers; ISAIAH BOWMAN, *The Andes of Southern Peru* (1916), a comprehensive view of geological structure and natural history; D.L. BRAMAO and LEMOS PETEZVAL, "Soil Map of South America," *Proc. Int. Congr. Soil. Sci.* (1960); GILBERT J. BUTLAND, *Latin America: A Regional Geography* (1960), a valuable contribution; FRED A. CARLSON, *Geography of Latin America*, 3rd ed. (1952), describes the structure of the continent and the natural regions; PRESTON E. JAMES, *Latin America*, 4th ed. (1969), a broad survey; CARL TROLL (ed.), *Geo-ecology of the Mountainous Regions of the Tropical Americas* (1968), a series of contributions to ecological problems in relation to relief and climate in the Andes.

Vegetation and animal life: MARSTON BATES, *The Land and Wildlife of South America* (1964), a book of interest to the general reader by a leading naturalist; ANGEL CABRERA and JOSE YEPES, *Historia natural Ediar. Mamíferos sudamericanos* (1940), the only comprehensive book on neotropical mammals and their life; CARL H. EIGENMANN and WILLIAM R. ALLEN, *Fishes of Western South America*, 2 vol. (1942), a key to neotropical fishes and their characteristics; THOMAS H. GOODSPEED, *Plant Hunters in the Andes*, 2nd ed. rev. (1961), contains firsthand data on high mountain vegetation; STEPHEN HADEN-GUEST, JOHN K. WRIGHT, and EILEEN M. TELLAFF, *A World Geography of Forest Resources* (1956), an inventory of world forestry containing a chapter on South America; ROBERT C. MURPHY, *Oceanic Birds of South America*, 2 vol. (1936), oceanography of the South American seas and life histories of all the marine species; PAUL W. RICHARDS, *The Tropical Rain Forest* (1952), a classic book on the rain forest and its characteristics throughout the world; RODOLPHE MEYER DE SCHAVENSEE, *A Guide to the Birds of South America* (1970), a key to bird life with the description and distribution of every species; G.G. SIMPSON, "History of the Fauna of Latin America," *Am. Scient.*, 38:361–389 (1950), a geological history of neotropical mammals and their distribution; FRANZ VERDOORN (ed.), *Plants and Plant Science in Latin America* (1945), contains contributions on vegetation, agriculture, and land use.

Natural resources: A.O. HERRERA, *Los recursos minerales de América latina*, rev. ed. (1965), a survey of the resources of the continent; CARL A. LAMEY, *Metallic and Industrial Mineral Deposits* (1966), a world survey of mineral resources, with much data on South America; ULRICH PETERSEN, "Metallogenic Provinces in South America," *Geol. Rdsch.*, 59:834–897 (1970), geological origin and localizations of the main sites where ores have been found.

Human resources: An excellent general survey of ethnic origins, geographic distribution of the various groups, and racial relations, may be found in MAGNUS MORNER, *Race Mixture in the History of Latin America* (1967); and in MARVIN HARRIS, *Patterns of Race in the Americas* (1964). Two major contributions on the South American Indians are JULIAN H. STEWARD and LOUIS C. FARON, *Native Peoples of South America* (1959); and CLARK WISSLER, *The American Indian*, 3rd ed. (1950). Steward's work is a general summary of the monumental *Handbook of South American Indians*, 7 vol. (1946–59). A typology of the ethnic origins and their distribution in the various South American subcultures is provided by CHARLES WAGLEY and MARVIN HARRIS, "A Typology of Latin American Subcultures," *Am. Anthrop.*, 57:428–451 (1955). On indigenous languages, in addition to chapters in the previously mentioned books, one of the classics is CESTMIR LOUKOTKA, *Classification of South American Indian Languages* (1968). The most complete statistics and interpretations on mass European immigration (before 1930) is found in WALTER F. WILLCOX (ed.), *International Migrations*, 2 vol. (1929–31, reprinted 1969). A more recent contribution is WILFRID D. BORRIE *et al.*, *The Cultural Integration of Immigrants* (UNESCO, 1959). In addition, the International Labour Office's publication, *Indigenous Peoples: Living and Working Conditions of Aboriginal Populations in Independent Countries* (1953), gives data and documents on the situation of the Indians in South America.

Political geography: J.P. COLE, *Latin America: An Economic and Social Geography* (1965), is a comprehensive treatise. For a perceptive interpretation of the Spanish in South America, see SALVADOR DE MADARIAGA, *El auge del Imperio Español en América* (1945; Eng. trans., *The Rise of The Spanish American Empire*, 1965). An important anthology, including sources and interpretation, covering the period from conquest to modern times, is LEWIS HANKE (ed.), *History of Latin American Civilization* (1967). An excellent history covering the same period is HUBERT HERRING, *A History of Latin America from the Beginnings to the Present*, 3rd ed. rev. (1968). A general textbook, covering various aspects of culture and society, from colonial times to our days, is HELEN M. BAILEY and ABRAHAM P. NASATIR, *Latin America: The Development of Its Civilization*, 2nd ed. (1968). An integrated overview of the social, economic, and political modernization of the region is given in GINO GERMANI, "Stages of Modernization in Latin America," in *Studies*

in Comparative International Development (1969–70). On the varieties of Spanish, see PEDRO HENRIQUEZ URENA, *Sobre el Problema del Andalucismo Dialectal de América* (1932).

Resource development and commerce: J.P. COLE (*op. cit.*), is a comprehensive study of the population, physical background, major historical events, institutions and main aspects of economic development, together with more detailed analyses for individual countries. See also the UNITED NATIONS, ECONOMIC COMMISSION FOR LATIN AMERICA (UN-ECLA), *Economic Survey of Latin America* (annual), a report covering major developments in the economic field and economic trends for each country—some issues also deal with specific aspects such as energy, transportation, or small industry; and the *Economic Bulletin for Latin America* (published twice yearly), which includes special articles as well as informative and methodological notes and updates the statistical data in the annual surveys. ALCIRA LEISERSON, *Notes on the Process of Industrialization in Argentina, Chile, and Peru* (1966), focuses primarily on the main political characteristics of these countries as they undergo industrialization, and analyzes the relationship between populism and the rise of the new technocratic elites. The report of the INTER-AMERICAN DEVELOPMENT BANK, *The Process for Industrialization in Latin America* (1969), refers particularly to regional integration, technological innovation, external markets for regional manufactures, external financing, and the role of government and private enterprise in industrial development. The *Report of the Latin American Symposium on Industrial Development* (UN, 1966), contains an evaluation of the process of industrialization, its past evolution, present characteristics, and future prospects. The PAN AMERICAN UNION, *Industrialization in Latin America: Priority Problems* (1967), analyzes the relationship between industrialization, import substitution, and employment. LEHMAN B. FLETCHER and WILLIAM C. MERRILL, *Latin American Agricultural Development and Policies* (1968), contains papers dealing with policies for economic growth in the agricultural sectors of eight countries. JOHN A. HOPKINS, *The Latin American Farmer* (USDA, 1969), is a thorough study of farm conditions in Latin America. ROBERT T. BROWN, *Transport and the Economic Integration of South America* (1966), is a very complete study of the various modes of transportation and their relation to economic integration; see also CHARLES J. STOKES, *Transportation and Economic Development in Latin America* (1968); PAN AMERICAN UNION, *General Problems of Transportation in Latin America* (1962); and the UN-ECLA, *Transport in Latin America*, 3 vol. (1965). DONALD W. BAERRESEN, MARTIN CARNOY, and JOSEPH GRUNWALD, *Latin America Trade Patterns* (1965), is a general historical survey of trade between the Latin American countries and a detailed analysis of more recent developments in trade among LAFTA members. Also of interest are UN-ECLA, *Regional Integration and the Trade of Latin America* (1968), a study of various possible forms of economic integration and of obstacles to achieving their objective; and its *Statistical Bulletin for Latin America*, a monthly publication giving statistical information on selected economic indicators; PAN AMERICAN UNION, *Latin America's Foreign Trade: Problems and Policies* (1966); the UNITED NATIONS, FOOD AND AGRICULTURAL ORGANIZATION, *Production Yearbook* and *Trade Yearbook* (both annual); and the INTERNATIONAL MONETARY FUND and INTERNATIONAL BANK FOR RECONSTRUCTION AND DEVELOPMENT, *Direction of Trade* (monthly).

Demography: An important series of studies of demographic trends may be found in the special issue of the *Milbank Memorial Fund Quarterly*, vol. 43, no. 4, pt. 2 (1965), on *Components of Population Change in Latin America.* A short general presentation by a major Latin American demographer is the article by CARMEN A. MIRO, "The Population of Latin America," in CLAUDIO VELIZ (ed.), *Latin America and the Caribbean* (1968). Another general article is KINGSLEY DAVIS, "The Place of Latin America in World Demographic History," *Milbank* (*op. cit.*), 42:19–47 (1964). Many demographic studies appear in the *Economic Bulletin for Latin America*, a journal published in English (twice yearly) by the United Nations Economic Commission for Latin America, Santiago de Chile. On fertility, see the special issue of the *Milbank* (*op. cit.*), vol. 46, no. 3 (1968), on *Current Research on Fertility and Family Planning in Latin America;* and O. ANDREW COLLVER, *Birth Rates in Latin America* (1965). On urbanization, internal migration, and living conditions in the cities, see PHILIP M. HAUSER (ed.), *Urbanization in Latin America* (1961); and GLENN H. BEYER (ed.), *The Urban Explosion in Latin America* (1967). Current demographic statistics appear in the *Statistical Abstract of Latin America* (annual).

(J.P.D./G.Ge./H.F.A.)

South American Indian Languages

South American Indian languages once covered and today still partially cover all of South America, the Antilles, and Central America to the south of a line from the Gulf of Honduras to the Nicoya Peninsula in Costa Rica. Estimates of the number of speakers in that area in pre-Columbian times vary from 10,000,000 to 20,000,000. At the start of the 1970s there were approximately 11,200,000, more than half of them in the central Andean area. Language lists include around 1,500 languages, and figures over 2,000 have been suggested. For the most part, the larger estimate refers to tribal units whose linguistic differentiation cannot be determined. Because of extinct tribes with unrecorded languages, the number of languages formerly spoken is impossible to assess. Only between 550 and 600 languages (about 120 now extinct) are attested by linguistic materials. Fragmentary knowledge hinders the distinction between language and dialect and thus renders the number of languages indeterminate.

Number of languages and speakers

Because the South American Indians originally came from North America, the problem of their linguistic origin involves tracing genetic affiliations with North American groups. To date only Uru-Chipaya, a language in Bolivia, is surely relatable to a Macro-Mayan phylum of North and Meso-America. Hypotheses about the probable centre of dispersion of language groups within South America have been advanced for stocks like Arawakan and Tupian, based on the principle (considered questionable by some) that the area in which there is the greatest variety of dialects and languages was probably the centre from which the language groups dispersed at one time; but the regions in question seem to be refugee regions, to which certain speakers fled, rather than dispersion centres.

South America is one of the most linguistically differentiated areas of the world. Various scholars hold the plausible view that all American Indian languages are ultimately related. The great diversification in South America, in comparison with the situation of North America, can be attributed to the greater period of time that has elapsed since the South American groups lost contact among themselves. The narrow bridge that allows access to South America (*i.e.*, the Isthmus of Panama) acted as a filter so that many intermediate links disappeared and many groups entered the southern part of the continent already linguistically differentiated.

See also NORTH AMERICAN INDIAN LANGUAGES and MESO-AMERICAN INDIAN LANGUAGES.

Investigation and scholarship. The first grammar of a South American Indian language (Quechua) appeared in 1560. Missionaries displayed intense activity in writing grammars, dictionaries, and catechisms during the 17th century and the first half of the 18th. Data were also provided by chronicles and official reports. Information for this period was summarized in Lorenzo Hervás y Panduro's *Idea dell' universo* (1778–87) and in Johann Christoph Adelung and Johann Severin Vater's *Mithridates* (1806–17). Subsequently, most firsthand information was gathered by ethnographers in the first quarter of the 20th century. In spite of the magnitude and fundamental character of the numerous contributions of this period, their technical quality was below the level of work in other parts of the world. Since 1940 there has been a marked increase in the recording and historical study of languages, carried out chiefly by missionaries with linguistic training, but there are still many gaps in knowledge at the basic descriptive level, and few languages have been thoroughly described. Thus, classificatory as well as historical, areal, and typological research has been hindered. Descriptive study is made difficult by a shortage of linguists, the rapid extinction of languages, and the remote location of those tongues needing urgent study. Interest in these languages is justified in that their study yields basic cultural information on the area, in addition to linguistic data, and aids in obtaining historical and prehistorical knowledge. The South American Indian languages are also worth studying as a means of integrating the groups that speak them into national life.

Research on the languages

Classification of the South American Indian languages. Although classifications based on geographical criteria or on common cultural areas or types have been made, these are not really linguistic methods. There is usually a congruence between a language, territorial continuity, and culture, but this correlation becomes more and more random at the level of the linguistic family and beyond. Certain language families are broadly coincident with large culture areas—*e.g.*, Cariban and Tupian with the tropical forest area—but the correlation becomes imperfect with more precise cultural divisions—*e.g.*, there are Tupian languages like Guayakí and Sirionó whose speakers belong to a very different culture type. Conversely, a single culture area like the eastern flank of the Andes (the Montaña region) includes several unrelated language families. There is also a correlation between isolated languages, or small families, and marginal regions, but Quechumaran (Kechumaran), for instance, not a big family by its internal composition, occupies the most prominent place culturally.

Most of the classification in South America has been based on inspection of vocabularies and on structural similarities. Although the determination of genetic relationship depends basically on coincidences that cannot be accounted for by chance or borrowing, no clear criteria have been applied in most cases. As for subgroupings within each genetic group, determined by dialect study, the comparative method, or glottochronology (also called lexicostatistics, a method for estimating the approximate date when two or more languages separated from a common parent language, using statistics to compare similarities and differences in vocabulary), very little work has been done. Consequently, the difference between a dialect and language on the one hand, and a family (composed of languages) and stock (composed of families or of very differentiated languages) on the other, can be determined only approximately at present. Even genetic groupings recognized long ago (Arawakan or Macro-Chibchan) are probably more differentiated internally than others that have been questioned or that have passed undetected.

Extinct languages present special problems because of poor, unverifiable recording, often requiring philological interpretation. For some there is no linguistic material whatsoever; if references to them seem reliable and unequivocal, an investigator can only hope to establish their identity as distinct languages, unintelligible to neighbouring groups. The label "unclassified," sometimes applied to these languages, is misleading: they are unclassifiable languages.

Problem of language names

Great anarchy reigns in the names of languages and language families; in part, this reflects different orthographic conventions of European languages, but it also results from the lack of standardized nomenclature. Different authors choose different component languages to name a given family or make a different choice in the various names designating the same language or dialect. This multiplicity originates in designations bestowed by Europeans because of certain characteristics of the group (*e.g.*, Coroado, Portuguese "tonsured" or "crowned"), in names given to a group by other Indian groups (*e.g.*, Puelche, "people from the east," given by Araucanians to various groups in Argentina), and in self-designations of groups (*e.g.*, Carib, which, as usual, means "people" and is not the name of the language). Particularly confusing are generic Indian terms like Tapuya, a Tupí word meaning enemy, or Chuncho, an Andean designation for many groups on the eastern slopes; terms like these explain why different languages have the same name. In general (but not always), language names ending in *-an* indicate a family or grouping larger than an individual language; *e.g.*, Guahiboan (Guahiban) is a family that includes the Guahibo language, and Tupian subsumes Tupí-Guaraní.

There have been many linguistic classifications for this area. The first general and well-grounded one was that by U.S. anthropologist Daniel Brinton (1891), based on grammatical criteria and a restricted word list, in which about 73 families are recognized. In 1913 Alexander Chamberlain, an anthropologist, published a new classification in the United States, which remained standard for several years, with no discussion as to its basis. The classification (1924) of the French anthropologist and ethnologist Paul Rivet, which was supported by his numerous previous detailed studies and contained a wealth of information, superseded all previous classifications. It included 77 families and was based on similarity of vocabulary items. Čestmír Loukotka, a Czech language specialist, contributed two classifications (1935, 1944) on the same lines as Rivet but with an increased number of families (94 and 114, respectively), the larger number resulting from newly discovered languages and from Loukotka's splitting of several of Rivet's families. Loukotka used a diagnostic list of 45 words and distingushed "mixed" languages (those having one-fifth of the items from another family) and "pure" languages (those that might have "intrusions" or "traces" from another family but totalling fewer than one-fifth of the items, if any). Rivet and Loukotka contributed jointly another classification (1952) listing 108 language families that was based chiefly upon Loukotka's 1944 classification. Important work on a regional scale has also been done, and critical and summarizing surveys have appeared.

Current classifications

Current classifications are by Loukotka (1968); a U.S. linguist, Joseph Greenberg (1956); and another U.S. linguist, Morris Swadesh (1964). That of Loukotka, based fundamentally on the same principles as his previous classifications, and recognizing 117 families, is, in spite of its unsophisticated method, fundamental for the information it contains. Those of Greenberg and Swadesh, both based upon restricted comparison of vocabulary items but according to much more refined criteria, agree in considering all languages ultimately related and in having four major groups, but they differ greatly in major and minor groupings. Greenberg used short lexical lists, and no evidence has been published in support of his classification. He divided the four major groups into 13 and these, in turn, into 21 subgroups. Swadesh based his classification upon lists of 100 basic vocabulary items and made groupings according to his glottochronological theory (see above). His four groups (interrelated among themselves and with groups in North America) are subdivided into 62 subgroups, thus, in fact, coming closer to more conservative classifications. The major groups of these two classifications are not comparable to those recognized for North America, because they are on a more remote level of relationship. In most cases the lowest components are stocks or even more distantly related groups. It is certain that far more embracing groups than those accepted by Loukotka can be recognized—and in some cases this has already been done—and that Greenberg's and Swadesh's classifications point to many likely relationships; but they seem to share a basic defect, namely, that the degree of relationship within each group is very disparate, not providing a true taxonomy and not giving in each case the most closely related groups. On the other hand, their approach is more appropriate to the situation in South America than a method that would restrict relationships to a level that can be handled by the comparative method.

At present, a true classification of South American languages is not feasible, even at the family level, because, as noted above, neither the levels of dialect and language nor of family and stock have been surely determined. Beyond that level, it can only be indicated that a definite or possible relationship exists. In the accompanying chart —beyond the language level—recognized groups are therefore at various and undetermined levels of relationship. Possible further relationships are cross-referenced. Of the 82 groups included, almost half are isolated languages, 25 are extinct, and at least 10 more are on the verge of extinction. The most important groups are Macro-Chibchan, Arawakan, Cariban, Tupian, Macro-Ge, Quechumaran, Tucanoan, and Macro-Pano-Tacanan.

Macro-Chibchan. Macro-Chibchan languages, which form the linguistic bridge between South and Central America, are spoken from Nicaragua to Ecuador. Spread compactly in Central America and in western Colombia and Ecuador, they include approximately 40 languages spoken by at least 100,000 speakers. The group is proba-

South American Indian Language Groups*

language	location
1. Alacalufan (47): Aksanas or Kaueskar, Alacaluf, Caucau or Caucawe	Chile
2. Andoque (9)	Colombia
3. Araucanian or Mapuche (32, 37)	Chile, Argentina
4. Arawakan (43, 44)	
A. Amuesha	Peru
B. Apolista or Lapachu†	Bolivia
C. Arauan: Araua†, Curina, Madihá, Paumarí, Yamadí	Brazil
D. Chamicuro	Peru
E. Maipurean	
1. Achagua, Amarizana, Capite Minanei, Cauyarí, Guarú, Guayupé, Maipure†, Piapoco, Resígaro, Tariana, Warakena†, Yucuna	Colombia
Anauya, Baré, Curipaco, Guinau, Mandawaca, Paraujano [Parauhano]	Venezuela
Araikú†, Aruan†, Cariay†, Carútana, Catapolítani, Cawishana†, Hohodene, Manao†, Mapanai, Marawá†, Mariate†, Maulieni, Moriwene, Pasé†, Siusí, Wainumá†, Wiriná, Yabaana, Yumana	Brazil
Arawak or Lokono	Guyana, Fr. Guiana
Goajiro	Colombia, Venezuela
Adzaneni, Ipeca	Brazil, Colombia
Island Carib	Dominica
2. Atorai, Mapidian	Guyana
Wapishana	Guyana, Brazil
3. Baníva, Yavitero	Venezuela
4. Bauré, Mojo or Ignaciano, Muchojeone, Pauna, Paicone	Bolivia
5. Campa, Machiguenga, Piro	Peru
Canamari, Chontaquiro [Chontakiro], Cuniba†, Cushineri†, Ipuriná	Brazil
Inapari	Bolivia
6. Caripuna, Marawan	Brazil
7. Chané†	Argentina
Guaná	Paraguay, Brazil
Quiniquiano, Tereno	Brazil
8. Paressí	Brazil
Sarave	Bolivia
F. Taino†	Antilles
G. Morique [Morike] or Mayoruna	Peru
5. Atacama or Cunza or Lincan Antai†	Argentina, Chile
6. Auake or Arutani	Brazil, Venezuela
7. Auishiri [Awishira]	Peru
8. Baenan	Brazil
9. Bora-Huitotoan (2)	
A. Boran: Bora or Miraña, Emejeite, Muinane	Colombia
B. Huitotoan [Witotoan]:	
Andoquero, Huitoto [Witoto]	Colombia
Coeruna†	Brazil
Ocaina, Orejone†, Nonuya or Achote [Achiote]	Peru
10. Canichana	Bolivia
11. Capixana or Canoe	Brazil
12. Cariban (70)	
A. 1. Acawai (Arecuna, Camaracoto, Ingarico) Waica	Venezuela
Taulipang or Ipuricoto or Pemon	Brazil, Venezuela
2. Apalai, Aracajú, Upurui	Brazil
Oyana	Surinam
Rucuyen	Fr. Guiana
3. Apiacá or Apingi†, Arara†, Pariri†	Brazil
4. Atroahi (Yauaperi), Quirixana, Waimiri [Waimiry]	Brazil
5. Bakairi† [Bacairi], Nahukua (Nahucua), Yaruma†	Brazil
6. Bonari†, Hishkariana, Parucoto, Waiboi, Waiwai	Brazil
7. Cachuene or Caxuiana, Mutuan†, Pauxí†, Saluma, Wayewé	Brazil
Chiquena [Chikene] or Shikiana	Brazil, Guyana
8. Carare, Opone	Colombia
9. Caribe or Calina or Galibi	Antilles, Guianas
Carif	Br. Honduras, Honduras
10. Carijona (Guaque [Guake], Umaua)	Colombia
11. Colima†, Muzo, Pijiao	Colombia
12. Cumanagoto† (Chayma†, Tamanaco), Tivericoto†	Venezuela
13. Keseruna, Macushí, Paraviyana†, Purukoto†, Zapara	Brazil
14. Mariquitare or Decuana	Brazil, Venezuela
Yecuana or Mayongong (Cunuana, Ihuruana)	Venezuela
15. Mapoyo or Nepoyo, Yauarana	Venezuela
16. Motilon	Colombia, Venezuela
17. Palmela† [Palmella]	Brazil
18. Panare	Venezuela
19. Patagon	Peru
20. Pawishana	Brazil
21. Pianocoto	Brazil
Tliometesen	Surinam
Trio (Ocomayana, Urucuyena [Urucuena], Wama)	Surinam, Brazil
22. Pimenteira†	Brazil
23. Yao†	Fr. Guiana, Trinidad
B. Chocó: Chamí, Sambú [Sambo], Waunana	Colombia
13. Cariri or Kiriri	Brazil
14. Catacao†: Catacao†, Colan	Peru
15. Catembri or Mirandela†	Brazil
16. Catuquina [Catukina]: Bendiapa (Canamari, Parawa), Catuquina [Catukina] or Wiridyapa, Catauxi, Tucundiapa [Tucundyapa]	Brazil
17. Cayuvava (70)	Bolivia
18. Chapacura: Chapacura or Huachi, Itene or Moré, Itoreauhip, Nape, Pacahanovo, Quitemo†	Bolivia
Cumaná (Abitana), Torá, Urupá (Yarú [Jarú]), Wañám [Wanyam]	Brazil
19. Chiquito or Tarapecosi (42)	Bolivia
20. Cholonan: Cholona or Seeptsá†, Híbito†	Peru
21. Cofán	Colombia, Ecuador
22. Culli or Ilinga†	Peru
23. Erikbaktsa or Canoeiro	Brazil
24. Gamela†	Brazil
25. Gorgotoqui	Brazil
26. Guahiboan (4): Chiricoa, Guahibo	Venezuela, Colombia
Churuya, Guayabero	Venezuela
27. Guaycurú-Charruan	
A. Charruan†: Chaná†	Uruguay
Charrúa† (Guenoa or Minuan†)	Uruguay, Argentina, Brazil
B. Guaycuruan: Abipón or Callaga†	Argentina, Paraguay
Caduveo or Mbayá or Guaycurú	Brazil, Argentina, Paraguay
Guachí†	Brazil
Mocoví	Argentina
Payaguá or Lengua	Paraguay
Toba-Pilagá	Argentina, Bolivia
28. Guamo†	Venezuela
29. Guató	Bolivia, Brazil
30. Guennaken or Gununa-Kune or Puelche†	Argentina
31. Huarian: Huari or Corumbiara, Masaca or Aicana	Brazil
32. Huarpean† (37): Allentiac†, Millcayac†	Argentina
33. Iranxe (4)	Brazil
34. Jirajaran: Ayomán, Gayón†, Jirajara	Venezuela
35. Kaliana or Sape	Venezuela
36. Koaia	Brazil
37. Jebero-Jivaroan	
A. Jeberoan or Cahuapanan: Cahuapana or Chuncho [Concho], Chayavita, Jebero [Chébero], Miquirá, Yamorai	Peru
B. Jívaroan: Aguaruna	Peru
Jívaro or Shuara (Achual, Huambisa [Wambisa])	Ecuador
38. Kukura†	Brazil
39. Leco	Bolivia
40. Lulean (48)	
A. Lule or Tonocoté†	
B. Vilela or Chunupí (Atalalá, Ocolé, Uacambabelé)†	Argentina
41. Macro-Chibchan	
A. Chibchan	
1. Abiseta or Orosi or Tucurrique [Tucurrike], Boruca or Turucaca†, Bribrí or Lari, Cabecar, Chiripó, Estrella†, Terraba or Brurán†, Tirub or Rayado†	Costa Rica
2. Andaquí	Colombia
3. Atanque or Busintana, Bintucua or Ijca, Cágaba or Koghi, Guamaca or Arsario, Tairona or Teyuna†	Colombia
4. Bairira or Cunaguasaya, Motilón or Dobocubí	Colombia
Mape	Venezuela
5. Betoi†	Colombia
6. Cara or Imbaya†, Cayapa or Nigua, Colorado or Campaz or Colima or Satxila, Pasto†	Ecuador
Cuaiquier, Muellama†, Telembi†	Colombia
7. Catio†, Nutabé†	Colombia
8. Chibcha or Muisca†, Tunebo	Colombia
9. Chimila, Malibú†	Colombia
10. Changuena†, Chumula† (Gualaca†)	Panama
11. Coconuco, Guambiana or Silviano, Moguex, Totoró	Colombia
12. Corobisi, Guatuso, Guetar, Suerre or Camachi	Costa Rica
13. Cuna	Panama
Cueva	Colombia
14. Guaymí, Move (Penonemeño†)	Panama
15. Panzaleo or Quito†	Ecuador
Paez	Colombia
16. Rama	Nicaragua
17. Sebondoy or Kamsá or Coche or Mocoa, Quillasinga†	Colombia
B. 1. Esmeralda or Atacame†	Ecuador
2. Yaruro	Venezuela
C. Itonama	Bolivia
D. Paya	Honduras
E. Sumo-Mosquito-Matagalpa	
1. Matagalpa† (Cacaopera)	Ecuador, Surinam
Jinotega or Chingo	Nicaragua
2. Mosquito [Miskito]	Honduras, Nicaragua
Ulua, Sumo	Nicaragua
F. Waican or Yanoaman: Karimé [Carimé], Pakidai-Surara, Paucosa	Brazil
Sanemá or Samatari (Pubmatari)	Venezuela
Shamatari, Shiriana or Casapare, Waica [Waika]	Brazil, Venezuela
G. Warao [Warrau] or Guarauno	Venezuela
42. Macro-Ge (70)	
A. Bororoan	
1. Bororo	Brazil
2. Otuké	Bolivia
B. Botocudo or Aymoré†	
C. Fulnió†	Brazil
D. Ge	
1. Akroá†, Xakriaba† [Shacriaba], Xavante [Shavante], Xerente [Sherenté]	Brazil
2. Apinayé-Kayapó, Eastern Timbira, Suyá	Brazil
3. Caingang [Kaingang], Xokleng [Shocleng]	Brazil
4. Southern Kayapó	
E. Jeikó [Jeico]†	Brazil
F. Kamakán†	Brazil
G. Karajá	Brazil
H. Kapoxo† (Kumanaxo†), Malalí†, Maxakalí†, Monoxo†, Patashó†	Brazil
I. Ofayé or Opayé-Shavante†	Brazil
J. Purí-Coroado†: Coroado†, Coropó†, Purí†	
43. Macro-Mayan (4): Uru-Chipaya (Uru, Chipaya)	Bolivia
44. Macro-Pano-Tacanan (4)	
A. Chon: Haush or Manekenken†, Ona or Shelknam†, Tehuelche, Teushen or Tehuesh†	Argentina

South American Indian Language Groups* (continued)

language	location
B. Mosetene: Chimane, Mosetene† [Moseten]	Bolivia
C. Pano-Tacanan	
1. Panoan: Amahuaca [Amawaca], Cashinahua [Cashinawa]	Brazil, Peru
Capanahua [Capanawa], Cashibo, Conibo-Shipibo (Chama, Setebo, Sensi), Marinahua [Marinawa], Marobo, Nocamán, Pano or Pánobo	Peru
Culino (Curina), Jaminahua, Mayoruna or Maruba, Nastanahua, Nixinahua, Parannahua, Poyanahua, Remo, Shaminahua, Tushinahua [Tushinawa], Waninahua or Catoquino, Yahuanahua (Yawanawa), Yumanahua	Brazil
Arazaire [Arasaire], Atsahuaca [Atsawaca] or Chaspa, Yamiaca or Haauñeiri	Peru
Caripuna	Brazil
Chácobo, Pacahuara [Pacawara]	Bolivia
2. Tacanan: Arasa, Cavineña, Chama or Esseejja, Guarizo, Huarayo (Tianinagua), Mabenaro, Maropa or Reyesano, Sapiboca [Sapiboka], Tacana (Araona, Toromona)	Bolivia
D. Yuracare	Bolivia
45. Makú	Venezuela, Brazil
46. Mascoy [Mascoi] or Lengua: Angaité (Sanapá), Kashiká, or Guaná, Lengua or Enslet or Cocoloth (Mascoy)	Paraguay
47. Mataco-Maccá [Macá]	
1. Ashluslay or Chulupí	Paraguay
Chorotí or Solote or Yofuaha	Paraguay, Argentina
Choropí (Suhin, Sotirai), Mataco or Mataguayo (Guisnay, Nocten, Vejoz)	Argentina
2. Enimagá or Cochaboth†	Paraguay
3. Maccá [Macá] or Towothli	Paraguay
48. Movima (27)	Bolivia
49. Munichi [Muniche]	Peru
50. Mura-Matanawí	
A. Bohurá, Mura, Pirahá	Brazil
B. Matanawí†	Brazil
51. Murato or Candoshi or Shapra	Peru
52. Nambikwara [Nambicuara]: Central, Eastern Nambikwara	Brazil
53. Omurano or Mayna†	Peru
54. Otomaco-Taparita†: Otomaco†, Taparita†	Venezuela
55. Pankarurú [Pancararú]	Brazil
56. Puinave-Maku: Makú, Marahan, Querari	Brazil
Puinave	Colombia
57. Puquina†: Pohena or Callahuaya†, Puquina	Bolivia
58. Quechumaran	
A. Aymaran: Aymara, Cauqui or Jaqaru	Bolivia, Peru
B. Quechuan: Almaguero, Inga	Colombia
Ancash, Ayacucho, Cajamarca, Chasutino, Huánuco, Junín, Lamano, Lima, Mayna, Pasco, Ucayali	Peru
Catamarca-La Rioja, Santiago del Estero	Argentina
Cuzco-Bolivian	Peru, Bolivia
Ecuadorian, Quijos, Tena	Ecuador
Tuichi	Bolivia
59. Sabelan: Sabela or Auca or Huarani, Tiwituey	Peru
60. Sáliva-Piaroan: Maco (Macu), Piaroa	Venezuela
Sáliva	Colombia
61. Sec or Sechura or Tallán or Atalán†	Peru
62. Simacu or Itucale or Arucuaya or Urariña	Peru
63. Tarariu (Tarairiu) or Ochukuyana† [Ochukayana]	Brazil
64. Taruma†	Brazil
65. Tikuna [Ticuna] or Tukuna [Tucuna]	Brazil, Colombia
66. Timote†	Venezuela
67. Tiniguan: Pamigua, Tinigua†	Colombia
68. Trumai	Brazil
69. Tucanoan	
A. Western: Amaguaje†, Coto, Piojé	Peru
Coreguaje [Correguaje], Dätuana, Icaguaje†, Macaguaje, Macuna, Särа, (Ömöa, Buhagana), Siona [Sioni], Tama, Tanimuca, Uantia, Yahuna, Yupua	Colombia
Coretu	Brazil
Secoya	Ecuador
B. Eastern: Bara, Erulia (Paneroa, Tsölöa), Karapana [Carapaná], (Möchda), Pamöá, Siana or Chiranga, Tatapuyo, Waiana, Yarutí or Patsoca	Colombia
Desana, Tuyuka (Tuyuca), Wanana (Waikina)	Colombia, Brazil
Kubeo (Cubeo), Tucano	Brazil
70. Tupian	
A. 1. (Tupí-Guaraní): Apiaká† [Apiacá], Awetí, Canoeiro, Kamayurá [Camayura], Kawaíb [Cawahíb] (Pawate, Parintintin, Wirafed) Kayabí (Cayabí), Shetà†, Takuñapé† [Tacunyapé], Tapirapé, Tenetehára (Anambé, Guayayara [Guajajára], Manajé, Tembé, Turiwara†, Urubú), Tupí-Guaraní (Tupí-nambá), Neengatú	Brazil
Oyampí-Emerillon†	Brazil, Fr. Guiana
Pauserna	Bolivia, Brazil
Guaraní	Argentina, Brazil, Paraguay
Kaiwá	Brazil, Paraguay
Chiriguano, Guarayú	Bolivia
Tapieté, Chané	Paraguay
2. Cocama	Peru
Omagua	Brazil, Peru
3. Guayakí	Paraguay
4. Mawé	Brazil
5. Sirionó	Bolivia
B. Arara, Ramarama (Itogapid), Urukú, Urumí	Brazil
C. Arikem†, Kabishiana†, Karitiana†	Brazil
D. Arué, Digüt, Mondé†	Brazil
E. Guaratégaya† (Amniapé†, Kanoe, Mekens), Kepkiriwat†, Makurap†, Tuparí, Wayoró (Apichum)	Brazil
F. Kuruaya† (Curuaya), Mundurukú	Brazil
G. Manitsawá†, Shipaya†, Yuruna	Brazil
H. Puruborá	Brazil
71. Tushá	Brazil
72. Tuyoneiri or Arasairi or Huachipairi	Peru
73. Uman or Huamoi	Brazil
74. Xukurú or Ichikile	Brazil
75. Yabutí†: Aricapú†, Mashubi†, Yabutí or Quipiu†	Brazil
76. Yagua: Masamae, Peba or Nijamo, Yagua or Mishara, Yameo or Camuchivo	Peru
77. Yámana† or Yaghan (41)	Chile
78. Yunca or Chimú or Muchic or Chincha†: Puruhá-Cañari† Yunca†	Ecuador Peru
79. Yurí†	Brazil, Colombia
80. Yurimangui†	Colombia
81. Zamuco: Chamacoco (Ebidoso, Tumrahá), Zamuco (Ayoré, Moro)	Paraguay
82. Záparo: Arabela, Iquito (Cahuarano), Shimigae† (Semigae)†, Andoa, Záparo†	Peru

*Only languages attested linguistically are included. Extinct languages are marked with †. A number in parentheses after the name of a group indicates a possible relationship with the group identified by that number. Languages are separated by commas, names in parentheses are of dialects, and names in brackets are alternative spellings. Except for Arawakan, Macro-Ge, and Tupian, most groupings are geographical, but those identified by capital letters represent in general markedly differentiated groups. Spelling follows the most common usage for each language or group, thus it is not consistent. Equivalent spellings: $b = v$; $g = j = y$; $gu = hu = u = w$; $i = y$; h = (nothing); $h = j$; $k = c$ (before *a, o, u*); $k = qu$ (before *i, e*); $sh = x = ch$; $s = z$; $\tilde{n} = nh = ny$; $x = j$. Names are arranged alphabetically within each subdivision.

bly more differentiated than a stock, languages not belonging to Chibchan being strongly differentiated. In the Colombian Andes a now extinct Chibchan language was the language of the highly developed Muisca culture. Important present-day languages include Guaymí (about 10,000 speakers) and Move (about 15,000) in Panama, Cuna (21,000) and Páez (20,000) in Colombia, and Cayapa, or Colorado (4,000), in Ecuador. A connection with Cariban has been suggested, and it is possible that such a relationship could be found through Warao (Warrau) and Waican (Waikan) on the one hand and through Chocó (Cariban) on the other.

Arawakan. Arawakan languages formerly extended from the peninsula of Florida in North America to the present-day Paraguay–Argentina border, and from the foothills of the Andes eastward to the Atlantic Ocean. More than 55 languages are attested, many still spoken. Around 40 groups still speak Arawakan languages in Brazil, and others are found in Peru, Colombia, Venezuela, Guyana, French Guiana, and Surinam. Taino predominated in the Antilles and was the first language to be encountered by Europeans; although it rapidly became extinct, it left many borrowings. As did most languages of the tropical forest, the Arawakan languages receded with the influx of Spanish and Portuguese, mainly through group extinction; thus, 14 groups became extinct in Brazil between 1900 and 1957. Important languages still spoken are Goajiro (45,000 speakers) in Colombia, Campa (25,000) and Machiguenga (6,000) in Peru, and Mojo (3,000) and Bauré (3,500) in Bolivia. Although most Arawakan languages have been recognized as such for a long time, they are greatly differentiated. They are most probably related to both the Macro-Pano-Tacanan and Macro-Mayan language groups.

Extent of the Arawakan languages

Cariban. Cariban languages, numbering approximately 50, were spoken chiefly north of the Amazon but had outposts as far as the Mato Grosso in Brazil. The group has undergone drastic decline, and only about 40,000 people speak Cariban languages today, mostly in Brazil and Colombia; they have disappeared from the Antilles and have been much reduced in Venezuela and the Guianas. The most important group today—Chocó in

western Colombia—is distantly related to the rest of the stock. Other languages are Carib in Surinam, Trio in Surinam and Brazil, and Waiwai, Taulipang, and Makushí (Macusí) in Brazil. A relationship with Tupian seems certain.

Tupian. With the exception of Emerillon and Oyampi of French Guiana and northeastern Brazil, Tupian languages were spoken south of the Amazon, from the Andes to the Atlantic Ocean and down to the Río de la Plata. There are around 50 attested languages related on the stock level and subdivided into eight families. Tupinambá, the language spoken along the Atlantic coast at the time of discovery, became important in a modified form as a lingua franca, and the closely related Guaraní became the national language in Paraguay, being one of the few Indian languages that does not seem to yield under the influence of Spanish or Portuguese. At the time of discovery, Tupí-Guaraní tribes were moving everywhere south of the Amazon, subjugating other tribes, some of which adopted Tupí-Guaraní. Both Tupí and Guaraní are among the languages that have exerted a great influence on Portuguese and Spanish. Tupí groups have declined markedly, 26 groups becoming extinct in Brazil between 1900 and 1957, and at least 14 languages disappearing during the same period. The westernmost language, Cocama in Peru, is still spoken by about 10,000 speakers, and Chiriguano in Bolivia has about 15,000 speakers. Other languages have a much smaller number of speakers; there are 14,000 speakers for the 26 surviving groups in Brazil. The total number of Indian speakers of Tupian languages is approximately 50,000, but there are also more than 1,000,000 culturally non-Indian speakers of Guaraní in Paraguay. Besides the connection with Cariban, further relationships possibly exist with Macro-Ge, various small families like Zamuco and Mataco-Maccá and isolated languages like Cayuvava.

The Tupí and Guaraní languages

Macro-Ge. Macro-Ge is geographically the most compactly distributed of the big South American language families. Ge proper extends uninterruptedly through inland eastern Brazil almost as far as the Uruguayan border. There are about ten Ge languages with a total of 10,000 speakers. Most of the other families, now extinct, were located closer to the Atlantic coast, from where they probably were displaced by Tupian expansion. The Bororan family is represented by Bororo in Brazil and by the Otuké language in Bolivia. It seems likely that Macro-Ge has its closest relationship with Tupian.

Quechumaran. Quechumaran, composed of the Quechuan and Aymaran families, is the stock with the largest number of speakers—5,000,000 for Quechuan and 1,000,000 for Aymaran—and is found mainly in the Andean highlands extending from southern Colombia to northern Argentina. This group has also resisted displacement by Spanish, in addition to having gained in speakers from the time of the Incas to the present as several other groups adopted Quechuan languages. Cuzco-Bolivian Quechua is spoken by well over 1,000,000 speakers, and there are around seven Quechuan languages in Peru with more than 100,000 speakers each. Although most Quechuan languages have been influenced by Spanish, Quechuan in turn is the group that has exerted the most pervasive influence on Spanish. No convincing further genetic relationship has been yet proposed.

Tucanoan. Tucanoan, spoken in two compact areas in the western Amazon region (Brazil, Colombia, and Peru), includes about 30 languages with over 10,000 speakers. One of the languages is a lingua franca in the region.

Macro-Pano-Tacanan. Macro-Pano-Tacanan, a group more distantly related than a stock, includes about 30 languages, many of them still spoken. The languages are located in two widely separated regions: lowland eastern Peru and adjoining parts of Brazil and lowland western Bolivia on the one hand, and southern Patagonia and Tierra del Fuego on the other. In the latter region the languages are practically extinct.

Other language groups

By number of component languages, or by number of speakers, or by territorial extension, the other language groups are not as significant as those just listed. Most of these small families and isolated languages are located in the lowlands, which form an arch centred on the Amazon from Venezuela to Bolivia and include the bordering parts of Brazil.

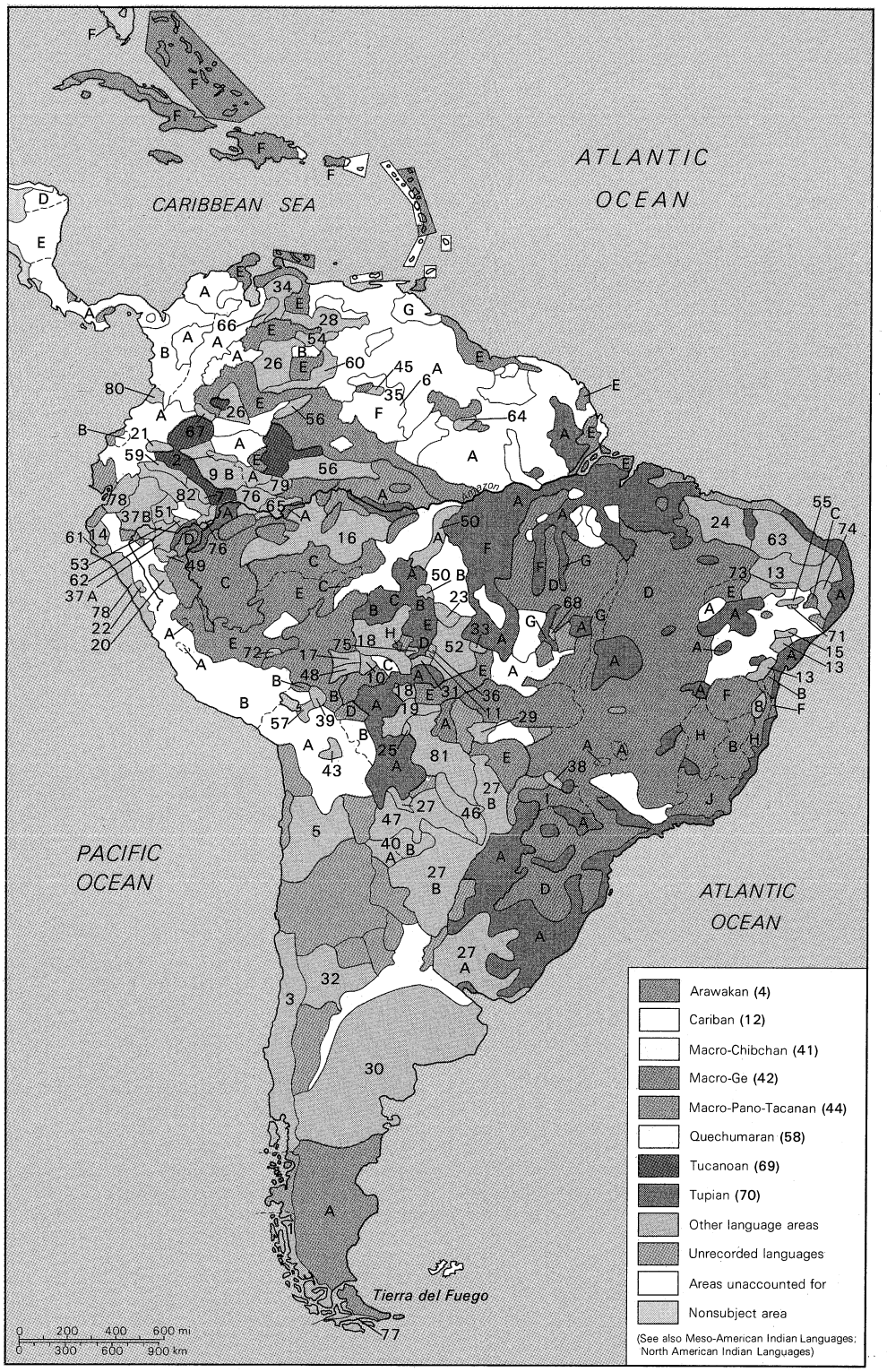
Lingua francas and cultural tongues. Lingua francas as well as situations of bilingualism arose mainly under conditions furthered or created by Europeans, although a case like that of the Tucano language, which is used as a lingua franca in the Río Vaupés area among an Indian population belonging to some 20 different linguistic groups, may be independent of those conditions. Quechua, originally spoken in small areas around Cuzco and in central Peru, expanded much under Inca rule, coexisting with local languages or displacing them. It was the official language of the Inca Empire, and groups of Quechua speakers were settled among other language groups, although the language does not seem to have been systematically imposed. The Spaniards, in turn, used Quechua in a great area as a language of evangelization—at one period missionaries were required to know the language—and continued to spread it by means of Quechua speakers who travelled with them in further conquests. During the 17th and 18th centuries it became a literary language in which religious, historical, and dramatic works were written. Today its written literary manifestations are not spontaneous, but there is abundant oral poetry, and in Bolivia radio programs are broadcast entirely in this language.

Dispersion of Tupí-Guaraní dialects, taking place shortly before the arrival of Europeans and even after it, resulted not from imperial expansion—as for Quechua—but from extreme tribal mobility and the cultural and linguistic absorption of other groups. Under Portuguese influence the modified form of Tupinamba known as *língua-geral* ("general language") was the medium of communication between Europeans and Indians and among Indians of different languages in Brazil. It was still in common use along the coast in the 18th century, and it is still spoken in the Amazon. Tupí, now extinct, was an important language of Portuguese evangelization and had a considerable literature in the 17th and 18th centuries. Another dialect, Guaraní, was the language of the Jesuit missions and also had abundant literature until the middle of the 17th century when the Jesuits were expelled and the missions dispersed. Nevertheless, Guaraní survived in Paraguay as the language of a culturally non-Indian population and is today the only Indian language with national, although not official, status—persons not speaking Guaraní being a minority. Paraguayan Guaraní is also a literary language, not so much for learned works—for which Spanish is used—but for those of popular character, especially songs. There is a more or less standardized orthography, and persons literate in Spanish are also literate in Guaraní. A great mutual influence exists between Guaraní and Spanish.

Grammatical characteristics. Diversity rather than common traits characterizes the grammar of South American Indian languages. Features commonly encountered seem to reflect facts of frequency in general typology rather than traits specific to this area. The greatest number of languages are probably suffixing languages like Quechumaran and Huitotoan, or use many suffixes and some prefixes like Arawakan and Panoan. Also very numerous are those languages having few prefixes and suffixes, such as Ge, Carib, or Tupian. Languages employing only prefixes to show grammatical distinctions have not been reported. There are a few with many prefixes but still more suffixes (Jebero, or Chébero); others, like Ona and Tehuelche, with almost no affixing, are also rare.

Grammatical diversity

Similarly, the complexity of words varies a great deal. In Guaraní words with three components and in Piro (Arawakan) words with six elements are of average complexity for the respective languages. In languages like the Cariban or Tupian ones, word roots are nominal (nouns) or verbal (verbs) and may be converted into the other class by derivational affixes; in languages like Quechua or Araucanian, many word roots are both nominal and verbal. Languages like Yuracare form many words by reduplication (the repetition of a word or a part of a word), a



Distribution of the South American Indian languages. The numbers and letters refer to the numbers and letters in the Table.

process that does not occur systematically in the Tupian languages. Compounding, the joining of two or more words to form new words, is a very widespread type of word formation, but it can be nearly absent, as in the Chon languages. Verb stems in which the nominal (noun) object is incorporated are also rather frequent. Many languages are of the agglutinative type (Quechuan, Panoan, Araucanian); *i.e.*, they combine several elements of distinctive meaning into a single word without changing the element. Others (Cariban, Tupian) show a moderate amount of change and fusion of the elements when combined in words.

Grammatically marked gender in nouns occurs in Guaycuruan (Guaicuruan), and a difference in masculine and feminine gender in the verb occurs in Arawakan, Huitotoan (Witotoan), and Tucanoan, but genderless languages are more common. Singular and plural in the 3rd person ("he, she, it") is not obligatorily distinguished in Tupian and Cariban, but languages like Yámana and Araucanian have singular, dual, and plural. A very common distinction is that between inclusive 1st person ("you and I," hearer included) and exclusive 1st person ("he and I," hearer excluded). Pronominal forms differentiated according to categories that indicate wheth-

er the person is present or absent, sitting or standing, and so forth occur in Guaycuruan languages and Movima.

Indication of case relationships

Case relations in nouns are generally expressed by suffixes or postpositions; the use of prepositions is rare. Possession is indicated predominantly by prefixes or suffixes, and systems in which possessive forms are the same as those used as the subject of intransitive verbs and as the object of transitive ones are rather common. Classificatory affixes that subclassify nouns according to the shape of the object occur in the Chibchan, Tucanoan, and Waican groups.

Very frequently the verbal forms express the subject, object, and negation in the same word. The categories of tense and aspect seem to be about evenly represented in South American languages, but the specific categories expressed vary a great deal from language to language: Aguaruna (Jívaroan) has a future form and three past forms differentiated as to relative remoteness, while in Guaraní the difference is basically between future and nonfuture. Other languages like Jebero express fundamentally modal categories. Very common are affixes indicating movement, chiefly toward and away from the speaker, and location (*e.g.*, in Quechumaran, Záparo, Itonama), and in some stocks like Arawakan and Panoan there are many suffixes in the verb with very concrete adverbial meaning, such as "by night," "during the day." Classificatory affixes indicating the way the action is performed—by biting, striking, walking—occur in Jebero and Tikuna (Ticuna). Actions done individually or collectively are differentiated paradigmatically in Carib, while in Yámana and Jívaro different verbal stems are used according to whether the subject or the object is singular or plural. There are also various languages (Guaycuruan, Mataco, Cocama) in which some words have different forms according to the sex of the speaker.

Equational sentences are very common. These are formed by juxtaposing two nominal expressions (nouns) without a linking verb, a fact that usually correlates with the absence of a verb "be" for expressing identification or location (*e.g.*, "John good man," "my house there"). Sentences in which the predicate is a noun inflected like a verb with the meaning "being" or "having" that thing designated by the noun also occur in Bororo and Huitoto (Witoto); *e.g.*, "I–knife" = "I have a knife." Sentences in which the subject is the undergoer of the action are frequent, but true passive sentences in which the undergoer and the agent are expressed are rare, though they do occur in Huitoto. Subordinate sentences are rarely introduced by conjunctions; subordination is usually expressed by postposed elements or special forms of the verbs such as gerunds, participles, or subordinate conjugations.

Phonological characteristics. As in grammar, there are no phonological features common to all South American languages that would be specific to them alone. The number of distinctive sounds (phonemes) may vary from 42 in Jaqaru (Quechumaran) to 17 in Campa (Arawakan). Jaqaru has 36 consonants, while Makushí (Cariban) has 11; some Quechuan languages have only three vowels, whereas Apinayé (Macro-Ge) has ten oral vowels and seven nasal ones. A dialect of Tucano (Tucanoan) exhibits three contrasting points of articulation, while Chipaya (Macro-Mayan) has nine. Many types of contrasting sounds occur although not with equal frequency. Voiceless stops (*e.g.*, *p, t, k*) occur everywhere, but voiced stops (*e.g.*, *b, d, g*) may be absent, and fricatives (*e.g.*, *f, v, s, z*) may be few in number. Glottalized voiceless stops —consonants made with simultaneous closure of the glottis and without vibration of the vocal cords—are rather common (Quechumaran, Chibchan), but not glottalized voiced stops (in which the vocal cords vibrate). Also less frequent are aspirated (Quechumaran) and palatalized sounds (Puinave); glottalized nasal sounds (Movima) and voiceless laterals (*l*-like sounds, as in Vilela) are rare. A distinction between velar and postvelar sounds occurs in Quechumaran and Chon, between velar and labiovelar in Tacana and Siona (Sioni); palatal retroflex consonants, made with the tip of the tongue turned up touching the palate, occur in Pano-Tacanan and Chipaya.

Great variety in distinctive sounds

Systems with nasal vowels are common (Macro-Ge, Sabelan), but in several languages (Tupian, Waican) nasalization is a feature not of vowels and consonants but of whole words. There is an apparent absence of front rounded vowels (*ü, ö*), but central or back unrounded vowels (*ï, ï*) are common. Systems with long vowels occur in Chipaya and some Cariban languages, and glottalized vowels occur in Tikuna and Chon languages. Very common are pitch-stress systems with high and low tones on stressed syllables; *e.g.*, in Panoan, Huitotoan, and Chibchan. More complex systems with three tones as in Acaricuara, four as in Mundurukú (Mundurucú), and five as in Tikuna are rare. Syllables are generally without complex consonant clusters.

The typology proposed by Tadeusz Milewski, a Polish linguist, classifies American Indian languages into three types: (1) Atlantic, with few oral consonants but complex systems of nasal consonants, and oral and nasal vowels, of which the Ge languages would be typical; (2) Pacific, with complex systems of oral consonants (many contrasting points and modes of articulation) but with few nasal consonants and few vowels, as exemplified by Quechumaran; and (3) Central, with consonant systems more like the Pacific type and vowel systems like the Atlantic, of which Chibcha would be typical. The typology is probably too gross to accommodate meaningfully every language type found in South America, but it holds to a certain extent, especially for the Atlantic type (Macro-Ge, Tupian, and Cariban).

Vocabulary. Indian languages vary significantly in the number of loanwords from Spanish and Portuguese. Massive borrowing has taken place in areas where languages have been in intense and continued contact with Spanish or Portuguese, especially where groups are economically dependent on the national life of the country and there is a considerable number of bilingual persons, as in Quechuan, or where no cultural differences correlate with language differences, as in Paraguayan Guaraní. Borrowings have not been limited to designations of artifacts of European origin but affect all spheres of vocabulary, having displaced native terms in many cases. Neither are they limited to lexical items; they include function elements such as prepositions, conjunctions, and derivative suffixes. Sound systems have also been modified. In some contact situations in which the Indian group displayed an antagonistic attitude toward the European conquest, purism developed and loans are comparatively few; *e.g.*, Araucanian. When contact has been frequent but superficial, loanwords are usually scant, but the meaning of native terms has shifted or new descriptive terms have been coined to designate new cultural traits, as in Tehuelche.

Loanwords from Spanish and Portugese

Borrowings among Indian languages may have been more numerous than yet reported, judging from the wide and rapid diffusion that loans from Spanish and Portuguese had through the central part of South America. Borrowings between Quechua and Aymara have occurred in great number, but the direction of borrowing is difficult to determine. Many Indian languages in the Andes and the eastern foothills have borrowed from Quechua either directly or through Spanish. In Island Carib (an Arawakan language), borrowings from Carib (a Cariban language) have formed a special part of the vocabulary, properly used only by men; these words were adopted after the Island Carib speakers were subjugated by Caribs.

In turn, some Indian languages have been a source of borrowing into European languages. Taino (Arawakan), the first language with which Spaniards had contact, furnished the most widespread borrowings, including "canoe," "cacique," "mais," and "tobacco," among many others. No other South American Indian language has furnished such widespread and common words, although Quechua has contributed some specialized items such as "condor," "pampa," "vicuña." The larger number of Arawakan borrowings results from these languages having been predominant in the Antilles, a region where Dutch, French, English, Portuguese, and Spanish were present for a long time. Cariban languages, the other important

group in that region, do not seem to have furnished many words, but "cannibal" is a semantically and phonetically modified form of the self-designation of the Caribs. The influence of some Indian languages on regional varieties of Spanish and Portuguese has been paramount. Thus Tupí accounts for most Indian words in Brazilian Portuguese, Guaraní in the Spanish of Paraguay and northeast Argentina; and Quechua words are abundant in Spanish from Colombia to Chile and Argentina. In addition, Quechuan and Tupí-Guaraní languages account for most place-names in South America.

Language and culture

No detailed studies are available concerning the relationship of the vocabularies of Indian languages to the culture. Certain areas of vocabulary that are particularly elaborated in a given language may reflect a special focus in the culture, as for example the detailed botanical vocabularies for plants of medical or dietary importance in Quechua, Aymara, and Araucanian. Shifts in cultural habits may also be reflected in the vocabulary, as in Tehuelche, which formerly had a vocabulary designating different kinds of guanaco meat that is now very much reduced, because the group no longer depends on that animal for subsistence. Kinship terminology is usually closely correlated with social organization so that changes in the latter are also reflected in the former: in Tehuelche, former terms referring to paternal and maternal uncles tend to be used indiscriminately, even replaced by Spanish loans, because the difference is not functional in the culture any more.

Proper names, to which different beliefs are attached, offer a variety of phenomena, among them the practice of naming a parent after a child (called teknonymy) in some Arawakan groups; the repeated change of name according to various fixed stages of development, as in Guayaki; word taboo, forbidding either the pronunciation of one's own name or the name of a deceased person, or both, as in the southernmost groups (Alacaluf, Yámana, Chon) and in the Chaco area (Toba, Terena); and the use of totemic names for groups, as in Panoan tribes.

Writing and texts. The existence of pre-Columbian native writing systems in South America is not certain. There are two examples, that of the Cuna in Colombia and an Andean system in Bolivia and Peru, but in both cases European influence may be suspected. They are mnemonic aids—a mixture of ideograms and pictographs—for reciting religious texts in Quechua and ritual medical texts in Cuna. The Cuna system is still in use.

Although the linguistic activity of missionaries was enormous and their work, from a lexicographic and grammatical viewpoint, very important, they failed to record texts reflecting the native culture. The texts they left for most languages are, with a few exceptions, of a religious nature. Most of the folklore has been collected in the 20th century, but many important collections (*e.g.*, for the Fuegian and Tacanan tribes) are not published in the native language but rather in translation. There are good texts recorded in the native language for Araucanian, Panoan, and Cuna, for instance, and more are being recorded by linguists now, though not necessarily analyzed from a linguistic point of view.

Efforts are being made in several areas to introduce literacy in the native Indian languages. For some, practical orthographies have existed since the 17th century (Guaraní, Quechua); for several others, linguists have devised practical writing systems and prepared primers in recent years. The success of these efforts cannot yet be evaluated.

BIBLIOGRAPHY. CESTMIR LOUKOTKA, *Classification of South American Indian Languages* (1968), contains a full list of languages, an extensive and detailed bibliography, the location and classification of languages, and a map; J.A. MASON, "The Languages of South American Indians," in J.H. STEWARD (ed.), *Handbook of South American Indians* (U.S. Bureau of American Ethnology Bulletin 143), 6:157–317 (1950), has an extensive bibliography, general information on language groups, a discussion of classifications, and a map. See also N.A. MCQUOWN, "The Indigenous Languages of Latin America," *Am. Anthrop.*, 57:501–570 (1955), a critical appraisal of classifications with useful lists of languages and families, the location and classification of languages, and a map; C.F. and F.M. VOEGELIN, "Languages of the World: Native America," *Anthrop. Linguistics*, vol. 6, no. 6 and vol. 7, no. 7 (1964–65), general information on American Indian languages, much information on groups and individual languages, and a discussion of classifications; T.A. SEBEOK (ed.), *Current Trends in Linguistics*, vol. 4, *Ibero-American and Caribbean Linguistics*, pt. 2, "Linguistics of Non-Ibero-American Languages" (1968), critical surveys of work done during the last 20 years; M. SWADESH, "Afinidades de las lenguas amerindias," in *Akten des 34. Internationalen Amerikanisten Kongress*, pp. 729–738 (1964), and J.H. GREENBERG, "The General Classification of Central and South American Languages," in *Men and Cultures: Selected Papers of the 5th International Congress of Anthropological and Ethnological Sciences, Philadelphia 1956* (1959), two recent classifications of South American languages.

(J.A.S.)

South American Nomad Cultures

In the past, South American nomadic hunters, gatherers, and fishers could be found all the way from Cape Horn to the Orinoco River in northern South America. The most variable groups were found in the southern half of the continent, occupying a variety of habitats and exploiting differing resources. With the technology known to them, food production was low, the population sparse, the social organization simple. Constant movement within prescribed territories prevented large permanent villages or the accumulation of material wealth.

TRADITIONAL CULTURE PATTERNS

Types of South American nomads. *Shellfish gatherers.* In the south, the Chono, Alacaluf, and Yaghan Indians occupied the whole Chilean archipelago down to Cape Horn. This is a rugged terrain of islands and fjords with heavy rainfall, an average winter temperature of 32° F (0° C), and an average summer one of 50° F (10° C). The dense forests make land travel extremely difficult and horticulture impracticable. The area is poor in game, fish, and edible plants. The archipelagic tribes thus depended on shellfish and seals or whales that had been stranded on the beaches. Travel was almost entirely by canoe.

Hunters and gatherers of the steppes and plains. The large area of the steppes and plains extends from Tierra del Fuego, in the south, through Patagonia, to the Pampas of northern Argentina and Uruguay. The Ona occupied the open country on the islands of Tierra del Fuego. The brush-covered, semi-arid Patagonian plateau was the home of the Tehuelche, while the Puelche and Querandí inhabited the flat grassy Pampas. The Charrúa lived in the grasslands north of the Río de la Plata. In aboriginal times, these people practiced no agriculture and had no domesticated animals, with the possible exception of the dog. Throughout the whole region the tribal groups depended on hunting guanaco, rhea (the South American ostrich), and smaller animals and on gathering some roots and herbs. The population was one of the sparsest in South America.

Hunters, gatherers, and fishermen of the Gran Chaco. The Gran Chaco extends northward from the grasslands of the Pampas to Paraguay and southern Mato Grosso in Brazil. It is an arid region covered with drought-resisting vegetation. The area is drained by the Paraguay River and its western tributaries, such as the Pilcomayo, Bermejo, and Salado rivers, that originate in the Andean foothills. During the summer months the Chaco experiences the highest temperatures in South America.

The people subsisted on plant food, supplemented near the rivers by fish at certain times of the year. The plant foods were supplied by such pod-bearing thorny bushes as the algarrobo and by many local trees. Some wild rice was also available. Honey and larvae were also eaten. In the southeastern part, guanaco and rhea were hunted. On the whole, however, the people depended primarily on plant foods, in contrast to the nomads to the south, who were essentially hunters.

In aboriginal times all the nomadic peoples of the Chaco travelled on foot or, in some cases, in canoes. The horse was introduced into the Chaco after the Spanish conquest, and its adoption by some tribes had far-reaching consequences in the area. It is convenient to separate



Distribution of South American nomadic peoples.
From J. Steward and L. Faron, *Native Peoples of South America*

the Chaco tribes of historic times into foot Indians and horsemen. Among the foot Indians were such groupings as the Zamuco, of the northeast, and the Mataco, of the central Chaco. Each such grouping consisted of a number of tribes. The mounted bands, who spoke Guaycuruan, consisted of such groups as the Abipón, Mocoví, and Mbayá.

Forest hunters and gatherers. North of the Chaco, the country merges gradually into the tropical forest zone, particularly in the western section of Bolivia. In Brazil the forest zone comprises columns of forests on both banks of the major rivers. There are also island forests—large patches of forest standing on a plain or plateau, evidently supplied by springs that flow the year round. Typical nomadic tribes in this area were the Sirionó of eastern Bolivia and the Nambikwara (Nambicuara) of Mato Grosso, Brazil, and the Guayakí of eastern Paraguay.

Aquatic nomads. In the marshes of the upper Paraguay River in Brazil an unusual tribe of Indians, called the Guató, lived most of their lives in canoes, fishing and hunting cayman and other aquatic animals. They built temporary shelters on small islands that stood slightly above flood level. There are also other aquatic nomads in the north of South America and in the Caribbean.

Economic system. Because of their nomadic habits, all the hunters and gatherers had very little in the way of such material equipment as weapons, textiles, clothing, ornaments, and buildings. Their technical processes were very simple and appear to have been invented long ago.

Housing

Shelter was provided by caves if available. In the colder climate of the south, the archipelagic tribes of Chile and the nomads of the Chaco made domed huts of bent poles covered with bark, skins, or brush. When the people moved on they left the frame for others to use, taking only the skin coverings with them. The Patagonians made

a skin-covered hut known as the *toldo*. The Yaghan used a conical tepee-like shelter or a double lean-to. The Nambikwara used a lean-to in the dry season or camped under trees, sleeping on fire-warmed ground. During the rainy season a larger double lean-to was used. There were no permanent settlements, although people sometimes gathered together to perform ceremonies and to feast when food was plentiful.

The forest hunters, such as the Sirionó and Nambikwara, wore no clothing. The southern nomads wore skin robes and crude moccasins. There was no sewn clothing. Earplugs, nose plugs, and lip plugs were widely used, except by the archipelagic people. Featherwork, armbands and leg bands, necklaces, and body painting were common in many areas. Some of these ornaments were used to distinguish bands or lineages and other groupings, but they were not used as status symbols.

Finger weaving of yarn spun from native cotton and palm frond fibres was practiced in the Chaco and among the Sirionó and Guató. The heddle loom, a later development, was known among the Sirionó, Nambikwara, and Chono. Long strips of fabric were woven for making armbands and leg bands and other decorations. Netting was used for making fishnets and bags for the transportation of goods, particularly in the Chaco.

Pottery was known to some of the nomads but was little used because pots were difficult to transport. Coiled basketry was widely used by the nomads. In the Chaco, twilled baskets made from palm fronds were used at campsites and abandoned when the people moved on. The Patagonian and Pampean hunters used containers made of skins.

Two methods of making fire were widespread. The first involved a spark with flint on iron pyrite. A later technique involved twirling a hardwood pointed stick in a socket in softer wood: dried pith was then placed around the drill, and by gently blowing on the spark the pith ignited. Meat and fish were cooked by placing them directly on coals or putting them into earth ovens, lined with heated stones and covered with earth and coals. The Chono boiled food by placing heated stones in tightly woven baskets. The hunting and fishing people used no salt, but the Chaco tribes, who depended primarily on plant food, traded for salt with the highland people. Some of the forest nomads used ashes in place of salt.

Weapons

Bows and arrows were used by all the nomads. Among the Patagonian and Pampean hunters, however, there is archaeological evidence to suggest that the bow and arrow was preceded by the bola. Before the introduction of the horse, guanaco and rhea were hunted by stalking, the hunter throwing the bolas around the neck or legs of the game. Bolas were made by attaching stone weights to two or three short cords that, in turn, were fastened to a longer lasso. With the coming of the horse after the Spanish conquest, the bolas became very important, for from horseback they could be easily swung to ensnare guanaco, rhea, wild cattle, and other large game. Among the Patagonians, Pampeans, and inhabitants of parts of the Chaco, it became the principal hunting device. Spears and the *atlatl*, or spear thrower, were used to some extent.

Among the forest nomads, such as the Sirionó and Nambikwara, the principal weapon for hunting and fishing was the longbow, which was six feet in length. The barbed arrows were from five to eight feet long. Because they had no canoes, both shot fish from the banks of a stream.

Among the archipelagic tribes of southern Chile it was predominantly the women who gathered shellfish on the beaches at low tide and who, from bark canoes, dived with a shell blade and a basket held in their teeth. The shellfish gatherers were careful not to exhaust the supply in one area. These people also always carried a fire on a clay platform in their canoes, both for warmth and for roasting shellfish over the coals. The men hunted cormorants, penguins, steamer ducks, petrels, and other marine birds at night with torches while they roosted, and clubbed them to death. Ducks and geese were lured by decoys, then captured with pole snares.

Seals and sea lions were harpooned in the water or clubbed on shore. Porpoises and sick whales were harpooned. Whale hunting was a cooperative enterprise involving many men, who risked their lives in flimsy bark canoes. Fish were sometimes found in shallows or in pools at low tide and, with the help of dogs, were driven into nets. Because the Indians had no knowledge of food preservation, they had to be constantly on the move to provide for their food supply.

Social organization. The typical organization among nomadic hunters, gatherers, and fishers was the band, which, depending on the resources, could be large or very small. Low productivity and the lack of developed transportation prevented the accumulation of a surplus to maintain permanent communities. There were no social or occupational specialists; every family produced its own equipment. Despite these general similarities there was wide diversity in social structure depending on the methods of obtaining food.

Among the Chono, Alacaluf, and Yaghan of the Chilean archipelago, the natives were dispersed in elementary family units of father, mother, and children with perhaps an elder or two. These small family bands, if they can be so called, moved from one beach to another. There were no permanent territorial claims to shellfish beds, although individual families repelled others while they were using a particular shellfish bed. Sometimes close relatives or friends would move together briefly; and at times a number of families would gather together to feast on a stranded whale or join in hunting seals or sea lions. The family was also the economic unit among the Guató and, during the dry season, among the Nambikwara. This, of course, does not mean that the people did not visit relatives when circumstances permitted or when certain religious and ceremonial activities demanded.

Multifamily bands. The hunting of guanaco and rhea among the eastern Yaghan and Ona and among the Patagonian and Pampean tribes was more productive when carried out cooperatively by a number of families banded together. Such bands numbered from 40 to 100 persons and had defined hunting territories, which the men defended against trespass. Chieftainship does not appear to have been hereditary but was ceded to a leader able in settling both internal disputes and conflicts between bands.

Composite bands. With the introduction of horses and cattle, a great change took place in the band organization. Horse nomadism permitted greater mobility, new techniques of hunting, and much larger bands. The former foot hunters joined into bands ranging from 500 to 1000 persons. They roamed over ill-defined areas hunting wild cattle and raiding Spanish settlements and other Indians lacking horses. Each of these bands consisted of a number of lineages under a leader of proven ability; a strong leader might attract a huge following, including members drawn away from other bands. Warfare between bands increased because of uncertainties over rights to territory.

The southern hunters of Patagonia and the Pampas were patrilineal (descent was reckoned in the male line) and patrilocal (a wife resided with her husband's lineage and band).

Forest nomads, such as the Guayakí and Sirionó, on the other hand, were matrilineal and matrilocal; that is, an individual traced his ancestry through his mother's lineage, and a man went to live with his wife's band. Matrilineal descent and matrilocal residence were associated with the importance of women in collecting wild plant foods.

Chaco tribes

Although little is known about the social structure of the Chaco tribes in aboriginal times, there appears to have been a contrast between the peoples of the dry western area and the wetter eastern area. Because the people in the west depended on water holes, they were forced to shift camp frequently as holes dried up. The groups nevertheless seem to have claimed territorial rights to gathering, fishing, and hunting areas. With the arrival of the Spaniards and an increase in warfare, the authority of the chiefs was strengthened, although chieftainship was rarely hereditary. In the eastern Chaco, on the other hand, the presence of fish runs in the larger rivers and the

practice of fairly productive cultivation permitted the settlements to be larger and less mobile. After acquiring the horse from the Spanish, however, the Mbayá and other Guaycuruan-speaking peoples gave up what little horticulture they practiced and became predatory nomads raiding Spanish settlements, taking cattle, and capturing slaves from more sedentary tribes. Other Chaco tribes, such as the Abipón, Mocoví, Toba, and Lengua also became horsemen and raiders. These tribes continued to move their camps in search of pasture for their herds of horses and cattle. Incipient class differences based on war honours and wealth appeared.

Mbayá raiders

The Mbayá were outstanding raiders in the Chaco. Although roaming over great areas, the warrior bands always returned to their base settlements, where they had permanent houses and kept their slaves and livestock. The Mbayá also exhibited the clearest form of social stratification, which, although pre-Spanish, crystallized with the coming of the horse and the intensification of warfare. Mbayá society became stratified into nobles, warriors, serfs, and slaves. The nobles were divided into those who inherited their titles and those upon whom titles were bestowed for lifetime only. The warrior class was basically hereditary, but other men demonstrating greatness in war could become members, thereby establishing new hereditary lines. The serfs, who served only the members of the noble class, were from subjugated peoples. The lower class was made up of captured and purchased slaves, who included not only Indians from neighbouring tribes but also mestizos from the Spanish settlements. Slaves could gain their freedom by marrying into the warrior class.

Family and kinship. Marriage among most nomadic tribes was consensual, similar to common-law marriage. It was easily entered into and easily dissolved, although there were strong forces supporting its continuance, especially whenever women play an important role as food gatherers. Noble classes, where they existed, as among the Mbayá, were practically endogamous; virtually all men, that is, married within their own class.

Marriage customs

Marriage among the shellfish gatherers of the Chilean archipelago, on the other hand, was a stronger institution. Postmarital fidelity was demanded. The family formed a strong autonomous unit that performed nearly all cultural activities on its own and cooperated with other families only briefly during sea hunts and initiation ceremonies. Marriage between known relatives was forbidden, but in practice this meant merely that first cousins and closer kin could not marry. A widow, however, could marry the brother of her deceased husband; and a widower could marry the sister of his deceased wife; these practices of levirate and sororate helped to maintain family alliances, when almost everything else tended to draw families apart.

The larger nomadic bands in South America practiced band exogamy; that is, a person in one band could marry only someone in another band. These marriages were not made at random, however, for (as among the Nambikwara) cross-cousin marriage was preferred; in a matrilineal society a man married his mother's brother's daughter; in a patrilineal society he married his father's sister's daughter.

Rites of passage. Birth ceremonies were simple family affairs. After the birth, both parents fasted for a few days and observed certain food taboos. Couvade was practiced; that is, the father stayed in the hut several days, mimicking labour, while relatives and friends provided essential needs.

Among the Sirionó a child was born openly in the communal house; and after birth the parents walked in the forest scattering ashes as a purification rite and then lit a new fire that signified new life.

Before the age of puberty, boys and girls learned by imitating older children and adults. Among the shellfish gatherers, children by the age of four began to gather shellfish and spear sea urchins close to shore, returning to camp to roast them and eat them. From an early age children thus took care of their food needs as far as shellfish were concerned. Boys and girls were separated after the age of seven. The boys played with bows and arrows. The girls learned to swim and dive. Males did not learn to swim or dive, since diving for shellfish was women's work. Corporal punishment was rare, but children were lectured by elders on manners and morals.

Socialization was formalized especially in the initiation rite, which marked the passage from youth to adulthood for both boys and girls. There was usually no fixed date, the time depending upon the number of neophytes and the opportunity to amass a sufficient supply of food for the feast.

Initiation rites

The initiation ceremonies began with the men preparing sealing clubs and shellfish poles in a special hut in which they painted their faces and participated in singing, dancing, and mummery. The men then went out to hunt seals on the coastal rookeries, and the women went for shellfish. The men then built a large hut, where they sang, danced, and instructed the young men in proper vocational and moral behaviour. Later, women joined in the ceremony instructing girls in the proper behaviour for women. Then followed a mock battle between the sexes. After a feast, the assembly disbanded.

Among the Patagonian and Pampean tribes, a special hut known as the pretty house was erected for initiation ceremonies (as well as for some other rites, such as first menses). Medicine men bled themselves and smeared the novices with blood. There was dancing by the men and singing by the women. Horses were killed and roasted, and horsemeat was passed out to the guests.

In the Chaco there was considerable variation in the details of the initiation rites, but the underlying purpose of education and socialization was the same as among the shellfish gatherers and the Patagonian and Pampean guanaco hunters. Boys went through several rites, and when blood was drawn from their genitals they were considered mature warriors. On the Bermejo and Pilcomayo rivers, girls' puberty rites were attended with singing and dancing designed to protect the girls from evil spirits. A girl was kept in isolation and observed a special diet. After the rites pubescent girls were allowed sexual liberty and often took the initiative in love affairs.

Among none of the nomadic peoples did marriages involve any special ceremonies; gifts, though, were exchanged between the bride's and the groom's parents. Death rites were more complex. Mourners painted their faces black, beat on the outside of the dead person's hut, fasted, and lamented. They also directed their anger at the supreme deity. In the Chilean archipelago, the dead person and his effects were either buried or cremated. Among the Patagonian and Pampean tribes, the corpse was left on a hilltop or placed in a cave; some belongings were placed near the body.

Religion. Among the Yaghan shellfish gatherers there was a belief in a Supreme Being who was not a creator but a ruler. He was one who gave life to men and who gave humans animal and plant foods. People prayed to this being for success in fishing and hunting. Among the Patagonian and Pampean tribes there was a belief in a supreme being who, after creating the world, did not enter further into human affairs. There was a belief in good and evil bush spirits.

The Chaco groups did not believe in a supreme being of any kind. Although celestial bodies sometimes were thought to affect human beings, they were not objects of worship. The Chaco people had great fear of the ghosts of the dead and disposed of the corpse as quickly as possible. The body was buried in a cemetery, and food offerings were made. The house and property of the deceased were burned.

Forest nomad ceremonies

Among such forest nomads as the Sirionó and Nambikwara, rituals and ceremonies were much less developed than in the Chaco. This no doubt was due to incessant search for food and the inability to accumulate surpluses for large-scale feasts. Although the Sirionó did not believe in a supreme being, they did consider the Moon to be a culture hero who gave them maize and manioc and other features of their culture. They also feared the ghosts of the dead and bush spirits.

Shamans, or medicine men, who were thought to receive

their curative powers through the ghosts of dead shamans and special guardian spirits, were important among all the tribal groups. Among the Chaco groups shamanism was very highly developed, both for curing illnesses and in working for the general welfare of the tribe. Sickness was caused, it was thought, by one of two means: mysterious foreign objects would magically penetrate the body, causing disease, or a person's soul would leave the body, leaving him ill. In the former instance, the shaman would suck out the foreign object; in the latter, he would go out at night and bring the wandering soul back.

CHANGES IN NOMADIC CULTURES

The nomadic hunters and gatherers today are marginal survivors who retain many archaic culture traits and share very few of the more recent inventions. In areas in which they have not been disturbed by the whites or in which they have had opportunity to withdraw into refuge areas, they have been able to maintain much of their original culture. In areas in which they have come into permanent contact with the whites, they have become rudimentary agriculturalists, building permanent houses, making pottery, and weaving.

Disintegrative factors are the effect of white man's diseases and miscegenation. The Chono in southern Chile became extinct in the 19th century, and the Alacaluf and Yaghan have mixed with whites. The horse nomads became cowboys on Argentine ranches and commonly intermarried with non-Indians. The same can be said of the Chaco tribes. The forest nomads are perhaps closest to the original pattern of nomadic culture. As the whites expand, the nomads will increasingly settle down as village dwellers and farmers on lands allotted to them, following the pattern that took place in North America.

BIBLIOGRAPHY. JULIAN H. STEWARD (ed.), *Handbook of South American Indians*, vol. 1, *The Marginal Tribes* (1946), is the outstanding authoritative work. JULIAN H. STEWARD and LOUIS C. FARON, *Native Peoples of South America* (1959), is a one-volume work summarizing the material in the handbook for the lay reader. ALLAN R. HOLMBERG, *Nomads of the Long Bow: The Siriono of Eastern Bolivia* (1950; paperback ed., 1969), is the best work on these people, a good example of the forest nomads. KALERVO OBERG, *The Terena and Caduveo of Southern Mato Grosso, Brazil* (1949), is a Smithsonian Institution monograph that includes an account of the Caduveo, a sub-tribe of the once powerful Mbayá. KALERVO OBERG, *Indian Tribes of Northern Mato Grosso, Brazil* (1953), includes a short account of the Nambikwara.

(K.Ob.)

South American Peoples and Cultures

The customs and social systems of South American peoples are closely and naturally related to the environments in which they live. These environmental relationships are mediated by the systems of technology that the people use to exploit their resources.

Four basic types of social and cultural organization of South American peoples emerge from the archaeological and historical records: (1) central Andean irrigation civilizations, (2) chiefdoms of the northern Andes and the circum-Caribbean, (3) tropical-forest farming villages, and (4) nomadic hunters and gatherers. Each type developed in its own fashion during thousands of years, and since the 16th century each has made a distinctive adjustment to the impact of European civilization.

Early men, hunters and gatherers with no knowledge of agriculture, gradually worked their way across the Bering Strait in pursuit of food and meandered over North and South America in small, migratory bands for thousands of years. They reached Tierra del Fuego in approximately 6000 BC, after passing through the bottleneck of Central America, dispersing in the rugged terrain of the northern Andes, following the resource-laden Caribbean coastline eastward, and filtering southward through the tropical lowlands now making up part of Venezuela, the Guianas, and Brazil. They also hunted game through the highland basins of the central Andes and hunted and fished along the west coast of South America until they reached land's end.

CULTURAL AND ECOLOGICAL AREAS

Hunters and gatherers. Peoples who led a nomadic hunting and gathering life inhabited the agriculturally marginal areas of South America and were peripheral to the centres of great cultural development. All of Argentina and the archepelagic zone of southern Chile comprised the habitat of such hunting and gathering peoples as the Chono, Alacaluf, and Yámana (Yaghan) of Chile, the Ona of the island of Tierra del Fuego, and the Tehuelche, Guennakin (Puelche), Charrúa, and Querandí of mainland Argentina. In the Gran Chaco region were the Guaycuruan-speaking Indians, the Abipón, Mataco, Vilela, and others, who were migratory peoples roaming the grassy plains of their small territories in search of rhea, guanaco, peccary, and jaguar. There are still hunters and gatherers in the tropical rain forests of Brazil and neighbouring countries, societies which are off the mainstreams of diffusion of ideas and technologies and which have remained at a hunting and gathering subsistence level. Many such peoples became extinct shortly after contact with Europeans, through warfare, enslavement, and disease. Others, such as the Guaraní Indians of Paraguay, made prolonged adjustments to European colonization and gradually mixed with the conquerors biologically and culturally.

In the tropical forests were the Jívaro, Yaruro, Makú, and many other small societies eking out a livelihood mainly by hunting, fishing, and gathering wild plants. They kept a wary eye on their more powerful neighbours, the village agriculturalists, who coursed the main rivers and their tributaries in canoes, searching for food and sometimes human heads.

Social behaviour of hunters and gatherers

The hunting and gathering peoples of aboriginal South America were organized into small social units made up of a single kin group or of several loosely linked groups of relatives. Members of these societies were differentiated almost entirely on the basis of their sex and age rather than on status characteristics of an economic, military, political, or religious nature, as in more complexly organized social systems. Behaviour was sanctioned by tribal customs that involved kinship rights and obligations and constituted the basis of morality. These peoples had very similar rites throughout the South American continent and similar beliefs in cures and magic. Their technology and material culture, though not homogeneous from one society to another, was always rudimentary and generally lacked agriculture, well-developed building arts, and manufacturing processes found among other South American Indians. They lived in marginal areas, exploiting the limited natural resources with elementary techniques. This kept their social units small, widely scattered, and simply organized.

Tropical-forest farming villages. The agricultural villagers of the tropical forests had more developed exploitative techniques than the hunters and gatherers. Farming, food storage, and canoe transportation along rivers made for greater economic sufficiency and the ability to live in larger and more stable units. The forest-dwelling agriculturalists included the bulk of the Arawakan-, Cariban-, and Tupian-speaking peoples, such as the coastal Arawak proper and those of the Greater Antilles, the Achagua, Guahibo, Palicur, and many other societies; the Carib of the Guianas, such as the Barama River Carib, the Taulipang, and the Makushí; the Tupians of the coast of Brazil, such as the Tupinamba; and inland groups among whom were the Mundurukú, Kawaíb (Parintintín), and their neighbours.

Society and economy of the tropical forest

Tropical-forest farming villagers, like hunters and gatherers, had sociocultural units consisting mainly of kin-based populations which were structured along lines of age and sex, without much in the way of economic, political, or religious grounds for social-status differentiation. Social controls were largely based on kinship rights and obligations of a moral nature, except in cases of certain military activities that were often under the temporary leadership of special chiefs. Their richer technology and production of agricultural surpluses enabled villages to remain in the same place for many years, even though the depletion of soils necessitated the periodic re-establish-



Distribution of South American indigenous peoples.

ment of new villages and the abandonment of older ones. Populations were larger and, of course, more concentrated. They were supported by a more adequate and dependable food supply, which included maize (corn), beans, squash, manioc, and tropical vegetables and fruits, as well as the riches of the rivers on which these peoples lived, such as turtles and the thousands of turtle eggs harvested annually and abundant fish and game. Hunting was important but subsidiary to agriculture. The rites of these peoples—those surrounding birth, puberty, initiation into men's secret societies, marriage, and death, and the shamanistic practices involved in the supernatural curing of illness—tended to be similar throughout the tropical-forest region. Many of the rites were similar to those of the simpler hunting and gathering peoples.

Chiefdoms of the northern Andes and the circum-Caribbean. In this extensive and geographically varied region there existed many peoples who lay in the main path of the Spanish conquistadores and who were overwhelmed by them. The Spaniards were attracted by the abundance of gold ornaments and religious objects displayed in the native villages and were excessive in their search for even greater wealth.

Among the chiefdoms were the Chibcha of highland Ecuador (the greatest chiefdom of them all) and the Coconuco, Pijao, Páez, Puruhá, Cana, and Palta of the northern Andes; the Jirajara and their neighbours, the Caquetío, Palenque, and Cumanagoto of northern Venezuela; and the Arawakan Taino of the Greater Antilles.

Though having a technology similar to that of the tropical-forest farming villages and sharing a basic material culture with them, the chiefdoms of the northern Andes and the circum-Caribbean areas had a still more productive food complex, which supported much denser populations in quite large and permanent villages and towns. Natural resources were more varied and abundant in the

regions that they inhabited, and farming was more productive.

Social and political life of the chiefdoms

Villages were composed of multi-kinship groups organized on the basis of social strata which had attributed statuses, rather than merely on the basis of kinship considerations such as age, sex, and the moral obligations these incurred. Some social ranks were hereditary, such as chieftainship and ritual office. Warfare was of great importance in many societies of this type. Participation in military activities insured upward social mobility for individuals and families and the eventual achievement of membership in the topmost strata of the village. War captives were taken as drudge servants and for sacrificial victims in religious rites. There was a foreshadowing of state institutions in the offices of priest, chief, military leaders, and nobles and captive slaves. In the chiefdoms, however, these institutions had not crystallized as they eventually did in the Andean kingdoms and empires. A major diagnostic feature of chiefdoms was their priest–temple–idol complex, a ritual organization of a different order of complexity from the supernatural beliefs and practices of the tropical-forest villagers and the hunters and gatherers.

Central Andean irrigation civilizations. First occupied by small groups of hunting and gathering peoples who filtered southward along the Pacific coast and through the highland basins thousands of years ago, the central Andes eventually became the seat of the highest form of civilization developed in native South America. The earliest archaeological evidence of agriculture in this region has a date of 2300 BC, which is probably much later than the first domestication of plants. With the spread of agricultural knowledge throughout the central Andes, populations increased in size and attained more settled and larger communities. A thousand years before the Spanish Conquest, the central Andes had the most developed agricultural and irrigational system in all of South America, the densest population south of Mexico, and the most efficient system of overland transportation in the Western Hemisphere. The combination of these features permitted the growth of true urban centres, an intricate class system, a strongly entrenched bureaucracy, and the extension of social controls over vast areas by means of political, religious, and military institutions.

Two of the most famous early kingdoms in the central Andes were the Tiahuanaco and the Chimú; the former spread its culture from what is today highland Bolivia northward to the vicinity of Lima and beyond. In the north of Peru arose the Chimú kingdom, which expanded southward and overlapped the northern extension of the Tiahuanaco culture, as the latter's influence had begun to decline. Following these two great cultural spreads and military conquests came the expansion of the Inca state. When Inca civilization reached imperial proportions it controlled the area occupied today by Peru, Ecuador, Bolivia, and the northern half of Chile. The expansion of the Inca Empire preceded the Spanish Conquest by slightly less than 100 years.

Irrigation, agriculture, and bureaucracy in Andean civilizations

All of these imperial states, as well as other smaller ones before them, shared a number of characteristics that set them apart from the chiefdoms and other peoples. They were based on state-controlled irrigation works, which made the production of huge agricultural surpluses possible. These surplus crops were controlled by the emperor and apportioned among the state, the church, and the populace according to a standard formula. As a result of an abundant food supply and surpluses that could be stored against adverse times, population steadily increased. There developed a rigidly hereditary class system—with the agricultural masses at the bottom and the Inca royal family at the top, with ranks of nobles, chiefs, lesser administrators, artisans, and others in between. The state waged war for territorial conquest and taxed the defeated peoples. It imposed the Inca religion, with its emperor-god and hierarchy of deities, its shrines and temples attended by priests and sacred virgins, and its ceremonial calendar. The Inca were masters of bureaucratic regimentation who ruled the lives of the commoners through political controls enforced by state machinery and statute law rather than by customary sanctions. Inca institutions overshadowed and to some extent replaced the traditional behaviour patterns of the thousands of farming communities that made up the empire.

Central Andean technology differed little from that of surrounding areas, except in metallurgical skills and in the building arts, but it was outstanding in the quality, variety, and excellence of its products, the most outstanding of which were produced for the state and the nobility by highly skilled artisans.

PEOPLE: LINGUISTIC AND PHYSICAL CHARACTERISTICS

In South America, native language families encompassed large blocks of territory and numerous societies. They cut across different cultural and social types and are found represented in different geographical and environmental surroundings. Languages may be grouped in many ways, but the major language groupings or families of South America may be conveniently divided into the Macro-Chibchan, Andean-Equatorial (including Tupian), Ge-Pano-Carib, and Hokan. This is the most simplified classification of South American Indian languages (see, however, SOUTH AMERICAN INDIAN LANGUAGES).

Racially, the Indians of North and South America belong to a Mongoloid subtype known as Amerind. They show certain diagnostic physical traits of Mongolian peoples, among which are the shovel-shaped incisor tooth pattern, the distinctive fold of the eyelid, and the darkening of the lower back known as Mongolian spot. Such traits were scattered in South America. Hair form is generally straight and hair colour black, but there are instances of brownish, reddish brown, curly, and even frizzy hair.

Great physical variation is to be expected in so large a population dispersed over an entire continent. Wide differences are apparent in body form, height, skin colour, and head shape. Of the more than 100 societies for which there are physical data of reliable value, height ranges greatly between means of about 57 inches (145 centimetres) and about 69 inches (175 centimetres). The peoples of smaller stature are found in the northwestern part of South America (for example, the Chocó and the Cuna) and continue southeastward in a broad band that extends into south central Brazil. They overlap considerably, however, with peoples whose mean height falls between 60 and 64 inches (152 and 163 centimetres). These peoples of moderate height extend farther south, down into Chile as far as the Isla de Chiloé, where there is again overlap with a very short people, the Chono. The next tallest groups of peoples comprise a narrow band from northern Brazil to southern Argentina. The tallest peoples in South America were the hunters and gatherers of the southern plains and Tierra del Fuego—the Puelche, Tehuelche, and Ona.

Population densities in the 16th century

In the 1500s, the central Andes, the area of greatest population density in South America (about ten persons per square mile), was sparsely populated compared to centres of Old World civilization. Yet its population of approximately 3,500,000, crowded into narrow coastal valleys and small highland basins on approximately 1 percent of Peru's total land area, constituted a much higher density than could be found in any other part of South America. The chiefdoms of the northern Andes, northern Venezuela, and the Antilles had an estimated total population of 1,900,000, with densities ranging from 6.6 to 1.1 persons per square mile (2.5 to 0.4 persons per square kilometre). The southern Andes was inhabited by the Atacama, Diaguita, and Araucanians, whose combined population was possibly 1,131,000, with a density range of 0.38 to seven persons per square mile. Tropical-forest peoples numbered about 2,200,000 and had a density of 0.6 per square mile. Hunting and gathering peoples of the Chilean archipelago, Patagonia, the Gran Chaco, and eastern Brazilian uplands had a combined population of less than 800,000 and a density range of 0.2 (Chilean archipelago) to 1.1 (western Chaco).

The population density of the central Andes was about 200 times greater than that of the hunters and gatherers, 20 times greater than that of the tropical-forest farmers,

and 30 to 40 percent greater than that of the Araucanians and the chiefdoms of the northern Andes and the circum-Caribbean.

FOUNDATIONS OF SOUTH AMERICAN CULTURES

Human life forms did not evolve in the New World, despite certain claims to the contrary which have never been taken seriously by most scholars. Man migrated from Siberia to Alaska, probably some 35,000 years ago, when there was a land and ice bridge between the two continents. He seems to have remained locked in the northwestern sector of North America for eons, held back by impenetrable glacial formations. When the glacial cap retreated and valleys opened up, a full-blown *Homo sapiens* (then existing as a hunter and gatherer) began to follow the southward progression of game animals, fanning out over North America and wedging through Central America into South America, again a process occupying thousands of years. Archaeological discoveries have unearthed man's skeletal remains in association with now-extinct species of animals and in geological deposits of the last phases of the Ice Age.

Early man. Archaeological evidence demonstrates that South America was occupied by early man at least 10,000 years ago, ample time for high civilizations to have evolved in the central Andes and for ecological adjustments to have been worked out elsewhere on the continent. Scientific dating techniques establish that agriculture was practiced along the Peruvian coast at least as early as 2300 BC. By 1000 BC agricultural societies flourished. This does not mean that all of South America had reached this stage of development nor that it was densely populated by farming communities. On the contrary, the continent was spottily inhabited by simply organized hunters and gatherers who then occupied the most favourable regions. As knowledge diffused from the central Andes to other parts of South America and as agriculture and other techniques were adopted by those peoples living in favourable environments, farming communities took form, and populations among them began to increase. Thus, on the foundation of early hunting and gathering societies, the more complex social and cultural systems gradually were built in those areas where agriculture developed; cultural growth and social complexity followed apace. Hunters and gatherers were pushed out of the farming regions to agriculturally marginal areas, where some of them are found today.

The original migrants to the New World had no knowledge of the domestication of plants or animals, with the exception of dogs, which were used in hunting. Recent discoveries in Mexico indicate that agriculture was independently discovered in the New World in roughly the same era that it was in the Near East (about 7000–8000 BC) and that New World civilizations were built on an indigenous agricultural base.

Independent American development

It is known archaeologically that cultural influences from Asia, as well as latter day migrations of people such as the Eskimo, continued to impinge on parts of the New World over the millennia, but New World cultural developments that culminated in the formation of high civilizations in Mexico and Peru were overwhelmingly the product of native, independent invention in almost all spheres of cultural and social life. Sporadic influences probably reached Peru and the western parts of the tropical forests from across the Pacific Ocean, but their effect on the course of cultural development in this hemisphere was negligible. Native America constituted a separate cultural unit, compared with the Old World.

The development of civilizations. The archaeological record for the central Andes shows a step by step development of cultural and social forms from a pre-agricultural, hunting and gathering baseline some 10,000 years ago to the Inca Empire in the 15th century AD. The record does not show any significant cultural influence on this development from transpacific contacts.

The evidence on early hunting and gathering peoples in Peru is still sparse. It is not yet possible to reconstruct social patterns, since most of the remains consist only of shellfish middens and small, widely scattered campsites along the coast. It was a period of thousands of years' duration, however, toward the end of which some knowledge of plant domestication reached the Peruvian coast.

The next major era is set off by incipient agriculture and is also characterized by the remains of small, hamlet-type communities along the Pacific Ocean near river mouths, where the alluvial soil was able to support crops. Technology remained simple, irrigation was not practiced, and population remained small.

After the passage of 1,000 years or so, marked developments appear in the archaeological record. These include many new crops, irrigation ditches that extended the arable area and controlled the supply of water, more and larger communities that attest to a growing population, and important temple mounds that formed the symbolic centres of theocratic government controlled by a priestly class. The formative era saw the development of the basic technologies and life-styles that were to become elaborated into even more complex cultural forms and state institutions. The emergence of city-states and empires in the central Andes is the result of local cultural–ecological adjustments of this sort, based on an irrigation agriculture that supported growing populations and necessitated controls in the hands of priests and nobles, with a warrior class subservient to the state.

Around 500 BC strong regional styles began to appear in the manufacture of utilitarian and luxury goods and public buildings. An abundance of large temple mounds, more extensive and intricate irrigation networks, cities, roads, bridges, reservoirs, and other works calling for mass labour and tight controls characterize this cultural florescence. It was capped by the crystallization of class-organized societies, supported by masses of farm families and conscript labour, defended by well-organized and well-disciplined troops, catered to by a large number of master craftsmen, and ruled and regulated by a class of priests and nobles.

Imperial expansion in the central Andes

During the last phase of the prehistoric era in the central Andes, which began around AD 1000, regional states came to be absorbed into vast empires, the best known of which was the Inca Empire. The Inca began their expansion in 1438 and completed it in 1532, by which time the Spaniards landed on the northern coast of Peru at what is now the seaport of Paita. The Inca spread their imperial bureaucracy from Ecuador to central Chile and implanted their religious beliefs and practices, as well as much of their culture and the Quechua language, in the process of empire building. Their achievement was cut short by the Spanish Conquest under Pizarro, at a time when the Inca Empire seemed on the verge of civil war.

The European conquest. A full appreciation of the force and nature of the European conquest of South America must take into consideration postcontact population trends among the indigenous societies. Today, there are at least as many people of overwhelmingly Indian ancestry as there were just prior to the European conquest, but the vast majority of these, approximately 7,000,000, live in the central Andes and represent a resurgence after a marked population decline following the conquest. Elsewhere in South America, Indian populations declined rapidly after contact with Europeans and, for the most part, have not increased appreciably since. This loss of Indian populations is related directly to the intensity of European exploitation and the density of the native populations in question, two principal factors in adjustments during the colonial period.

Population decline was heaviest along the South American coastlines and major rivers, where Indian concentrations were greatest. Along the coasts of Brazil, the Guianas, Venezuela, Colombia, and Ecuador, where Europeans came in great force, the Indians were killed in large numbers, died in the course of enslavement, succumbed to new diseases, or fled into the hinterlands in depleted numbers. Conditions were similar along the great river systems, where native populations declined sharply in the first decades after coming into contact with Europeans, their places being filled in the labour pools of colonial society by African slaves, who have made a great contribution to South America's mixed population.

Spanish influences

Indians who survived European intrusions are those small communities in the marginal, unattractive areas scarcely touched by soldiers and settlers. South of the tropical-forest area, in Argentina and Uruguay, where Indian populations were small and scattered, the coastal groups were again the first to succumb to conquest. In the Gran Chaco, resistance to Spanish settlement was fierce and temporarily successful, but, in time, these Indians were nearly wiped out by disease in mission centres and elsewhere, and the survivors were absorbed into the gaucho population that developed along with Argentine cattle raising. In Chile, the Atacama and Diaguita Indians were rapidly suppressed and absorbed by the Spaniards, as were the northern Araucanians (Picunche). The southern Araucanians (Mapuche and Huilliche) held out against white subjugation and developed a military organization to defend their heartland until the latter decades of the 19th century. The Indians of southernmost South America—the Puelche, Tehuelche, Ona, Yaghan, and Alacaluf—have become virtually extinct.

In contrast to the rest of South America, the highland populations of the Andes are today larger than at the time of conquest. They have maintained great cultural stability, have survived epidemics, and continue to live in small farming and pastoral communities established centuries ago. Their population is steadily and rapidly increasing, and there is great population pressure on arable land, which constitutes a national problem in Bolivia and Peru.

Effects of colonialism. The kinds of changes induced by European conquest varied according to the intensity of settlement and exploitation, the density and organization of Indian populations, and the ecological adjustments made by the conquerors. Three examples of these variables may serve to indicate general trends.

Peru. Inca culture and society were deeply affected by the Spanish Conquest and the intensive settlement that followed. Spanish patterns of bureaucratic government replaced those of the Inca Empire, land use and ownership changed radically, tribute and forced labour threatened the agricultural base of the old society, ancient deities succumbed to Roman Catholicism, community and domestic life were geared to the demands of the new colonial regime.

Economic and religious effects of European influences

Inca agriculture underwent great change through the introduction of European crops demanded by Spanish overlords. Indians were parcelled out among the settlers as tribute producers, menial labourers, and house servants. The abuses of exploitation were so great that very quickly most of the land was alienated from the Indians, who became a large, landless, and rootless population available for conscript labour in service of the colony.

The Spaniards imposed the Roman Catholic religion and tried to stamp out native beliefs and practices, a work of long duration that has not been wholly successful. Although the Inca state religion was totally suppressed with relative ease, an almost incalculable number of cults of lesser deities persisted in the villages.

The Inca upper classes were most readily assimilated into Spanish colonial society, whereas the agricultural masses retained much of their traditional culture. The native nobles entered the administrative ranks of colonial society and adopted Spanish dress and other customs. Artisans, servants, and others in direct contact with the settlers also became rapidly acculturated to a colonial way of life. Where native communities remained outside the main force of the colonial economy and where communal land was retained, the traditional culture was preserved somewhat intact, with customs of land use, ownership, family organization, marriage practices, and some home industries surviving into the 20th century. Today, villages have economic links with the cities through the production of marketable crops and may now be considered as peasant communities in a national economy.

Chile. The Spaniards conquered the northern half of Chile several years after having become established in Peru. They brought the Picunche under their control with relative ease by 1544 and used them to placer mine gold in the rivers, perform agricultural labour on settlers' farms, and build and provide services in colonial towns, cities, and military outposts.

In response to the colonists' demands for more Indian labour, Spanish troops attempted to conquer the southern Araucanians, the Mapuche and Huilliche. These Indians rebelled against harsh treatment at the hands of the Spaniards and succeeded in burning all their outposts and settlements and driving them north again. The history of northern Chile, after that, is one of peaceful colonization and the assimilation of the Indian population into a colonial labour force. Mapuche–Huilliche territory, however, remained a frontier zone for centuries. The Mapuche and Huilliche were placed on reservations after they sued for peace in 1884. Today, they continue to live a difficult life in reservation territory. Their 300,000 population gives them numerical superiority in several Chilean provinces, where they are beginning to take a more active political role in their own behalf.

Panama. The Chocó Indians of the tropical forests of Provincia de Darién and nearby Colombia survived the Spanish intrusion because they had nothing of value to the Europeans and were bypassed. In turn, the Chocó were not especially warlike and avoided the dangers of contact.

The Chocó retain many of their traditional values and ways of life. They emphasize magical curing, observe age-old marriage practices, and live in pole and thatch houses built on pilings along rivers, where they have small groves of plantains and also grow manioc, cacao, and other tropical crops in jungle clearings. Most Chocó have no knowledge of Spanish or of Catholicism. Men wear breechclouts and women wear cotton skirts and some jewelry. They are by no means integrated into national life and prefer to live apart in the densely forested areas where they may hunt and fish unmolested.

BIBLIOGRAPHY. JULIAN H. STEWARD (ed.), *Handbook of South American Indians*, 7 vol. (1946–59), is a monumental compilation of articles on South American ethnography, archaeology, physical anthropology, and languages. JULIAN H. STEWARD and LOUIS C. FARON, *Native Peoples of South America* (1959), is a synthesis and updating of the *Handbook* written in a consistent theoretical framework. LOUIS C. FARON, *Mapuche Social Structure* (1961), is a study of the reintegration of these people of Chile since the time of conquest.

(L.C.F.)

South American Tropical Forest Cultures

Problems of generalization

The tribal cultures of South America are so various that they cannot be adequately summarized in a brief space. The mosaic is baffling in its complexity: the cultures have interpenetrated one another as a result of constant migratory movements and through intertribal relations, leading to the obliteration of formerly significant differences, and to new cultural systems made up of elements of heterogeneous origin. Hundreds of languages, in very irregular geographic distribution, with innumerable dialects, are or have been spoken in the tropical area of South America. Thus, only the broadest generalizations can be made; one can mention certain cultural manifestations that are present in a great number of groups, even though varying in their actual expression, and illustrate them with specific examples—but always with the qualification that in a neighbouring tribe or group a distinctly contrasting idea or institution may exist.

The innumerable native peoples differ in their patterns of adaptation to their natural environment. Whether they live in the rain forest, in the gallery forests lining the rivers, in the arid savannas, or in the swamps, however, they share a common cultural background; they often combine fishing and hunting with rudimentary farming. Most are relatively sedentary, but some are nomadic or semi-nomadic. Greater differences are sometimes found among neighbouring groups living in the same forest than between some forest and savanna peoples. And some tribes, when migrating to open areas, maintain to a great extent the forest characteristics of their culture.

On the banks of the great rivers but also in zones between the forest and the savanna live tribes who gain



Distribution of South American tropical forest cultures.

their subsistence from farming and fishing. Hunters and gatherers, almost all of whom also practice some farming, have settled near the heads of rivers, in open land, or in gallery forests.

Tribes speaking related languages are scattered over a large part of the continent. The tribes of the Arawak and the Carib linguistic families are most numerous in the Guianas (French Guiana, Guyana, Surinam, and the adjacent regions of Venezuela and Brazil) as well as in other parts of the northern Amazon, but the former have representatives as far south as the Chaco and the latter as far south as the upper Xingu. The Tupí tribes extend to the south of the Amazon valley.

The Ge family includes groups most of which are located in the semi-arid lands of central Brazil. In the extreme northwest of Brazil and in the jungles of eastern Peru and Bolivia live the Pano tribes. The Jívaro of Ecuador are famous headhunters. They cut off the enemy's head, separate the soft part from the skull, and, with the help of hot sand, reduce it to the size of a fist without altering the physiognomy. They attribute great magical power to these trophies, or *tsantsa*.

CULTURAL PATTERNS

A characteristic feature of the tropical forest cultures is their combination of farming with hunting, fishing, and gathering. Before the arrival of the white man, the Indians of the tropical forest had no domestic animals except the dog. These people did not write or erect stone buildings as did the Indians of Middle America. They did not form states with centralized political organizations. They had no castes of warriors or priests. Their utensils and instruments were almost all of vegetable or animal origin, since in large sections of the area stones for making axes, arrowheads, and other objects were quite scarce. One finds evidence of metalwork only in the regions near the Andean civilizations, although objects of copper and other metals occasionally found their way across the continent, through channels of trade.

Social organization. In almost all of the tropical forest area the population density was low, averaging probably less than one person per square mile.

Populous centres existed only along the coast and the main rivers, particularly the Amazon; the latter, according to reports by early white explorers, was fringed with Indian villages. For the most part, however, the Indians were dispersed throughout the vast territory in innumerable tribes and tribelets. This is why a classification by languages and cultures gives only a vague idea of the complex picture of the forest populations. Peoples having the same dialect and culture might exist as separate groups, even as enemies, like the extinct Tupí of the

Brazilian coast. While some Indians considered themselves primarily members of their local group, others, like the Xerente (Sherenté), gave greater value to a common language and culture than to village divisions. But differences in dialect and culture often imposed obstacles to the recognition of tribal solidarity.

There were no permanent political associations or confederations encompassing tribes of different languages and cultures. At most, some tribes might form ephemeral confederations for warfare against a common enemy. Certain close relations sometimes existed between groups of diverse origin, especially through tribal intermarriage. The best known examples are along the Río Negro in northwest Brazil, where numerous populations, mostly Arawak and Tucano, are united in a vast network of interethnic relations. At the headwaters of the Xingu, a complex system of intertribal institutions also exists among formerly autonomous groups.

Marriage customs

Few tropical forest tribes are strictly monogamous. Marriage with two or more women is usually restricted, however, to chiefs and other men of prestige. Polygamy is perhaps most accentuated among the Jívaro, where headhunting has killed off many of the men; it frequently takes the form of marriage with two or more sisters. Examples of polyandry are rare.

The choice of a partner is sometimes limited by the division of the tribe into clans and segments to which an individual belongs by heredity and within which marriage is prohibited. In some cases, for example in the Guianas and in the Río Negro region, the individual must find a mate outside his village or even outside his tribe (exogamy). The Terena of the southern Mato Grosso divide themselves into endogamous groups: the man and wife must come from the same group, called by ethnologists a moiety. Marriage between cross cousins, that is, between children of siblings of different sex, is considered ideal in many tribes; that of parallel cousins, children of siblings of the same sex, is frequently prohibited.

Kinship groups and household communities are based predominantly on the principle of lineage, that is, on relation through either the male or female line. Communities of extended patrilineal families were typical of the Tupí-Guaraní. In many Amazon tribes and in others farther north, the lineages or groups of lineages are patrilineal exogamous clans. Tribes with matrilineal clans, although less numerous, can be found throughout South America. In some tribes the clans number 40 or more, as among the Mundurukú; they are generally organized into two groups so that the whole tribe comprises two exogamous moieties. The duel divisions of the Ge Indians, often not related to kinship and marriage, are mainly ceremonial.

The tropical forest cultures do not exhibit much social stratification. When there is a servile element, it is normally made up of ethnic outsiders who do not constitute a class as such. War captives may be reduced to slavery, as among the northern Carib and Arawak, the Huitoto, and the Mundurukú. Among the extinct Tupí of the Brazilian coast, slavery was the fate of those destined for ritual sacrifice. In many cases, chiefly among the northern Carib, slavery has primarily an economic function: the captives form a servile group known as *peito*—the same term that is applied to a fiancé during the period in which he is obliged to work for his future father-in-law. The Rucuyen, a Carib tribe of French Guiana, for some time maintained in servitude a great number of the Oyampī, their Tupí neighbours. In the northwest Amazon valley, Arawak and Tucano tribes hunt and enslave Makú men, who are forced to work in their gardens; the Makú women and children are used as domestic servants.

A tendency to form a class of nobility has been found in many Arawak groups, who not uncommonly impose themselves over other tribes by means of intermarriage, especially among families of chiefs. In some regions, relatives of chiefs constitute a kind of nobility. In tribes divided into clans, it is common to attribute superior status to a certain clan or even to scale them in hierarchic order. Nevertheless, the local or tribal community is essentially egalitarian.

Education and socialization

The children learn through play and imitation. Boys acquire skill in the use of weapons by practicing with small bows or blowguns made by their fathers. Girls learn to cook in little clay pots and to weave on small looms; they help their mothers in the preparation of manioc flour. The young also participate in the general religious life. The transmission of moral standards is rarely of a formal nature, and there is little punishment or repression.

The institution of the couvade is found throughout the forest culture. The father of a newborn infant must observe a rigorous diet for a week or so after the infant's birth. It is based on the idea of a mystical relation between the father and the child.

Puberty rites are often quite elaborate. In many tribes, such as the Guaraní, the symbol of masculine maturity is the labret, an ornament worn in a perforation of the lip; the ritual is preceded by an instruction period during which the boys, isolated from the community, learn the religious chants and dances, and it culminates with the perforation of their lower lips. Initiation rites may be limited to boys or to girls or may be for both sexes. The initiation of the boys is generally done collectively for those who have reached the eve of sexual maturity, while that of the girls is normally held individually on the occasion of the first menstruation.

In many of the Indian cultures these rites take a central position among other important ceremonies such as funerals and fertility rites. In the Guianas and the northwest Amazon region, the initiation of the boys is very complex. The Yurupary celebration inducts the boys into the secret society of mature men. Special rites are revealed to them; they are shown the sacred trumpets or the masks representing ancestral spirits. They are subjected to violent whippings, which they must tolerate without the least expression of pain. In the Guianas, the ritual torture consists of the stings of hornets or the bites of poisonous ants. The girls' initiation, generally more developed in the Amazon area near the Andes, is also frequently accompanied by difficult tests. Among the Tikuna (Tucuna) and other Amazonian groups, all the hair of the girl is pulled out; its regrowth symbolizes the emergence of a new adult personality.

The initiation of the boys assumes great importance in the social structure of some Ge groups of central Brazil, whose complex of rites begins at ten years of age and continues in cycles until 20. In one such tribe, the Xerente, candidates spend three years isolated from community life preparing themselves for manhood. In these Ge groups, those who have been initiated together form a distinct set of persons who feel united the rest of their lives.

Economic systems. Most of the tropical forest Indians are neither entirely sedentary nor entirely nomadic. Some wandering bands do not remain in the same place for more than a few days. Some farming populations are more or less attached to the earth. But even the latter make seasonal moves, especially those in semi-arid regions. The semi-nomadic tribes live in villages during the rainy season and go hunting in dry spells—*e.g.*, the Xavante and other Ge—or break up into little bands for gathering, as do the Nambikwara (Nambicuara). The Karajá (Carajá) of the Araguaia build their villages in rows of houses on high ground near the river, but in the dry season they move down to the long beaches. Most of the villages of the tropical forest farmers are not permanent; after some years they have to move because of soil exhaustion.

Housing

While the bands of gatherers rarely exceed a few dozen individuals, the farmers' villages have been known to include as many as 2,000. As a rule they are much smaller, dividing whenever the population becomes too large. A characteristic arrangement is the circular village of houses placed around a central plaza. This is found, for example, in the upper Xingu, in various Ge tribes, and among the Bororo of the Mato Grosso. The plan of the Bororo village, like that of the Ge, is a real map of the social structure. Each household represents a particular segment of the local group, such as an extended family or

a patrilineal or matrilineal clan. The centre of the plaza is often occupied by the men's house, where the men spend the night and the greater part of the day, and which is at times the locus of ceremonial activities.

The house reflects the economic organization and social structure. Designs range from the simple shelter of the Guayakí and the wind screens of the Nambikwara up to large communal houses containing 200 or more individuals, even the entire tribe. The latter, known as *malocas*, have been found in the Guianas, northwestern Amazonia, and in some regions farther to the south in the area of the Purus and the Guaporé rivers. The Tupinamba houses are reported to have measured up to 20 metres in length. Houses on piles are found in marshy and swampy locations, for example among the Warao (Warrau) and other Indians of Venezuela but sometimes also among tribes that inhabit dry lands and savannas. The Mura, who live on the Madeira and Purus rivers, and the Guató of the upper Paraguay River, who spend a good part of the year on rivers and lagoons, fishing and hunting aquatic animals, have made their canoes into dwellings. At other times they live in small shacks at the water's edge.

Most houses are made of rough wood, covered with palm leaves or grass. The great circular *malocas* with conical roofs in southeastern Venezuela merit special attention for their size and solidity. Although there are no walls in the *malocas*, the space is customarily divided according to social distinctions, giving a specific place to each family and sometimes even to each of its members. The furniture is very rudimentary. Some Indians sleep on mats or on platform beds, but more of them use hammocks which are found throughout the tropical region.

A great variety of economic systems is found in the tropical forest. The tribes cannot accurately be classified as hunters and gatherers on the one hand or as farmers on the other. The differences lie rather in the emphasis given to agriculture than in the presence or lack of it. The Guayakí of the forests of eastern Paraguay are one of the few tribes without any agriculture; they feed on wild honey and larvae, catch fish with arrows, and hunt jaguars and armadillos. The Sirionó of Bolivia and most of the Makú (a denomination that comprises rather heterogeneous Amazonian groups) are nomads who hunt, fish, and gather. A few Makú groups, however, influenced by their neighbours, have become more or less sedentary farmers. The same holds for the Shirianá and Waica of the Orinoco–Amazon headwaters.

Production and technology

The crops are chiefly bitter manioc as well as other tubers and roots, and, in the western regions, maize (corn). Some Ge tribes grow mainly sweet potatoes and yams. The forest is cleared by felling the trees (the stone axe has now been everywhere replaced by the iron axe) and, when the underbrush is dry, setting fire to it. The same plot is used for several (but never more than six) consecutive crops and then left fallow for several years until it is covered by new vegetation. The group must therefore move periodically. The slash-and-burn system does not, except in the more fertile lowlands, permit the growth of dense populations. It does, however, provide a seasonal food surplus that might in many cases, considering the available techniques, be increased. But the Indian has no incentive to store up goods in a generally egualitarian society, since goods are not a source of prestige.

The tropical forest Indians are highly inventive. They have developed many types of harpoons, arrows, traps, snares, and blowguns. In fishing they employ a variety of drugs that stun or kill the fish without making them inedible. The bow and arrow are today known everywhere; in some Amazon regions they have replaced the spear-thrower, a device still in use in certain western tribes. The bow and arrow are the principal weapons of warfare, although some groups fight with clubs and lances.

The techniques of basketry have a wealth of variations, mainly in the Guianas; the northwest Amazon region, and among the Ge peoples. Along with many kinds of baskets and hampers, these folk plait sifters, traps, fans, mats, and other household articles out of palm leaves and shafts of *taquara*, or bamboo.

The potter's wheel was traditionally unknown, but coiled ceramics reached a high degree of development, particularly among the Arawak and Pano tribes. Among nomadic groups pottery is either nonexistent or very rudimentary; instead, the nomads use gourds, calabashes, baskets, and fibre pouches.

Spinning and weaving, though well-known, remain at an elementary level since most tropical forest Indians, instead of dressing, prefer to paint the body and to embellish it with all sorts of adornments. From cotton, growing wild or planted, they make tunics, as well as belts of various types, skirts, and particularly hammocks. They use simple spindles, which they whirl like tops. The most common loom is the heddle loom: the threads of the weft separated by heddles, are wound around a vertical frame. In regions close to the Andes, especially in eastern Bolivia, the Indians make cloth of beaten bark.

The roots of the manioc or cassava plant is a staple of the Indian diet, and its processing requires a number of implements including baskets and sifters, graters made of planks with little stones embedded in them, the *tipiti* (a plaited cylinder used to squeeze the prussic acid from the grated pulp), great clay pots for preparing the flour, and earthen fryers for making flat cakes.

Land is generally owned by the group occupying or exploiting it—a band, a village, or a clan—and parcelled out to families or other small units for hunting, fishing, or planting. Collective tribal land or territory exists only in rare cases, when the solidarity between the various groups of a people is particularly strong. There are rigorous norms for the distribution of game among the hunter's family and among other families to which he is associated by certain ties; the hunter himself may receive a rather small share. Cleared land almost always belongs to the family using it, but when necessary others may have access to its products. Generosity is greatly valued: the generous person is good, the miser bad. This also holds for intertribal relations, when gifts are exchanged on the occasion of visits or celebrations.

Weapons and household utensils are the property of individual men and women, but canoes and other objects used collectively are not. Body adornments generally belong to the wearer. Intangible property may belong to the clan or other social unit, but it may also be individually owned, as in the case of the name or ritual functions among Ge tribes, and magical–religious chants among the Guaraní.

Brisk trade among tribes is carried on in parts of the Guianas, in northwest Amazonia, and in upper Xingu. Indians of the upper Orinoco export *urucu*, a red dye, to groups living downriver. The Arawak frequently trade ceramic wares produced by their women; they also supply blowguns in exchange for poisonous curare and barter manioc graters. Carib tribes often trade cotton products. Some groups specialize in the manufacture of canoes, which are much in demand by neighbouring groups. The most complex trading system is that of the upper Xingu; it includes a dozen tribes, each with its own products. Commerce contributes significantly toward reducing cultural differences among the tribes, the more so because it is accompanied at times by ceremonial activities through which religious ideas and practices, as well as elements of social organization, are transmitted from one tribe to another.

Belief and aesthetic systems. The jungle Indians believe that their well-being depends on being able to control innumerable supernatural powers, which in personal or impersonal form permeate or inhabit objects, living beings, and nature in general. Through shamanistic rites or collective ceremonies, man must encourage and maintain his harmonious integration in the universe, controlling the forces that govern it; their beneficial or harmful effects are largely determined by human action. In most of the cultures, magical measures and precautions are more important than the religious cult as such. The strength and health of the body, the normal growth of children, the capacity to procreate, and even psychic qualities are obtained by magical means. For the individual these means may include the perforation of the lips,

Magic

nasal septum, or ear lobes, the painting of the body, and the use of various adornments. A little stick passed through the nasal septum, such as that used by the Pawumwa of the Guaporé River, prevents sickness. The hunter or fisherman, in order to be successful and not to be *panema* (unlucky), as they say in many Amazonian regions, takes precautions such as scarring his arms or abstaining from certain foods. The magical devices of the hunter, the fisherman, and the warrior are considered much more important than their ability. Arrows must be treated by rubbing with a certain drug, since their magical attributes are believed to be more effective than their technical properties.

Stimulants and narcotics are of great importance in the magic and religious practices of most tropical forest Indians. Secular use of drugs is much rarer. Tobacco is known by almost all tribes. The Tupinamba shaman fumigates his rattle with tobacco, which he believes contains an animating principle that confers on the rattle the faculty of "speaking," that is, of revealing the future. Alcoholic beverages, consumed mainly in religious festivals, are obtained by fermentation of manioc, corn, and other plants. They are unknown among the Ge, in the upper Xingu, and in some regions of Bolivia and Ecuador. Coca leaves are chewed, especially in the sub-Andean regions. Infusion of maté is taken in the Paraguay area, as well as by the Jívaro and other groups of Ecuador. Hallucinogens are used mainly in the Amazon–Orinoco area; they include species of *Banisteriopsis* (a tropical liana), from which is made a potion that produces visions. In certain tribes the use of this drug is restricted to shamanistic practices; in others, as in the Uaupés River area, it is an essential element in religious festivals in which the community revives its mythic tradition. Other narcotics in ritual use, among them the *yopo*, or *paricá* (*Piptadenia*), known among many northern groups, are often breathed in the form of snuff, which partners blow into one another's nostrils; the Omagua of the upper Amazon used it as an enema.

Some magical practices are reserved for the shaman, who acquires his status by natural endowment, by inspiration, by apprenticeship, or by painful initiation. The shaman may practice medicine, perform magic rites, and lead religious ceremonies. Rarely, however, is he a priest in the usual sense of the term. In many groups his influence is superior to that of the political chief; in some, as among the Guaraní, the two roles may coincide. Not uncommonly, his influence continues even after his death: in the Guianas and elsewhere, his soul becomes an auxiliary spirit of his living colleagues, helping them in their curing practices and in the control of harmful spirits; among the Rucuyen, the bodies of common individuals were cremated, while that of the shaman was kept in a special place so that his soul might live on.

The shaman as healer

In curing the sick, the shaman must remove the object causing the sickness: a small stone, a leaf, an insect, any substance that has been sent through the black magic of an evildoer. The cure consists of massages, suction, blowing, and fumigation. If the illness comes from loss of the soul, the shaman must search for and recover it. If it comes from a bad spirit, he tries to overcome the evil influence with the help of one or more auxiliary spirits.

The soul has its seat in the bones, the heart, the wrist, or in other parts of the body. Some Indians believe that two or more souls are responsible for various vital functions. One finds also the idea of a purely spiritual soul. The Guaraní believe that man has an animal soul governing his temperament and his instinctive reactions but that he also has a second, spiritual one, sent by a divinity at the moment of conception. Thanks to his second soul, man thinks, speaks, and is capable of noble sentiments. After death this second soul returns to live among the gods, while the other soul wanders the Earth as a ghost menacing the living.

Nature is believed to be peopled by demons and spirits that are beneficial or malevolent, depending on man's behaviour. Besides the soul that gives life to every living thing, many plants and animals have a "mother" or "master," as do manioc, maize, and game animals.

The mythology of almost all tribes includes a creator of the universe and of man. This creator seldom sustains interest in his handiwork, and thus there is usually no cult attached to him. Social institutions, customs, knowledge, techniques, and cultivated plants are deeds or gifts of a culture hero or a pair of them, sometimes twin brothers who may represent the sun and moon. A number of myths are told about these figures; sometimes the pair consists of a hero and a trickster who opposes him.

Ceremonial practices vary, depending on the tribe and its way of life. Some great collective ceremonies have been associated with war, as among the northern Carib and the coastal Tupí, both famous for cannibalism, and the headhunting Mundurukú and Jívaro. Ceremonies are often believed to be indispensable for regulating the course of the sun and the moon, the sequence of the seasons, the fertility of plants, the procreation of animals, and the very continuity of human life. Their objective may also be to commune with the dead or with mythical ancestors; when they are connected with the disposal of the dead, they are at the same time passage rites, by means of which the spirits of the dead are made harmless. Among the Guaraní, most religious ceremonies mean profound spiritual communion with the gods.

Corpses are commonly disposed of by ground burial within or without the house. Urn burial has also been known, especially among Tupí groups; some groups have been known to unearth bones, clean them, and then rebury them. The Tarariu (Tarairiu) of northeastern Brazil and some Pano broiled the flesh of their dead and mixed the pulverized bones and hair with water or with a manioc-base beverage that they drank. Tribes of the Caribbean coast, after drying the body by fire, allowed it to decompose and later added the powder to a drink. In other northern regions, one still finds the custom of cremating the cadaver and consuming the charred and crushed bones in a banana mush.

Aesthetic expression

Artistic efforts are most commonly applied to decoration, whether of the human body, objects of practical or ritual use, or even houses. The most common body adornments are paint and feather ornaments. Tattooing has also been practiced, especially among the Mundurukú and many Arawak tribes. Magical and religious ideas are usually expressed in these adornments. The Carib tribes of the Guianas and some Tupí were outstanding in featherwork. The plumed mantles of the Tupinamba, the delicate and elaborate adornments of the Caapor of Maranhão state, and the rich and varied ones of the Mundurukú are much celebrated.

The design of ornaments is almost always geometrical, with characteristic patterns for particular tribes; the styles vary with the cultural areas.

Masks, generally used in ceremonial dances, are restricted to the tribes of certain areas: the Cágaba of northern Colombia, the Tikuna and others of the northwestern Amazon, the Tucuna and their neighbours of the upper Amazon, the Guartegaya and Amniapé (Amniepe) of the upper Madeira, the tribes of the upper Xingu, the Karajá and the Tapirapé of the Araguáia River area, some Ge of central Brazil, and the Chiriguano of southern Bolivia. The masks represent the spirits of plants, fish, and other animals, as well as mythical heroes and divinities. They are highly stylized in form but, on occasion, naturalistic in expression.

The Waurá women of the upper Xingu are famous for their pots and animal-shaped bowls. Of the historic tribes, the Tapajó of the Amazon had the richest style in ceramics, excelled only by the archaeological remains of the Ilha de Marajó. Among some groups in the Guianas and western Amazonia, artistic activity includes wood carving.

THE IMPACT OF WHITE CIVILIZATION

Disruption and crisis

The great majority of the Indian groups are closed societies, despite intense intertribal relations in certain areas. They lack institutions that would enable them to become part of a larger, all-embracing, and technologically superior society. Their relations with the whites have placed them in a situation of dependency that usually ends with their breaking up. The white man's prohibition of war-

fare, headhunting, cannibalism, polygamy, and other institutions that have profound meaning in tribal life sets in motion a process of social disorganization. In addition, numerous tribes have been extinguished by violent destruction, slavery, the loss of lands required for subsistence, epidemics, and by marriage with whites and blacks. Innovations introduced by the whites often have harmful effects: *e.g.*, iron utensils, which not only subject the Indians to those who supply them but also change the traditional division of labour in tribal society, and clothing, which alters personal hygiene and makes its wearers susceptible to new diseases.

Contact with the white man creates a profound crisis for the tribal leadership. Either the chief of the group is deprived of his authority, since the conditions for realizing the values essential to tribal life no longer exist, or he becomes despotic and a tool of the whites, using his power to benefit himself at the expense of his community.

Agricultural tribes are sometimes able to adapt to the new conditions for a time by trading their products, especially manioc flour. The sale of products such as fur skins, babassu nuts, copaiva oils, and carnauba wax helps in certain cases, as with the Tenetehara of Maranhão state, to maintain economic stability without breaking up the community organization. This is impossible, however, when the Indian undertakes to collect rubber for commercial firms, since this obliges the tribe to split into family units and to spread over vast areas; the result is an enormous cultural impoverishment. The transformation of the Indian into a labourer generally leads to the rupture of tribal bonds, much misery, and the disappearance of the tribes as ethnic entities.

There have been cases in which the Indian successfully integrated into the regional economic system as a paid worker or even as an independent producer. The Terena, an Arawak group of southern Mato Grosso, work on cattle breeding farms, an activity they learned long ago while vassals of the Guaycurú, who had become horse breeders after the Spanish conquest. The Goajiro of Colombia, another Arawak group, own great herds of cattle.

The disruption and crisis that follow inevitably upon contact with the whites, are, however, less serious when the Indian has had earlier contact with a hybrid population carrying a cultural system that already incorporates many of the elements of civilization. These mixed cultures, such as those on the Brazilian–Paraguayan frontier and in some parts of Maranhão state, act as a kind of sieve between the system of tribal life and that of the white invader. In the past such cultures took numerous solutions, especially of an adaptive kind, from the Indian culture, helping to give the Indian a feeling of worth when facing the invader. Today he no longer has this privilege, since the hybrid cultures already possess whatever Indian elements they can use.

Messianic movements

The cultural crisis the Indians undergo in contact with the whites brings about sporadic messianic outbreaks. Since the Indians face a problem for which there often seems to be no solution, they may appeal to the supernatural and wait for a miracle to happen. They hope for a return to the "lost paradise," that is, to the old life without the white man, whose expulsion or destruction may be part of the miracle. At the same time they want to avail themselves of some of the advantages of civilized life. Insofar as civilization seems an ideal impossible to realize without recourse to supernatural aid, because of the Indian's economic and social inferiority, the messianic miracle in many cases takes the form of a social and cultural revolution: in the new era the Indian will dominate over the white man and will have all those things in the civilized world that to him symbolize superior status.

In certain regions of the tropical forest these movements appear from time to time. Along the Río Negro in northwest Brazil, there have been several messiahs since the end of the last century. These leaders combined elements of their tribal religion with teachings and rites of Christian origin, although the predominant note was always hostility to the whites. Such movements have also occurred among the Tikuna of the upper Amazon; in one in 1956 the leaders proclaimed, among other things, that a city would appear suddenly in the middle of the forest, lighted by electricity and providing all the comforts of modern civilization. In 1963 the Canela, a Ge tribe of Maranhão state, had a messianic movement announcing that, when the new day came, the civilized people would be obliged to live in the forest or in the savanna, hunting with bow and arrow, while the Indians would become rich farmers. In this, as in other cases, the miracle was to be brought about by the great hero of tribal myth. The Guaraní of Paraguay and adjacent Brazilian regions are most famous for their frequent messianic movements, the fundamental myth of which is a cataclysm that will destroy the world; the Indian will find salvation in a distant paradise called the Land Without Evils. Probably the messianic tradition of the Guaraní dates from before the coming of the whites, but it seems to have undergone great expansion since then.

BIBLIOGRAPHY. JULIAN H. STEWARD (ed.), *Handbook of South American Indians*, 7 vol. (1946–59), is a noteworthy systematic description and interpretation of the native cultures of South America, including those of the tropical forest; an abridged version is JULIAN H. STEWARD and LOUIS C. FARON, *Native Peoples of South America* (1959). JANICE H. HOPPER (ed.), *Indians of Brazil in the Twentieth Century* (1967), is a collection of essays by American and Brazilian specialists. BETTY J. MEGGERS, *Amazonia: Man and Culture in a Counterfeit Paradise* (1971), is a comparative study of Amazonian tribes and their adaptation to the natural environment. More specialized works include GERARDO REICHEL-DOLMATOFF, *Desana: simbolismo de los indios tukano del Vaupés* (1968; Eng. trans., *Amazonian Cosmos: The Sexual and Religious Symbolism of the Tukano Indians*, 1971); IRVING GOLDMAN, *The Cubeo: Indians of the Northwest Amazon* (1963); and ALLAN R. HOLMBERG, *Nomads of the Long Bow: The Siriono of Eastern Bolivia* (1950). The complex social organization of the Ge tribes is studied in CURT NIMUENDAJU, *The Apinayé* (1939, reprinted 1967), *The Šerente* (1942), and *The Eastern Timbira* (1946); and in DAVID MAYBURY-LEWIS, *Akwẽ-Shavante Society* (1967). A general view of the religious ideas of indigenous peoples may be found in RAFAEL KARSTEN, *Studies in the Religion of the South-American Indians East of the Andes* (1964). Acculturation among tribes in contact with whites is discussed in CHARLES WAGLEY and EDUARDO GALVAO, *The Tenetehara Indians of Brazil: A Culture in Transition* (1949, reprinted 1969); ROBERT F. MURPHY, *Mundurucú Religion* (1958) and *Headhunter's Heritage: Social and Economic Change Among the Mundurucú Indians* (1960); and JAMES B. WATSON, *Cayuá Culture Change: A Study in Acculturation and Methodology* (1952).

(Eg.S.)

South Asian Cultures

The subcontinent of South Asia lies south of the Himalayas and Afghanistan. It comprises the countries of India, Pakistan, Bangladesh, Sri Lanka (Ceylon), Nepal, Sikkim, and Bhutan. While these countries together may be thought of as constituting a single region, they contain a multitude of ethnic groups. The peoples of the Himalayas have Mongoloid features. The upper-caste groups of the Indo-Gangetic Plains are basically Mediterranean in appearance. The tribal and semi-tribal populations of middle and southern India have dark skins, flat noses, and wavy or curly hair. The Pashtun (Pathan), Kashmiri, and Sindhi of the northwest are tall with long heads, white complexions, and wavy hair. The Kadar of Cochin, the Andamanese and Onge of the Andaman Islands, and the Vedda of Sri Lanka have the Negrito characteristics of dark skins, small stature, and frizzly hair. The tribes of the jungles of Bihār, Orissa, and Madhya Pradesh share certain characteristics with the aboriginals of Australia and are sometimes referred to as Proto-Australoid; some of them (the Oraon and the various tribes of Gonds) speak Dravidian languages, while others (the Mundas, the Santāl, and the Ho) speak Austro-Asiatic languages.

The South Asian population may be divided into tribal and nontribal peoples. The tribes are somewhat detached from the mainstream of civilization, having been isolated in the forests and hills. They constitute about 6 or 7 percent of the total population of India and varying proportions of the populations of the other countries.



Selected ethnic communities of South Asia.

A vast number of languages are spoken in South Asia, belonging to four different families of languages (*see* INDO-IRANIAN LANGUAGES; DRAVIDIAN LANGUAGES; AUSTRO-ASIATIC LANGUAGES; SINO-TIBETAN LANGUAGES). In India alone there are more than 100 separate languages, of which 15 have been designated as national languages.

The South Asians are predominantly religious in their outlook. Their religious orientation pervades every sphere, from social and economic life to science, literature, and art. Whether one thinks of Hinduism in India, Islām in Pakistan, Buddhism in the Himalayan states, or tribal religions in the hilly areas, one cannot escape the fact that the traditional way of life is largely determined by ritual and belief. Along with this religious outlook there appears to go a tendency to fatalism, a sense that every course of action as well as its outcome is up to the gods. This is true not only of Hindus and Buddhists but even of Muslims; as S.M.H. Zaidi observed in *The Village Culture in Transition: A Study of East Pakistan Rural Society* (1970), the Muslims "have a predominantly fatalistic attitude to change, believing that no human hand without the will of God can change anything."

Another characteristic of the South Asian peoples is that they are predominantly peasants. More than 90 percent of them live in villages and depend on simple agriculture for their livelihood. The average person lives his whole life in the same village. There is much interaction between the village societies and the important religious cities such as Vārānasi (Benares), Mathurā, Ayodhyā, Gayā, Prayāga (Allāhābād), Rāmeswaram, Madurai, Cape Comorin (the southernmost tip of India), Dwārkā, and Kāmākhya.

SOCIAL ORGANIZATION

Caste groups and tribes. South Asian society is characterized by a hierarchical social system with a vast efflorescence of caste groupings. A caste group is hereditary, usually localized, and associated with a particular occupation; its members do not marry outside the group. There are complex relationships of superiority and inferiority among caste groups, including restrictions on eating and drinking with one another. At the bottom are the so-called scheduled castes, or untouchables; in India these amount to about 15 percent of the population. These groups engage in various "impure" occupations such as scavenging, sweeping, cattle herding, and leather working.

At the lower levels of society, caste groups are not easy

Tribal organization

to distinguish from tribes. Many scholars, in fact, have suggested that caste groups evolved from primitive tribal groups. One difference is that while caste groups are endogamous (the members of a group do not marry outside it), the marital rules among tribal groups are somewhat more complex. Each tribal group confines its marriage relations to the same tribe or subtribe. But each tribe is divided into clans or moieties, often named after animals, birds, or plants, and these groups are exogamous; *i.e.*, the members of a clan must marry outside it. Such division of a tribe into several social groups or clans is found both in matrilineal tribes like the Garo and Khāsi of Meghalaya and in patrilineal tribes like the Nāga, the Munda, the Santāl, the Bhīl, and the Gond. Some tribes of southern India have a complex social organization. The Urāli, for example, who are matrilineal, have two sets of clans: brother clans, among whom marriage is prohibited, and brother-in-law clans, among whom marriage is sanctioned.

Even among the tribes of the Himalayas, where there is a great tolerance of casual sex relations, both premarital and extramarital, the rule of clan exogamy is strictly followed. Among the Sherpa of Nepal, the term for clan is *ru* ("bone"), implying that the members belong to the same father's bone or blood. A recent study of this tribe failed to reveal even a fleeting amorous attachment between members of the same clan and no marriage between members of brother clans. The Dimasā Kachāri of Assam have different sets of clans for different sexes: there are 40 male clans and 42 female clans, all practicing the system of double descent, by which the male child is given the clanship of his father while the female child inherits the mother's clan. Very few tribes lack the totemic clan division. The Saora (Savara) and Māler of middle India have territorial divisions rather than clans, and their marriage relations are exogamous within these divisions. The Saoras are divided into 17 such exogamous units.

Certain tribes of middle India have clans organized in groups or moieties. Among the Gond of Bastar, for example, the moieties are composed of 96 and 69 clans. The Rāj Gond Ādilābād have four groups of clans, and the Muria have five.

While most tribes lead a peaceful life, there have been some warring groups. The Baluchi nomadic groups used to fight among themselves over pasture land, leadership rights, and women. Descriptions of the Nāga tell of their weapons, their methods of warfare, and their custom of headhunting.

Caste and social structure

The basis of the social structure in South Asia is the caste system. Although it is Hindu in origin, the caste system has had a pervasive influence on all South Asian society, whether based on tribalism, Islām, Buddhism, Sikhism or Christianity.

In Nepal both Hindus and Buddhists are subject to the code of the caste system. Among the Hindus of Nepal there are high castes such as the Brahmins, Ṭhākurs, and Chetris. The followers of Buddhism in Nepal are also divided into three main caste groups. There are the Barnas, the Udās, and a mixed caste of heterodox Buddhists. The Barnas enjoy a status equal to that of the Hindu Brahmins and observe the same rules of commensality and intermarriage. The Udās are inferior to the Barnas in the caste hierarchy. The third group of mixed castes includes numerous occupational castes (barber, carrier of light, etc.) and professional castes considered to be equivalent to the untouchables among the Hindus. The tribes also have several categories: the Gurung and Magar are at the top; the Newār are in second place followed by the Kiranti, the Khambu, the Limbu, and the Yākhas; below them are the Sūnwar and Tamang (Mūrmi), who are given approximately equal status, while the Tibetan tribes including the Sherpa are ranked slightly below them; at the bottom of the scale are the Tharu, the Thami, the Hayu, the Thakali, and numerous other minor tribes.

In Sri Lanka the Goyigama (cultivators) enjoy the highest position, while the castes of potters and washermen are ranked low. These caste groups are endogamous.

The caste hierarchy, even in the traditional system, was not entirely rigid. It was possible for caste groups to move upward, although slowly. Even tribes have been able to acquire an appropriate place in the caste hierarchy. The Rāj Gond of central India, for instance, successfully claimed for themselves the rank of Kṣatriyas (a ruling and warrior class) on the basis of their acquisition of political power. Recent researches have revealed how caste-like groups such as the Bhumij, the Mahali, the Bauri in West Bengal, the Cheros, the Kharwār, and the Manjhi in Bihār, as well as various sections of the Bhīl and the Gond, have been assigned appropriate status in the Hindu caste systems in their regions.

Islām's claim of equality for all those who profess the faith has been greatly weakened in South Asia because of caste practices. The Muslim caste system, unlike the Hindu, is not based on purity and pollution concepts; nor is it sanctioned by religious ideas. Among the Muslim castes, the highest are the sayyids and the lowest are the sweepers. In Bangladesh the landholding groups of foreign (Persian or Arabian) descent are generally ranked high, while those who do manual labour that involves ritual contamination are ranked lowest.

The hierarchical system in Pakistan may meaningfully be compared to the Hindu caste system. In Swāt in northern Pakistan there are seven social groups; they are, in order of descending rank: (1) persons of holy descent, (2) landowners and administrators, (3) priests, (4) craftsmen, (5) agricultural tenants and labourers, (6) herders, (7) the despised group.

Sikhism also, despite its emphasis on equality, has a caste structure. The Sikhs are broadly divided into Sardars and Mazhabis, the former consisting of higher caste groups and the latter of sweepers. Caste restrictions are also rigidly observed among the Christians of the west coast, particularly in marriages.

The joint family

Family and marriage. The joint family is another basic institution in the societies of South Asia. In India the joint family has existed since prehistoric times. The coming of Islām failed to modify the structure of this most ancient institution. Despite the industrial revolution and other modernizing influences, the joint family continues to exist on a large scale.

There are two types of joint families. In the northern type, the family unit includes three or four generations of males and their dependents: one's grandfather and his brothers, one's father and his brothers, one's own brothers and cousins, one's sons, and the nephews and wives of all these male relatives plus one's own unmarried sisters and daughters. The southern type of joint family includes one's mother's mother and her sisters and brothers, one's mother and her sisters and brothers, one's own brothers and sisters, one's mother's sister's sons and daughters, and the children of one's sisters. The husbands of the women of the family live in the houses of their mothers, occasionally visiting their wives and children. Throughout Nepal the members of several generations in the father's line maintain common arrangements for cooking and have joint ownership of property under a single head. As in India, in some families the relatives on the mother's side may also live with the household in an extended family. Throughout Nepal, the kitchen is regarded as a symbol of family unity.

In Bangladesh the largest unit of family organization is the patrilocal extended family, in which two or three generations may live together. The extended family includes two or more nuclear or joint families united by male kinship bonds and living in a cluster of adjacent dwellings. It may have as many as 50 or 100 members or as few as seven or eight, the average being about 20. The extended family includes a number of component property-holding units such as the nuclear family, the joint family, and the irregular family (centred around a widow or a widow and her unmarried children). As long as the father remains alive after his son's marriage, the joint family is the rule rather than the exception. It is clearly related to the system of inheritance, by which land is allotted to the son only after his father's death. The joint family is slightly less frequent in Bangladesh than in West Bengal but more common than in Pakistan.

In Sri Lanka the nuclear family prevails. A Sinhalese family generally consists of a husband and wife and their unmarried children. It leads an economically self-sufficient life under the benevolent and mildly patriarchal rule of the father-husband.

Most tribal societies in South Asia have a nuclear family.

Patterns of family organization can also be viewed in terms of the marriage relationship, the kinship system, and the position of the women in the family. The organization of the family throughout northern India is that of the patriarchal family. The traditional rule of avoiding marriage with somebody who is removed by less than seven degrees from the father and five degrees from the mother is practiced by all castes, from the highest to the lowest. In southern India a man preferably marries his maternal cross-cousin—*i.e.*, his mother's brother's daughter.

Marriage patterns

Among the tribes of the high Himalayas, such as the Sherpa, the Bhutia (Bhotia), and the Lepchā, it is common to have more than one husband, usually brothers. The Sherpa believe that a polyandrous union is desirable because it prevents the fragmentation of property and fosters solidarity among brothers. It has been observed that since such a marriage improves a woman's economic prospects and assures her the advantage of a younger husband in later years, she prefers to marry two brothers. The Sherpa style of polyandry differs from the Tibetan practice that permits any number of younger brothers to share an elder brother's wife; this is also prevalent among the Khāsa, the Toda, and the Nāyar, who allow more than two younger brothers to share the same woman.

The custom of paying a bride price is prevalent among almost all the tribes of South Asia. The bride price may be livestock, food, clothing, or even service by the bridegroom in the bride's father's house. The ease of acquiring wives and the freedom of sex relations in the tribal communities make divorce an easy process.

In Islāmic law, marriage is essentially a contract rather than a sacrament, as it is among the Hindus. The rules of marriage prescribed by the Qur'ān have been modified in parts of India by the customs of the particular region. Marriage among the Muslims as among the Hindus is an important family affair, arranged by the elders of the family. Marriage is possible between almost any boy and girl under Muslim law, even within an extended family and between cousins. While traditional Muslim law does not sanction dowry, the Hindu influence compels it. The bridegroom's family expects the bride's father to give enough in dowry to establish a home for the couple. A Muslim can divorce his wife at his own will, but no such privilege is accorded to the wife. The custom of *pardah* (wearing the veil) is observed strictly among the Muslims of South Asia.

SOCIALIZATION AND EDUCATION

The birth of a child in the societies of South Asia is considered a divine blessing. A childless woman is viewed unfavourably. In Hindu society a very important ceremony called *śrāddha*, conducted after a person's death, can be performed only by a son. It is believed that without this ceremony the departed soul is not likely to attain salvation. To the peasantry of South Asia, a male child means a pair of extra hands to help on the farm. Girls are not highly regarded; having a female child is preferred only to remaining childless.

Rites of passage

In Hindu society an individual passes through a series of rites from his conception to his death. Several rites are performed before the birth of a child to ensure his safe arrival. After the child is born, other ceremonies are conducted to provide the incarnating soul with a good body and prepare it for life.

Next are the rites known as *vidyārambha* (commencement of education), *upanayana* (the sacred-thread ceremony), *vedārambha* (commencement of Vedic recitation) and *samāvartana* (convocation). Traditionally a Hindu boy was taught by the guru in the *gurukula*, where he led a life of rigorous discipline and purity. He was then ready to undergo the ceremony of marriage and become a householder in his joint family. A man's death is mourned by the members of the family, his kinsmen, and other members of his larger communities. A series of death rituals are observed by his eldest son in cooperation with the kin groups to ensure the deceased freedom from rebirth. The Indian joint family continues to practice some of the traditional rites or *saṃskāras*, especially the birth rites, the sacred-thread ceremony, the marriage ceremony, and the death rites.

Most of the Muslim communities are converts from the Hindu religion, and there is considerable similarity in the methods of child rearing. One important difference is the Muslim ritual of circumcision that usually takes place between the ages of three and five.

The tribal communities have relatively informal relationships among kinsmen, in contrast to the highly formalized Hindu relationships. The youth dormitory so common among the tribes is the traditional institution of the unmarried boys and girls; it provides training in folk songs and dances, the moral code, customs and manners, and sexual relations. The young learn the arts of life mostly by observation and imitation.

ECONOMIC FACTORS

Villages and settlements. Most of the people of South Asia still live in villages and follow the traditional pursuits of an agricultural and pastoral economy. In the high Himalayas the tribesmen live with their flocks in summer and winter encampments, each remaining unoccupied for some months in a year. The pastoral nomads of the Baluchi tribe in Pakistan generally have only temporary encampments. The Dafla, Apa Tani, Gallong, and other tribes of Assam site their villages on the highest parts of the hills, from whence the whole valley below is visible. Many tribes site their villages in accordance with the will of God as interpreted by their shamans. Hundreds of villages of the Santāl, Munda, and Ho tribes in middle India are in the forest clearings. In southern India many tribal villages are located in dense forest, and their economy revolves around the forest. In western Rājasthān there are dispersed settlements in desert or semidesert areas.

In the deltaic plains of West Bengal and Bangladesh, the countless rivers and streams have given birth to compact, nucleated settlements. More typical of the area are clusters of houses scattered along the banks of streams and rivulets. Each cluster usually forms a separate residential unit consisting of a number of huts built around a courtyard.

The linear settlement is much in evidence in the plain areas of Orissa, Madhya Pradesh, Andhra Pradesh, Kerala, and Tamil Nadu. The abundance of river valleys and the prevalence of the *jajmānī* system (in which lower castes perform specialized artisan services for the landowners and cultivators) serve to maintain the traditional linear settlement pattern. The linear settlement is also found in fishing villages on the sea, or estuaries, or along big rivers.

Villages vary from more than 1,000 people to less than 50. The villages of the rich agricultural plain areas are usually larger, while those of the hills and forests are smaller. Some tribal villages are temporary encampments of five to ten huts providing accommodation for up to 30 persons. The tribal villages of the Hill Māler, the Hill Khariā, or the Bondo highlanders in middle India are slightly bigger, composed of 10 to 30 huts accommodating 30 to 100 persons. Large villages like those of the Waromung in Nāgāland, having a population of 1,600, are exceptional.

The villages of the settled cultivators are large and somewhat elaborate in layout. Depending on the extent and fertility of the land, the population of an average village varies from 100 to 500 persons. Hindu villages, especially those in the Indo-Gangetic Plain, may have as many as 5,000 inhabitants, divided into a number of hamlets. The central part of a village is occupied by the dominant castes, while the untouchables live in a far-removed area. Every village has a temple located near the

settlement of the high castes. The houses of the landlords are elaborate, while those of the tenants are simple. Most houses are built of mud and tiles, with courtyards and parlours.

Food getting. Almost all of South Asia lies in the monsoon belt, and its agriculture depends on the monsoon rains. The vagaries of the weather pose a great problem to the peasantry. When there is a lack of rain they turn to the rain god, Indra. In Nepal the guardian deity Machendranāth, who presides over agricultural affairs, is elaborately worshipped; the ritual of asking for rain is observed collectively by the villagers, who go from shrine to shrine in a procession, shouting, "O Lord Mahādeo, give us rain!"

Agricultural organization

Agricultural technology is still very simple. While among the forest tribes the primitive occupations of hunting, food gathering, and collecting still predominate, the basis of the South Asian economy is the plow. The Indo-Gangetic Plain, extending from the Punjab in the west to Bangladesh in the east, along with the coastal deltaic regions of eastern India and Sri Lanka, is the largest rice-growing zone in the world. In Nepal, Sikkim, and Bhutan, intensive plow cultivation is practiced in the valleys and plains, but a primitive slash-and-burn type of cultivation is found in the hills. Pastoral pursuits are supplemental to agriculture in most South Asian villages. There are, however, communities, such as those of the Pathan and Baluchi in northwestern Pakistan and of the Gūjar and Gaddi in the western Himalayan districts, whose entire economy revolves around pastoral nomadism. During the summer they move up the mountains in search of pasture, returning in wintertime to the foothills.

The economy of the tribal peoples is based on hunting and gathering. They eat edible roots, fruits, vegetables, honey, insects, fish, and various kinds of birds and game. Some tribes such as the Birhor in Bihār, the Chenchu in Andhra Pradesh, the Juang in Orissa, the Kadar in Kerala, and the Paliyan and Paniyans in Tamil Nadu depend on the forest as their primary source of food. The tribes of the Assam Hills, of Orissa, or Madhya Pradesh, and of Nepal, who practice shifting cultivation (moving from one location to another), supplement their economy with occasional hunting and gathering. Even agricultural tribes such as the Munda, the Oraon, and the Ho of the Chota Nāgpur Plateau depend on the forest to supplement their harvests.

The hunting tribes use various types of bows and arrows and trapping devices. They also have simple missiles such as the throwing stick of the Oraon and the nonreturning boomerang of the Bhīl. Slings are also used to throw stones and hardened clay balls. Spears with iron heads and wooden shafts are very common. The Buna use multipronged harpoons with barbs for fishing. The Jarawa of the Andaman Islands poison their spearheads.

There are many forms of land tenure in South Asia, ranging from common ownership by villagers to individual ownership by the head of a nuclear family. In the past, many landlords gained their property under British occupation when they were given tax-collecting rights; the tax-farming system led to a great stratification of the peasantry. At the same time, the pressure of the population upon the land produced a fragmentation of holdings in some areas. Since independence there have been various attempts at land reform.

The *jajmānī* system

A common institution of Indian village life is the Hindu *jajmānī* system, practiced all over India under different names. Each caste group within a village gives certain standardized services to the families of other castes. A member of a carpenter or a barber caste works for a particular family or group of families with which he has hereditary ties. The association passes from father to son. The system serves to reduce the need for money and to minimize competition. It also assures a stable labour supply for the dominant agricultural caste in a particular region by limiting the mobility of the lower castes. If a worker leaves his village, he must supply a substitute, usually a member of the same joint family; but the system gives him rights and advantages that make him reluctant to leave and endanger his social and economic security.

Another traditional element of economic organization is the local market. The weekly or biweekly markets are located in open areas between villages where the means of communication provide a central place for the economic exchange of the region. In the nonmonetary economy of the tribal areas of India and Bangladesh, villagers exchange their produce for clothes, utensils, spices, and other necessities. Occupational specialists such as barbers, tailors, blacksmiths, and potters provide their goods and services. The agricultural produce from the surrounding regions is collected here and sold. The intertribal market is also a centre of social, religious, and political life.

Moneylending is an integral part of the agrarian economic structure of South Asia. In Bangladesh, the so-called middle class farmers lend out their surplus money. In Indian villages the moneylenders come mainly from business castes and traders who have settled in the tribal villages.

RELIGION, MAGIC, FOLKLORE, AND ART

The worship of spirits

Life in South Asia revolves around religious belief and practice. Belief in supernatural entities is strongest among the tribal communities; they believe that disease and death, failure in hunting or cultivation, and all the other hazards of life are caused by ghosts and spirits who can be propitiated with the sacrifice of animals. A study of the religion of the Ho and other tribes of the Mundari group has shown that this belief in the power of spirits to control the nature and destiny of man is basic to their religion. The Buddhist Sherpa observe certain rites to protect themselves and their land from evil spirits. Among the Muslims of South Asia, the matrix of primitive faith is found in the worship of *pīr*s, or saints, a practice for which there is no authority in the Qur'ān.

The tribes and also the Hindus worship certain forest and tree spirits, water spirits, and other spirits in natural objects. Among the agricultural tribes, every village has in its vicinity a grove (*sarna*) reputed to be a remnant of the primitive forest left intact for the local gods when the clearing was done. This sacred grove is the chief village shrine among many Mundari and Dravidian tribes of middle India. These tribes have special festivals associated with trees. In the festival of Karma, a branch of a sacred tree is cut and planted in the village assembly ground, and the unmarried young of both sexes dance around it. The Karma festival is also observed by Hindus in north Indian villages.

The Hindus consider the pipal tree (*Ficus religiosa*) sacred, believing that its trunk provides a habitation for Brahmā, its twigs for Viṣṇu (Vishnu), and its leaves for other gods. Water is poured on its roots by pious people after their morning bath. The bel tree (*Aegle marmelos*) is associated with the Śiva (Shiva) cult. The small plant tulsi (*Ocimum sanctum*) is associated with Viṣṇu, and its leaves are offered to him during worship. The presence of a tulsi plant in the courtyard is the mark of a pious Hindu house.

The worship of rivers and springs is of great significance among the tribal people, who believe that the indwelling spirit causes the water to move. The tribes of Chota Nāgpur offer sacrifices to the water spirits. The people of the plain consider some of the rivers, particularly the Ganges, to be sacred. The spirit of the Ganges is recognized by many Hindus as one of the wives of Śiva, and there are periodic rituals involving bathing in the Ganges and other rivers. The non-Aryan hill tribes worship mountains.

Worship of the sun has been common among both the tribes and the Hindus. The supreme spirit of the tribes of middle India is Singh Bonga (Sun God). Some temples have been erected to the sun, notably those at Konārak near Puri in Orissa and at Deo near Gayā.

Many tribal peoples and Hindus also worship the mother earth. Pious Hindus offer a prayer to her before walking in the early morning, and some farmers offer a sacrifice in her name before digging the land. Hindu mythology and folklore hold that the mother earth rests on the hood of a cobra; he is also considered sacred, and a

festival is devoted to him. Some deities are associated with certain diseases, as the goddesses of smallpox (Sītalā) and of cholera (Olābibi), who are worshipped by singing devotional songs in their praise.

According to the scriptures, the Hindu pantheon has 330,000,000 deities, although even the most detailed texts have never tried to list all of them or even all the categories of them. There are many shrines to gods such as Śiva, Viṣṇu, Gaṇeśa, and their wives, as well as to the mother goddess Śakti (Kālī or Durgā). A shrine may be a pile of stones set up under the village tree or an elaborate iconographic creation in stone, marble, bronze, copper, silver, or gold. The latter are found at places of pilgrimage. The ancient *Kṣetra* (sacred grounds) celebrated in the scriptures have played a significant role in bringing together peoples from north and south and from different sociolinguistic groups.

Types of priesthood

Diverse types of priesthood are found among the tribal and Hindu communities of India. There used to be tribes such as the Korwa and other primitive communities that had no priests because they had no gods, their worship being confined to the spirits of their ancestors. Some tribes such as the Munda, the Oraon, and the Gond have their traditional priests, who perform ceremonial observances. There are also shamans who are said to have the power to cure diseases or other evils caused by malignant spirits. They give statements believed to be oracular.

In the Hindu religious system the highest caste group of Brahmins are traditionally and professionally the priests. Every orthodox Hindu of clean caste must have a Brahmin family priest to whom he pays customary dues in return for services at weddings, funerals, and other ceremonial occasions. The sacred specialists popularly known as *paṇḍā* are found at places of pilgrimage. Hindu religious organizations also sanction a number of ascetic orders, mainly at places of pilgrimage.

Islām as a religion is optimistic and transcendentalist; it does not favour either asceticism or extremes of ecstasy as does Hinduism. Three types of religious leaders are commonly found in the Muslim rural areas of South Asia: *mullahs*, *mawlānās*, and *pīrs*. A *mullah* is a prayer leader and religious functionary who performs various ceremonies for the villagers, usually in association with a particular mosque or sector of the village. His office is hereditary, and he exercises control over secular affairs through discourses on Fridays on nonreligious matters. The *mawlānā* is better educated than the *mullah* in theology and serves as an interpreter of the Qur'ān. The *mullah* ordinarily supports himself entirely through his religious occupation, but the *mawlānā* generally earns his living from a secular occupation. A *pīr* is a holy man who commands a following of both Hindus and Muslims in the South Asian countries.

Religious amalgams

The remnants of Buddhism in South Asia are found in the Himalayan states of Sikkim, Bhutan, and Nepal, as well as in the Indian territories of Himachal Pradesh, Kashmir, and North Bengal. The Sinhalese of Sri Lanka also practice Buddhism. The Himalayan Buddhists worship Hindu deities and observe certain Hindu ceremonies; sometimes they employ a Brahmin priest. Certain Himalayan tribes such as the Lepchā, the Bhutia, the Limbu, and the Lahaula, who practice Buddhism, combine it with animistic beliefs. The Chakmā of Bangladesh profess to be Buddhists, but their religious practices are mixed with the folk Hinduism of West Bengal. The same fusion of beliefs appears among the Magh of Chittagong, who have added the worship of Śiva and Durgā to the Buddhist observances.

Several other religions, including Jainism, Sikhism, Christianity, and Zoroastrianism, are also found in South Asia. Jainism is the chief religion of the district of Jodhpur and of certain trading and business castes. It may be regarded as a sect of Hinduism, and intercaste marriages are frequent. Sikhism, like Jainism, has a close affinity with Hinduism. Sikhs are disciples of their ten gurus, some of whose writings are compiled in the sacred book the *Adi Granth*. The Sikhs are concentrated in Punjab, though migrants are found all over India. One of the most remarkable figures in the religious history of India was Kabīr (d. 1518), a mystic poet who established a new religious sect linking Hinduism and Islām. Groups of his followers are scattered through northern India.

Religion has been both the inspiration and the matrix of South Asian art, which is discussed at length in the article SOUTH ASIAN PEOPLES, ARTS OF.

The cultures of the tribes of all India, the Himalayan states, and Sri Lanka are characterized by dances and songs. Most of these reflect the community spirit of the tribes and their patterns of life. The village *akharā* (dancing ground) and the youth dormitory, found in most of the tribes of India, are important in the development, maintenance, and transmission of the tribal dance pattern. The numerous traditional festivals are inseparably linked with singing, dancing, and drumming.

CULTURES IN TRANSITION

The traditional cultures of South Asia are rural, caste oriented, conservative, and religious in outlook. They have altered somewhat in the last 150 years under the impact of British rule and in response to modern education, technological development, and political change.

The tendency of the lower castes to rise upward accelerated under British rule. This process of upward mobility has been variously named Brahmanization, Sanskritization, and Kṣatriyaization. It may be observed both in tribal and in nontribal communities. Some of the tribes have been greatly influenced by the Kṣatriya *varṇa*, or class. This *varṇa* includes a number of caste groups, headed by the artistocratic Rājput lineages and differing widely in status. Some tribal and lower caste groups have been able to distinguish themselves by using the class name Kṣatriya or claiming Rājput descent. They marry their daughters to men of Rājput lineage. There are also examples of certain tribal groups claiming the status of Brahmins in the Hindu caste system. The Bauri of West Bengal claim to be Brahmins by virtue of their observance of the mourning period of 11 days, wearing of the sacred threads, and worship of Hindu deities.

Lower caste Hindus have made even more vigorous efforts to raise themselves in the caste hierarchy. Organized violation of caste rules by the so-called untouchable castes developed under the leadership of B.R. Ambedkar, who finally despaired of eliminating untouchability from Hinduism and led his followers into Buddhism. A similar turn to Buddhism took place quietly in Sri Lanka. There are instances in India of lower caste groups seeking to adopt the customs and rights of higher caste groups and being prevented from doing so. Such efforts have led to the partial breakdown of the *jajmānī* system in the villages and have also been a factor in the migration of the numerically weaker castes to the towns.

Some of the tribes in the hilly and forested areas of India were converted to Christianity by European missionaries, and made a conscious effort to abandon their tribal traditions. After India gained its independence, however, some of them foresook their Christianity in order to participate in certain political and economic privileges enjoyed by the traditional tribal groups. In northeastern India, where Christianity is strong among the tribes, they pressed demands for separate statehood within India, resulting in the establishment of Nāgāland, Meghalaya, and other tribal states.

The process of modernization and Westernization has impinged upon every sector of the South Asian cultures. The rate and direction of change differ from region to region. South Asia has been selective in what it has accepted from the West, mingling its acquisitions with many traditional institutions. Thus the new industrial towns are treated as modern places of pilgrimage; the secularization of society is pursued in terms of religious tolerance; and the parties in democratic elections contend in terms of old values. South Asia is clearly in a period of transition from traditional societies to modern ones, and the course it is following is unique.

BIBLIOGRAPHY

Ethnic stratification: D.N. MAJUMDAR, *Races and Cultures of India*, 4th rev. ed. (1961); GOVERNMENT OF INDIA, *Handbook on Scheduled Castes and Scheduled Tribes* (1968).

Himalayan tribes: TARAKCHANDRA DAS, *The Purum: An Old Kuki Tribe of Manipur* (1945); VERRIER ELWIN, *A Philosophy for NEFA*, 2nd rev. ed. (1959); CHRISTOPH VON FUERER-HAIMENDORF, *The Apa Tanis and Their Neighbours* (1962) and *The Sherpas of Nepal* (1964); GEOFFREY GORER, *Himalayan Village: An Account of the Lepchas of Sikkim* (1938); J.H. HUTTON, *The Angami Nagas* and *The Sema Nagas* (both 1921); D.N. MAJUMDAR, *The Himalayan Polyandry: Structure, Functioning, and Culture Change* (1962); J.P. MILLS, *Lhota Nagas* (1922) and *Ao Nagas* (1926); GOPAL S. NEPALI, *The Newars* (1965); W.H. NEWELL, *A Study of Gaddi* (1961); R.N. PEHRSON, *The Social Organization of the Marri Baluch* (1966); SACHIN ROY, *Aspects of Padam-Miniyong Culture* (1960); and S.K. SRIVASTAVA, *The Tharus: A Study of Culture Dynamics* (1958).

Tribes of middle India: E.T. DALTON, *Descriptive Ethnology of Bengal* (1872); VERRIER ELWIN, *The Muria and Their Ghotul* (1947) and *The Religion of an Indian Tribe* (1955); D.N. MAJUMDAR, *The Affairs of a Tribe: A Study in Tribal Dynamics* (1950); T.B. NAIK, *The Bhils* (1956); S.C. ROYA, *The Mundās and Their Country* (1912, reprinted 1970) and *The Orāons of Chōtā Nāgpur* (1915); and L.P. VIDYARTHI, *The Māler* (1963).

Tribes and castes of southern India: M.N. SRINIVAS, *Marriage and Family in Mysore* (1942) and *Religion and Society Among the Coorgs of South India* (1952); and EDGAR THURSTON, *Castes and Tribes of Southern India*, 7 vol. (1909).

Village life: MARY JANE BEECH *et al.*, *Inside the East Pakistan Village* (1966); S.C. DUBE, *Indian Village* (1955); PAUL G. HIEBERT, *Konduru: Structure and Integration in a South Indian Village* (1971); OSCAR LEWIS, *Village Life in Northern India: Studies in a Delhi Village* (1958); BRYCE RYAN *et al.*, *Sinhalese Village* (1958).

The caste system: ANDRE BETEILLE, *Caste, Class and Power* (1965); J.N. BHATTACHARYA, *Hindu Castes and Sects* (1896); G.S. GHURYE, *Caste and Race in India* (1932); J.H. HUTTON, *Caste in India* (1946); E.R. LEACH (ed.), *Aspects of Caste in South India, Ceylon and North-West Pakistan* (1960); and M.N. SRINIVAS, *Caste in Modern India and Other Essays* (1962).

Family and kinship: K.M. KAPADIA, *Hindu Kinship* (1947); and, for a comparative treatment, IRAWATI KARVE, *Kinship Organization in India*, 2nd rev. ed. (1965).

Cultural change: M.C. SEKHAR, *Social Change in India: First Decade of Planning* (1968); M.N. SRINIVAS, *Social Change in Modern India* (1966); and L.P. VIDYARTHI, *Culture Change in the Tribes of Modern India* (1968).

(L.P.V.)

South Asian Peoples, Arts of

South Asia, consisting of the huge subcontinent of India, includes Pakistan, Bangladesh, Sri Lanka (formerly Ceylon), as well as the nation of India itself. In spite of differences in physical appearance, complexion, stature, and other ethnological features, the people of the entire region of South Asia are unified by a common cultural and ethical outlook; a wealth of ancient textural literature in Sanskrit, Prākrit, and regional languages is a major unifying factor. Music and dance, ritual customs, modes of worship, and literary ideals are similar throughout the subcontinent, even though the region has been divided into kaleidoscopic political patterns through the centuries.

The common cultural outlook

The close interrelationship of the various peoples of South Asia may be traced in their epics, as in the *Rāmāyaṇa* and the *Mahābhārata*. Kinship between the gods and heroes of regions far distant from each other is evident, and the place-names themselves often evoke common sources. Moreover, there have been continual attempts to impose a political unity over the region. In the 3rd century BC, for example, the emperor Aśoka had almost all of this region under his sway; in the 11th century AD, Rājendra I Cōḷa conquered almost the whole of India and a good portion of Southeast Asia; and the great Mughal Akbar again achieved this in the 16th century. Though the expansion and attenuation of boundary lines, the bringing together or pulling apart politically of whole regions, have characterized all of South Asian history, the culture has remained essentially one.

Geographic influences

The geography of the region encouraged a common adoration of mountains and rivers. The great Himalayas, which form the northern boundary, are the loftiest of mountains and are conceived to be the embodiment of nobility, the abode of immaculate snow, and the symbol of a cultural ideal. Similarly, the great rivers such as the Brahmaputra and the Indus are regarded as the mothers of their respective regions, assuring prosperity through their perennial supply of water.

The association of lakes and springs with water sprites and sylvan fairies, called *nāga*s and *yakṣa*s, is common throughout the region. Karkoṭa, the name of an early dynasty, itself signifies *nāga* worship in Kashmir. Sculptures of *nāga*s and *yakṣa*s found in widespread sites suggest a common spirit of adoration, as do sculptures, paintings, temples, and religious texts that for centuries were preserved within an oral tradition without losing their immaculate intonation. The same classical dance is seen in sculpture in Gandhāra in Pakistan, in Bhārhut in the north, and in Amarāvatī in the south.

Interrelationships of the arts

The relation of the various arts to each other is very close in South Asia, where proficiency in several arts is necessary for specialization in any one. Thus, it is believed that without a good knowledge of dance there can be no proficiency in sculpture, for dance, like painting or sculpture, is a depiction of all the world. For its rhythmic movements and exposition of emotion, dance also requires musical accompaniments; hence, knowledge of musical rhythm is essential. For the stirring of emotion either in music or in dance, knowledge of literature and rhetoric is believed to be necessary; the flavour (*rasa*) to be expressed in music, dance, sculpture, or painting requires a literary background. Thus all the arts are closely linked together.

The arts were cultivated in South Asia not only as a noble pastime but also in a spirit of dedication, as an offering to the Almighty. Passages in literature refer to princes studying works of art for possible defects. One inscription that mentions the name of the *sūtra-dhāra* ("architect") of the 8th-century Mallikārjuna temple at Pattadakal epitomizes the accomplishments and ideals, in both theory and practice, of the artist.

Artists traditionally have enjoyed a high position in South Asian societies. Poets, musicians, and dancers held honoured seats in the royal court. An inscription mentions the appreciation bestowed by Rājendra Cōḷa on a talented dancer, and the architect of the temple at Tiruvorriyūr, who was also patronized by Rājendra, was eulogized for his encyclopaedic knowledge of architecture and art. Nonetheless, the folk arts were closely linked with the elite arts. Tribal group dances, for example, shared common elements with classical art, dance, and music. Among the artistic traditions of the Indian subcontinent, sculpture in the round (*citra*) is considered the highest artistic expression of form, and sculpture in relief (*ardhacitra*) is next in importance. Painting (*citrābhāsa*, literally "the semblance of sculpture") ranks third. Feeling for volume was so great that the effect of chiaroscuro (*i.e.*, use of light and shade to indicate modelling) was considered very important in painting; a passage from a drama of the 5th-century poet Kālidāsa describes how the eye tumbles over the heights and depths suggested in the modelling of a painting. A classical text on art, *Citrasūtra* enumerates noteworthy factors in paintings: the line sketch, firmly and gracefully drawn, is considered the highest element by the masters; shading and depiction of modelling are valued by others; the decorative element appeals to feminine taste; and the splendour of colour appeals to common taste. The use of a minimum of drawing to produce the maximum effect in suggesting form is considered most admirable.

Portraits play an important role in the visual arts of South Asia, and there are many literary references to the effective depiction of portraits both in painting and in sculpture. A 6th-century text, the *Viṣṇudharmottara*, classifies portraiture into natural, lyrical, sophisticated, and mixed, and men and women are classified into types by varieties of hair—long and fine, curling to right, wavy, straight and flowing, curled and abundant; similarly, eyes may be bow-shaped, of the hue of the blue lotus, fishlike, lotus-petal-like, or globular. Artistic stances are enumerated, and principles of foreshortening are ex-

plained. Paintings or sculptures were believed to take after their creators, even as a poem reflects the poet.

Assimilation of foreign influences

Although South Asia has continually been subjected to strong outside influences, it has always incorporated them into native forms, resulting not in imitation but in a new synthesis. This may be seen even in the art of the Gandhāra region of Pakistan, which in the 4th century BC was immersed in Greco-Roman tradition. In the sculpture of this period Indian themes and modes have softened the Western style. Foreign influence is evident after the invasion of the Kuṣāṇas from Scythia in the 1st century AD, but the native element predominated and overwhelmed the foreign influence. During the Mughal period, from the 16th century, when Muslims from Central Asia reigned in South Asia, the blend of Iranian and Indian elements produced a predominantly Indian school that spread throughout the region, making it a unified cultural area under imperial rule. The influence of Islāmic art was enhanced by the second Mughal emperor, Humāyūn, who imported painters from the court of the Shāh of Persia and began a tradition that blended Indian and Persian elements to produce an efflorescence of painting and architecture.

Social influences on the arts

Art in all these regions has reflected a system of government, a set of moral and ethical attitudes, and social patterns. The king's desire to serve the people and to take care of them almost as his offspring is evident as early as the 3rd century BC. The ideal of the king as the unrivalled bowman, as the unifier, as the tall and stately noble spirit, as the sacrificer for the welfare of the subjects, and as the hero of his people (who conceive of him as on a stately elephant) is comprehensively illustrated in a magnificent series of coins from the Gupta Empire of North India of the 4th–6th centuries. The concepts of righteous conquest and righteous warfare are illustrated in sculpture. The long series of sculptures illustrating the history of the South Indian Pallava dynasty of the 4th–9th centuries gives an excellent picture of the various activities of government—war and conquests, symbolic horse sacrifices, the king's council, diplomatic receptions, peace negotiations, the building of temples, appreciation of the fine arts (including dance and music), the coronation of kings, and so on, all clearly demonstrating what an orderly government meant to the people. Similarly, moral and ethical attitudes are illustrated in sculptures that lay stress on *dharma*—customs or laws governing duty. The doctrine of *ahiṃsā*, or non-injury to others, often is conceived in the symbol of a deer, and the ideal of a holy place has been conceived as a place where the deer roams freely. The joy in giving and of renunciation is clearly indicated in art. Sculptures illustrate simple and effective stories, as from the *Pañca-tantra,* one of the oldest books of fables in the world. The spirit of devotion, faith, and respect for moral and ethical standards pervading the social structure in South Asia through the centuries is continuously represented in South Asian painting and sculpture.

(C.S.)

This article is divided into the following sections:

I. Literature
- Sanskrit, Pāli, and Prākrit literatures: 1200 BC–AD 1200
  - Sanskrit: formative period (1200–400 BC)
  - Sanskrit: epic and didactic literature (400 BC–AD 1000)
  - Pāli and Prākrit literature (*c.* 200 BC–AD 200)
  - Classical Sanskrit *kāvya* (200–1200)
- Dravidian literature: 1st–19th century
  - Early Tamil literature (1st–10th century)
  - Period of the Tamil Cōḷa Empire (10th–13th century)
  - 14th–19th century
- Indo-Aryan literatures: 12th–18th century
  - Hindi
  - Bengali
  - Assamese
  - Oriya
  - Marathi
  - Gujarati
  - Punjabi
  - Kashmiri
- Islāmic literatures: 11th–19th century
  - Arabic
  - Turkish
  - Regional languages
  - Persian
  - Urdu
- Sinhalese literature: 10th century AD to 19th century
- Modern period: 19th and 20th centuries
  - Bengali
  - Assamese
  - Hindi
  - Gujarati
  - Marathi
  - Punjabi
  - Rajasthani
  - Tamil
  - Malayalam
  - Kannada
  - Urdu
  - English
  - Sinhalese

II. Music
- Folk, classical, and popular music
  - Rural areas
  - Classical music
  - Nonclassical music of the cities
- Antiquity
  - Vedic chant
  - The classical period
- Medieval period
  - Precursors of the medieval system
  - Further development of the *grāma-rāgas*
- The Islāmic period
  - Impact on musical genres and aesthetics
  - Theoretical developments
- The modern period
  - Rhythmic organization
  - Musical forms and instruments
  - Interaction with Western music

III. Dance and theatre
- Influence on Southeast Asia
- The performing arts in India
- Indian dance
  - Classical dance
  - Folk dance
  - Modern Indian dance
  - Dance-training centres
- Indian theatre
  - Classical theatre
  - Folk theatre
  - Modern theatre
- Sri Lanka (Ceylon)
  - Dance
  - Masked drama
- Dance and theatre in Kashmir
- Pakistan
  - Folk dance
  - Performing arts in the Punjab
  - Theatre in Pakistan
- Bangladesh

IV. Visual arts
- Visual arts of the Indian subcontinent
  - General characteristics
  - Architecture
  - Sculpture
  - Painting
  - Decorative arts
- Visual arts of Sri Lanka (Ceylon)
  - Architecture
  - Sculpture
  - Painting

I. Literature

The peoples of South Asia have had a continuous literature from the first appearance in the Punjab of a branch of the Indo-European-speaking peoples who also settled all of Europe and Iran. In India this branch of Indo-Aryans, as they are usually called, met earlier inhabitants with different languages and no doubt a different culture —possibly a culture akin to that of the Indus Valley civilization, which had a script, and perhaps a literature of its own, of which nothing is known. Certain to have been settled in India were peoples who spoke languages of Dravidian origin, as well as other languages, called Munda, now preserved only by aboriginal tribes, which show affinities with the languages of Southeast Asia.

The earliest literature is of a sacred character and dates from *c.* 1400 BC in the form of the Ṛgveda. This work stands at the beginning of the literature of the Veda, or canonical Hindu sacred writings, which as a whole is roughly contemporary with the settlement of the Indo-Aryan peoples in the Punjab and farther east, in the mesopotamia of the Ganges and Yamunā rivers. The language of the Ṛgveda, which is a compilation of hymns to the high gods of the Aryan religion, is complex and archaic. It was simplified and codified in the course of the centuries from 1000 to 600 BC, which saw the development of prose commentaries called the *Brāhmaṇas*, *Āraṇyakas*, and *Upaniṣads*. While there must have been a long tradition of grammarians, the final codification of the language is ascribed to Pāṇini (5th or 6th century BC), whose grammar has remained normative for the correct language ever since. This language is called Sanskrit (Tongue Perfected). Sanskrit has had a scarcely interrupted literature from *c.* 600 BC until today, but its greatest efflorescence was in the classical period, from the 1st to 7th centuries AD. Because it was identified with the Brahminical religion of the Vedas, reform movements such as Buddhism and Jainism disdained the use of Sanskrit and adopted literary languages—amalgams of different dialects of the parent language—of their own, Pāli in Buddhism and Ardhamāgadhī in Jainism. These languages, usually called Prākrits—that is, derivative as well as more "natural" languages—produced a vast and, again, mostly sacred literature. In a further development of these dialects, the early beginnings can be seen of modern Indo-Aryan languages of northern India: Bengali (also the language of Bangladesh), Hindi (the official language of the Republic of India since 1947), Rajasthani, Punjabi, Gujarati, Marathi, Kashmiri, Oriya, Assamese, and Sindhi, each of which produced a literature of its own. Their names are derived from the regions in which they are spoken, regions with uncertain boundaries, where the different dialects fused at the borders. They all retained a close family resemblance that made bilingualism easy and a fact of Indian literary life.

Far more marked was the difference between Indo-Aryan speech and the languages of the Dravidian family, which are structurally wholly different, though in time a measure of convergence took place. Among them, the oldest recorded is Tamil, now the language of Tamil Nadu (Madras) state and of northern Ceylon (now called Sri Lanka), whose literature goes back to the early centuries of the Christian Era. Later to be put to literary uses were the cognate Telugu (Andhra Pradesh), Kannada (or Kanarese, Mysore state), and Malayalam (Kerala state) languages.

Influence of Sanskrit

In spite of this linguistic differentiation, the literatures composed in all of these languages reflect, in different degrees, the monumental influence of Sanskrit literature, Sanskrit being the universal Indian language of culture. This influence was one of both substance and form: in substance it provided the basic themes of literary enterprise, notably through the epics, the *Mahābhārata* and *Rāmāyaṇa*, the Hindu popular texts of the *Purāṇas*, especially the *Bhāgavata*, and the mythological repertory that came with Sanskritic Hinduism; in form, Sanskrit belles lettres bequeathed models of literary composition, and Sanskrit poetics provided the aesthetic theory underlying the models. The impact of Islām created a new language, Urdu (from Persian: Camp), based on Hindi; Urdu was the lingua franca of the army. Urdu was used later for literature and at present is the mother tongue of most Indian Muslims and their brethren in Pakistan. Its influence, however, does not compare with that of Sanskrit.

Comparable to the impact of Sanskrit, but far more alien, is that of English, which began to assert itself in the 18th century. The language brought with it new literary forms that were gradually adapted to the old ones, producing new genres—without necessarily giving up the older ones—in the local languages and giving rise to an interesting literature in the English language. Once more, a universal cultural language to a large extent unified aims in the scattered languages; English still plays this role, though it appears to be slowly declining.

SANSKRIT, PALI, AND PRAKRIT LITERATURES: 1200 BC–AD 1200

Sanskrit: formative period (1200–400 BC). The oldest document in the literature of South Asia is the Ṛgveda, or Veda of the Stanzas (*c.* 1400 BC), the fundamental text of Brahminical Hinduism. Not literary but religious–magical in its purposes, it is mostly a compilation of hymns, dedicated to a number of gods of the Vedic religion. They have the regular structure of an invocation: the attention of the god is evoked; a brief account of some of his feats is given, to hold his attention; and an exhortation for his help concludes the hymn. The poets, of whom little is known, appear to have come at the close of a priestly poetical tradition, rivalling one another in allusions to obscure exploits, in language often opaque and at times intended to mystify. Nevertheless, the Ṛgvedic hymns include lines of great beauty. They may occur in a riddling verse, such as "When the ancient Dawns first dawned, the great Syllable was born in the footsteps of the Cow," alluding to the birth of speech at the beginning of creation. Or they may occur in poetry addressed to a deity whose beauty inspires the poet to well-turned lines. To the Dawns, for example: "They approach equally in the east, spreading themselves equally from the same place./ The Goddesses waking from the seat of order, like herds of kine set loose, the Dawns are active"; or to the goddess of the night: "Night coming on, the goddess shines/ In many places with her eyes:/ All-glorious she has decked herself./ Immortal goddess, far and wide/ She fills the valleys and the heights:/ Darkness with light she overcomes."

Nonsacred verses are very rare in the Ṛgveda, but, when they occur, they can be quite powerful, as in a hymn of a gambler, who is speaking:

> It pains the gambler when he sees a woman,
> Another's wife, and their well-ordered household:
> He yokes these brown steeds early in the morning,
> And, when the fire is low, sinks down an outcast.
>
> "Play not with dice, but cultivate thy cornfield;
> Rejoice in thy goods, deeming them abundant:
> There are thy cows, there is thy wife, O gambler."
> This counsel Savitri the kindly gives me.

Influence of the Ṛgveda on Sanskrit poetry

Although not literary in purpose, the Ṛgveda had a decisive influence on the form of Sanskrit poetry: except for narrative verse, the basic unit of all subsequent poems (no matter how many verses they consist of) is the single stanza that contains one complete thought.

The second Veda (1200 BC), the Yajurveda (Veda of the Yajus [Formulas]), contains sacred formulas recited by a group of priests at the great Vedic sacrifices; and the third (*c.* 1100 BC), the Sāmaveda (Veda of the Chants), is in essence an anthology of the Ṛgveda. More literary interest attaches to the fourth Veda (1200 BC), the Atharvaveda (an *atharvan* was a special priest), which contains hymns, incantations, and many magic charms.

The succeeding literature (*c.* 1000–700 BC), the *Brāhmaṇas* ("Disquisitions About the Ritual"), continues not the poetry but the liturgical concerns of the Ṛgveda. They were written in a dry, expository prose, so that only their narrative portions have any literary interest. Much the same is true of the next layer of Vedic texts (800–600 BC), the *Āraṇyakas* ("Books Studied in the Forest"). But the picture changes in the *Upaniṣads* (650–500 BC; "Collections of Esoteric Equations"). These prose texts at times convey the actual mode of teaching of a revered sage, in a style that can be strikingly intimate:

The *Upaniṣads*

> "Bring me a fruit of this banyan tree here." "Here it is, sir." "Split it." "It is split, sir." "What do you see inside?" "These rather tiny seeds, sir." "Come, split one of them." "It is split, sir." "What do you see in it?" "Nothing at all sir." He said to him, "That very tininess which you no longer perceive, by virtue of this very tininess does this banyan tree stand so big."

While the older *Upaniṣads* are in prose, the later ones, dating from around 400 BC, mark a shift back to verse. They are the oldest examples of didactic verse, a genre that later gained enormous popularity.

The contribution of late-Vedic texts to later literature is pre-eminently that of the development of an expository

prose style and the evolution of a sacred language, which, in order to be effective, must be completely correct. Thus, the Vedic religion evolved a science of phonetics and, later, of grammar, which was summed up in the 5th or 6th century BC by the grammarian Pāṇini in *Aṣṭādhyāyī* ("Eight Chapters"), a book that was to become basic to Sanskrit education. This language, Sanskrit, remained the language *par excellence* for later literature and was used for literary purposes until the 13th century and, epigonically, until today.

Sanskrit: epic and didactic literature (400 BC–AD 1000). After the formative period of the Vedic age, literature moved in several different directions. The close of the Vedic period was one of great cultural renewal, with the founding of the new monastic religions of Buddhism and Jainism (6th century BC) and the more slowly emerging rearticulation of Brahminism into Hinduism. Neither the earliest Buddhists nor the Jains availed themselves of Sanskrit in their preachings, apparently viewing the language as the preserve of a Brahmin elite. Sanskrit continued in derivative works of Vedic inspiration and above all in the *Mahābhārata* and the *Rāmāyaṇa*.

Mahābhārata. From references in Vedic literature it appears that side by side with the ritual texts there flourished a more secular literature carried on by bards. Originally charioteers to noblemen and thus witnesses of their feats, they chronicled the martial history of the families to which they were attached. From these beginnings, part chronicle, part panegyric, developed the epic style.

Like most Sanskrit poetry, the *Mahābhārata* consists of couplets, two successive lines with the same metre. Generally, one metre is used throughout the poem, though for stylistic effects other metres may be interspersed. The epic metre, or *śloka*, is a very fluid one that lends itself excellently to improvisation. The *Mahābhārata* is the longest poem in history, with about 100,000 couplets, more than seven times the size of Homer's *Iliad* and *Odyssey* combined. Its characters go back to around 1000 BC, but in its present form the epic could not have been composed before 400 BC. From that time until AD 400, it underwent continuous elaboration, by insertions of episodes (one of which is related in the religious poem called the *Bhagavadgītā*), accounts of separate adventures of the heroes, tales generated by their ancestors, and so on; and in the end it became a storehouse of general Hindu lore, with lengthy didactic books inserted.

The main narrative of the *Mahābhārata* recounts the growing up of two sets of cousins, both of whom aspire to a throne, the title to which is clouded. The protagonists, the Pāṇḍavas, stake their possessions in a dice game with the antagonists, the Kauravas, who are in effective control of the realm; they lose, and must live for 13 years in exile. This the five brothers do, along with the wife they hold in common. Upon their return from exile, they are refused their promised share of the kingdom, and, though parleys are held, war is inevitable. All of the Indian dynasties and tribes take sides in a war that lasts for 18 days, which only seven warriors, among them the Pāṇḍavas, survive. Noteworthy is the picture of gloom and doom that the *Mahābhārata* draws: there is little extolling of the heroic virtues of prowess and gallantry; rather, the wastefulness and bloodshed of war are pointed up, prefiguring a later concern with *āhiṃsā*, or nonviolence.

This summary does no justice to an extremely complex story with hundreds of participants, but it sketches the general outline of epic events. The main story has an unmistakable epic and heroic tone, and some of the events and encounters are completely comparable to those in epics of other peoples. But narrative and stylistic unity are disrupted by the inserted quasi-related and unrelated secondary episodes, each of which has a style of its own, ranging from light badinage to sonorous morality tales. It was in these episodes that the *Mahābhārata* lived on and greatly influenced succeeding literature; the story of Śakuntalā, for example, which the great 5th-century classical poet Kālidāsa embroidered, the slaying of Śiśupāla, the battle of the hero Arjuna with the mountain man, the story of Nala, and so on. But the most celebrated episode surely is the *Bhagavadgītā*.

The *Bhagavadgītā*

The influence of the *Bhagavadgītā* ("Song of the Lord") has mainly been on the development of Hindu religion and philosophy. Still, it is open to doubt whether it would have exerted this influence were it not for its poetry. Like most of the *Mahābhārata*, the style is simple and direct, not given to embellishment; nevertheless, the poem often reaches the height of expressiveness, as in its evocation of the theophany of Kṛṣṇa (Krishna) as Viṣṇu (Vishnu), in the 11th of its 18 chapters. It led to imitations such as the *Īśvaragītā*, ("Song of the Lord [Śiva]"), also in the *Mahābhārata*, in which the god Śiva (Shiva) is celebrated.

Rāmāyaṇa. While the unity of the *Mahābhārata* has been disrupted by interpolations, the unity of the second epic, the *Rāmāyaṇa*, has been remarkably preserved. It is less an epic than a romance, recounting the story of prince Rāma and his wife Sītā. The first book, a later addition, tells of the youth of the prince, who later, by the trickery of one of his father's wives, is excluded from the throne to which he is heir. He goes into voluntary exile in the forests with his wife and his brother Lakṣmaṇa. There a demon, Rāvaṇa, abducts Sītā to his island kingdom of Laṅkā. In the course of his quest for her, Rāma allies himself with a monkey nation, whose general, Hanumān, later revered as a god, discovers Sītā on Laṅkā. A monumental battle ensues. "As the sky can only be likened to the ocean and the ocean to the sky, so the battle of Rāma and Rāvaṇa can only be likened to the battle of Rāvaṇa and Rāma." After his victory, Rāma is restored to the throne, but (in what appears to be a later addition) the populace accuses Sītā of misbehaviour, probable adultery, while in Laṅkā. Rāma thus abandons her to a hermitage (the sage of the hermitage, Vālmīki, is credited with the authorship of the *Rāmāyaṇa*), where she gives birth to their twin sons. Ultimately, Rāma takes Sītā and his sons back. In the later additions, the first and probably the last books, King Rāma is accepted as an incarnation of the god Viṣṇu, rather than merely a perfect man and hero.

It is the main story of the romance that has made an indelible impression on Indian culture, morally as well as literarily. Rāma is the perfect, just king; Sītā, the model of an Indian wife; Lakṣmaṇa (the brother), the paragon of fraternal love; and the monkey Hanumān, the epitome of a servitor's loyalty. It was translated into and adapted in many modern Indian languages, and (like parts of the *Mahābhārata*) it found its way into Java. Vālmīki himself was hailed by later classical poets as the first true poet (*kavi*), and indeed much of his work has a poetic freshness and literary intention that is largely absent from the *Mahābhārata*. Vālmīki's great tools are metaphor and simile, as is also true of later literature. He delights in description of pastoral scenes, in lamentations and grand martial spectacles, and in the idyll of the hermitage, which depicts a serene sage leading a life of quiet meditation and living on simple forest fare in a tranquil woodland close to a sacred river. And the entire work is suffused with a confident, unwavering morality, for which the heroes of the *Mahābhārata* are still searching.

Harivaṃśa and Purāṇas. The role of the *Mahābhārata* as the storehouse of Hindu lore was supplemented by the *Harivaṃśa* ("Genealogy of Hari"—that is, the god Viṣṇu), which deals with the ancestry and exploits of Kṛṣṇa, the Pāṇḍavas' friend and adviser in the epic but now wholly deified and identified with the great god Viṣṇu. Then, from perhaps the 4th century, the literature of the *Purāṇas* took over. Encyclopaedic works, often of considerable length, the *Purāṇas* deal with the mythology of time and space and of deities, with sagas of great heroic dynasties, and with legends of saints and ascetics; their interest is largely religious. Aesthetically, the most important of them is the *Bhāgavata-Purāṇa* (9th or 10th century), which celebrates the blessed lord (*bhagavat*) Viṣṇu in his many theophanies but is particularly evocative in its celebration of Viṣṇu's incarnation as Kṛṣṇa and the playful story of his youth. The influence of the *Bhāgavata-Purāṇa*, particularly the 10th book, on Indian religion, art, and literature has been monumental. In the opinion of one scholar, this book constitutes the greatest poem ever written; and so it is in the popular estimation

of the Hindus. It was adapted in many Indian languages and provided themes and scenes for the flourishing miniature styles of the Middle Ages.

Pāli and Prākrit literature (c. 200 BC–AD 200). No more than the Vedic literature do the literatures of early Buddhism and Jainism have a literary intention. Their texts, written in dialects other than Sanskrit, articulate the teachings of the religious founders and their successors. Because they were transmitted orally for a considerable time before they were written down in the form they would retain, they underwent the inevitable censorship of the centuries, both negative in the form of documents dropped out of use and positive in the form of newer documents added. The dates given here are only approximations of the time of the documentary fixation of the dates.

Buddhist texts. The earliest records of Buddhism are not textual but inscriptional, in the famous edicts of the Mauryan emperor Aśoka, who reigned *c.* 269–232 BC. Among these inscriptions on stone, the so-called 13th rock edict—in which Aśoka, after the massacre of the Kaliṅgas (modern Orissa), abjures war—is the most moving document of any dynastic history. The inscriptions were written in a variety of Prākrits; that is, Indo-Aryan languages closely cognate to, but considerably later than, the earliest stabilized Sanskrit.

Pāli, the language of textual literature

The vehicle of the extant textual literature is the Pāli language, which is held to be a western Indian dialect on a substratum of several central and eastern ones. It was the language in use by the Theravāda school of Buddhism; but, since that school became the dominant one among many in early Buddhism, the Pāli language is often identified with the Buddha's own speech. Most of the canonical literature is exclusively of religious interest, but interspersed in it are works of considerable literary interest.

Foremost perhaps are discourses put into the Buddha's mouth—for example, his sermon "In the Deer Park"—and no doubt deriving from fairly accurate memories. With their straightforward, lively, and incisive style, homely similes, and simple humour, they are excellent examples of the homiletics of early Buddhist preaching. Incorporated in the canon, too, are more general works of literature. The *Dhammapada* ("Verses on the Buddhist Doctrine") is a fine example of the moralistic, aphoristic strain in Indian literature, in which virtue is extolled and vice condemned. It has remained a work of considerable diffusion in all Buddhist countries, and, as in the case of the *Bhagavadgītā* in Hinduism, much of its popularity is due to its literary style. The *Suttanipāta* collection of the Buddhist canon, composed in a more formal style, contains 55 narrative and didactic poems, in the form of dialogues and ballads; they are composed in a metre akin to the Sanskrit *śloka*. Of great interest are the *Theragāthā* and the *Therigāthā* ("Hymns of the Senior Monks" and "Hymns of the Senior Nuns"), which give at times a vivid insight into the ambience in which a conversion to Buddhism took place: a monk celebrates his newfound freedom in an idyll of the hermit's life; and a nun reminisces over the pains of deserting her home and child, yet without regrets, since she has won the freedom of Buddhism. The prosodic variety of Buddhist lyrics is great; about 30 different metres can be distinguished. Pāli poems, with their new metres (often based on a musical phrase), stylistic features, figures of speech, and choice diction, foreshadow classical *kāvya* literature in Sanskrit, whose extant specimens date from a later period.

The *Jātakas*

Of great importance is a huge volume called *Jātakas* ("Birth Stories"), recounting some 500 episodes supposedly having occurred in the Buddha's earlier lives. Only those parts in archaic verse are canonical; the prose portion was written later (*c.* 3rd century AD), probably in Ceylon. The *Jātakas* consist of fairy tales, animal stories and fables (the future Buddha may be incarnate in an animal), ballads, and anecdotes. Though their setting is often imaginary, they provide significant material for the historian of society and culture. These mostly short tales abound in moving, delicate, often rustic touches that have made them the delight of the Buddhist world. Their themes are illustrated in bas-reliefs of Buddhist shrines (or *stūpas*) at Bhārhut and Sānchi and monumentally on the great *stūpa* of Java, the Borobuḍur.

Of considerable literary as well as historical interest is the Pāli text *Milinda-pañha* ("The Questions of Milinda"). Milinda is identical to the Greek Menander, the name of a Bactrian Indo-Greek king (*c.* 140–110 BC) who was skeptical of the verities of Buddhism and was enlightened by the teaching of an elder, Nāgasena. The extensive Buddhist erudition that the sage displays is artfully presented in the form of simile and parable, and the work has contributed importantly to the edification of audiences in the countries where Buddhism came to be established. The style, in spite of the repetitions so typical of Buddhist doctrinal texts, is lively and presents the reader with an invaluable picture of contemporary Indian life.

Jaina texts. Less interest attaches to Jaina canonical works, which were written in an adapted and stabilized literary dialect called Ardhamāgadhī (Semi-Māgadhī, Māgadhī being the dialect of the ancient kingdom of Magadha, in present day Bihār). The belletristic contribution of Jaina literature is discussed below.

Classical Sanskrit kāvya (200–1200). Prepared for by the systematization of the Sanskrit language by Pāṇini, the development of the great epics, notably the *Rāmāyaṇa*, and the refinements of prosody represented by the Pāli lyrics, there arose, in the first centuries AD, a Sanskrit literary style that governed canons of taste for a millennium and remained influential far later through modern Indian languages and their literatures. The style, called *kāvya*, is characterized by an extremely self-conscious effort on the part of the writer to compose poetry pleasing to both the ear and the mind. It evolved an elaborate poetics of figures of speech, among which the metaphor and simile, in their many manifestations, predominate; a careful use of language, governed by the stated norms of grammar; an ever-increasing tendency to use compound nouns instead of drawing on the quite plentiful possibilities of Sanskrit inflection; a sometimes ostentatious display of erudition in the arts and sciences; an adroitness in the use of varied and complicated, if appropriate, metres—all applied to traditional themes such as the epic had provided and to the rendering of emotions, most often the love between men and women.

Characteristics of the *kāvya* style

The style finds its classical expression in the so-called *mahākāvya* ("great poem"), most akin to the epyllion ("miniature epic") art form of the Alexandrian poets (a school of Greek poets, *c.* 3rd–1st centuries BC); the strophic lyric (a lyric based on a rhythmic system of two or more lines repeated as a unit); and the Sanskrit theatre. It can also be extended to narrative literature, especially the prose novel. The great masters in the *Kāvya* form (which was also exported to Java) were Aśvaghoṣa, Kālidāsa, Bāṇa, Daṇḍin, Māgha, Bhavabhūti, and Bhāravi.

The earliest surviving *kāvya* literature was written by a Buddhist, Aśvaghoṣa, said to have been a contemporary of the Kuṣāṇa (Kushan) king Kaniṣka (1st century AD). Aśvaghoṣa's work also marks a shift away from the Pāli of the Theravāda branch of Buddhism back to the more and more accepted Sanskrit of the Mahāyāna branch. Two works are extant, both in the style of *mahākāvya:* the *Buddhacarita* ("Life of the Buddha") and the *Saundarānanda* ("Of Sundarī and Nanda"). Compared with later examples, they are fairly simple in style but reveal typical propensities of writers in this genre: a great predilection for descriptions of nature scenes, for grand spectacles, amorous episodes, and aphoristic observations. The resources of the Sanskrit language are fully exploited; stylistic embellishments (*alaṅkāra*) of simile and metaphor, alliteration, assonance, and the like are employed, often quite felicitously. The original *Buddhacarita*, rediscovered in 1892, had been known from Tibetan and Chinese translations. The Sanskrit text is fragmentary, breaking off in the 14th canto (major division of the poem) with the enlightenment of the Buddha, while the other versions take the story through the Buddha's Nirvāṇa. Though intended to instruct the reader to turn away from the sensuous life and follow the Buddha's

path, the work is at its best in descriptions of that very life. This is even more apparent in the *Saundarānanda,* which recounts a well-known story of how the Buddha converted his half-brother Nanda, who was deeply in love with his wife, Sundarī, and with the good life, to the monastic life of austerity. In his mastery of the intricacies of prosody and the subtleties of grammar and vocabulary, Aśvaghoṣa shows himself the complete forerunner of the Hindu *mahākāvya* authors.

The mahākāvya. In its classical form, a *mahākāvya* consists of a variable number of comparatively short cantos, each composed in a metre appropriate to its particular subject matter. The subject matter of the *mahākāvya* itself is taken from the epic, which is not, however, followed slavishly. Most *mahākāvya*s display such set pieces as descriptions of cities, oceans, mountains, the seasons, the rising of the sun and moon, games, festivals, weddings, embassies, councils, war, and triumph. It is typical of the genre that, while each strophe, or stanza, is intended to be part of a narrative sequence, it more often stands by itself, a discrete unit conveying one idea or developing one image. In this, the tendency of the Ṛgvedic stanza (see above *Sanskrit: formative period* [*1200–400 BC*]) continues in the classical literature. Although the lines of the classical stanza are long enough to convey their meaning quite explicitly, it is the pride of the poet to suggest rather than to express. Sometimes this is done by simple collocation of words: for example, in the first line of Kālidāsa's *Meghadūta* a *yakṣa* (a mischievous elf-like creature) is afflicted by a curse, "the more painful because it spelt separation from his beloved"; the next word notes that he had been negligent in his duties; taken together, the two words, though syntactically unrelated, suggest that it was his amour that made him neglect his duties. Another common suggestive device is the double meaning, or play on words. These double meanings often add a certain graceful playfulness to the poetry, reminding one that the poem was written first of all to give pleasure to the man of taste.

Traditionally there are six model *mahākāvya*s, three by Kālidāsa and one each by Bhāravi, Māgha, and Śriharṣa, to which sometimes the *Bhaṭṭikāvya* is added.

The works of Kālidāsa

Nothing is known with certainty of the life of Kālidāsa, the greatest of Sanskrit poets, but there is substantial agreement that at one time he lived in Ujjayinī (Ujjain, in the present state of Rājasthān), the capital of Avanti and an important centre of Sanskrit culture in a commercially busy area. His name, which means Servitor of Kālī, indicates that he was a follower of that goddess, whom he was to celebrate as Pārvatī, the daughter of the mountain, in the *Kumārasaṃbhava.* Probably he lived during the reign of Candra Gupta II Vikramāditya (*c.* 380–*c.* 415), and there are reports that he died, by the hand of an envious courtesan, while a guest of King Kumāradāsa of Ceylon.

Compared with those of others, Kālidāsa's style might be called simple, but it is a very studied, very felicitous simplicity, hiding the actual complexity of his constructions. In two of his *mahākāvya*s, Kālidāsa draws on epic lore. The first, and probably earlier one, is the *Kumārasaṃbhava* ("Birth of the War God"), which describes the courting of the ascetic Śiva, who is meditating in the mountains, by Pārvatī, the daughter of the Himalayas; the destruction of the god of love (after his arrow has struck Śiva) by the fire from Śiva's third eye; and the wedding and lovemaking of Śiva and Pārvatī, which results in the conception of the war god. The original is in eight cantos, but a sequel was added by an imitator. The second *mahākāvya,* the *Raghuvaṃśa* ("Dynasty of Raghu"), deals with themes from the *Rāmāyaṇa:* it describes the vicissitudes of the Solar dynasty of the ancient Indian barons, culminating in the *Rāmāyaṇa* story of Rāma and Sītā. The *Raghuvaṃśa* is famous for its beautiful descriptions and incidental narratives, which give the poem a somewhat episodic character; among them are a description of the six seasons (spring, summer, rainy, autumn, winter, and dewy) and the story of a young hermit who went to the river to fill a water jar for his parents and was killed by a stray arrow.

Unique in Sanskrit love poetry is Kālidāsa's *Meghadūta,* in which the poet tries to go beyond the strophic unity of the short lyric (see below), which normally characterizes love poems, by stringing the stanzas into a narrative. This innovation did not take hold, though the poem inspired imitations along precisely the same story line. The *Meghadūta* is the lament of an exiled *yakṣa* who is pining for his beloved on a lonely mountain peak. When, at the beginning of the monsoon, a cloud perches on the peak, he asks it to deliver a message to his love in the Himalayan city of Alakā. Most of the poem, composed in an extremely graceful metre, consists of a description of the landmarks, cities, and the like on the cloud's route to Alakā. It must be considered among the finest poems, if not the finest poem, written in Sanskrit. Kālidāsa also wrote for the theatre (see below) and was no doubt the most versatile author of Sanskrit literature; his works became well-nigh canonical models.

Bhāravi (6th century) probably hailed from the south during the reign of the Pallava dynasty. He took up a *Mahābhārata* theme in his *Kirātārjunīya* ("Arjuna and the Mountain Man"), recounting the Pāṇḍava prince Arjuna's encounter and ensuing combat with a wild mountaineer who in the end proves to be the god Śiva. Bhāravi's language and style are more difficult than Kālidāsa's, but the poem is highly regarded in Indian literary tradition.

Māgha, who wrote in the 8th century, was a conscious rival of Bhāravi, whom he attempted to surpass in every respect. His *Śiśupālavadha* ("The Slaying of King Śiśupāla") is based on an episode of the *Mahābhārata* in which the rival King Śiśupāla insults the hero-god Kṛṣṇa, who beheads him in the ensuing duel. Māgha is a master of technique in the strict Sanskrit sense of luscious descriptions; intricate syntax; compounds that, depending on how they are split, deliver quite different meanings; and the full register of stylistic embellishments.

To some critics, the preoccupation with technique, the triumph of form over substance, appears to have spelled the doom of the *mahākāvya.* A curious but entirely Sanskritic phenomenon, for example, is the *Bhaṭṭikāvya,* a poem by Bhaṭṭi (probably 6th or 7th century). It again deals with the story of Rāma and Sītā, but at the same time it illustrates in stanza after stanza, in exactly the proper sequence, the principal rules of Sanskrit grammar and poetics. Less artificial is the *Naiṣadhacarita* ("The Life of Nala, King of Niṣadha"), written by the 12th-century poet Śrīharṣa and based on the story of Nala and Damayantī in the *Mahābhārata.* An example of another kind of excess indulged in by *mahākāvya* writers is the *Rāmacarita* ("Deeds of Rāma"), by the 12th-century poet Sandhyākāra, which celebrates simultaneously the hero-god Rāma and the poet's own king, Rāmapāla of Bengal. Many other works were written in this style, and, even today, one may encounter a *mahākāvya* treatment of a great man such as Mahatma Gandhi or Jawaharlal Nehru.

Jayadeva's *Gītagovinda*

Difficult to classify is the work of the 12th-century Bengal poet Jayadeva, who wrote the *Gītagovinda* ("Cowherd Song"). The basic structure of this long poem, in which the poet recounts the youthful loves of the cowherd hero and god Kṛṣṇa, largely based on the story of the *Bhāgavata-Purāṇa,* is that of the *mahākāvya.* Generously interspersed between cantos, however, are erotic–religious lyrics of extremely musical assonances, which were, and still are, sung. Jayadeva's work, rather lacking in the grammatical rigidity of the other *mahākāvya* writers, has been extremely popular and affords a fine example of the devotional lyric (see below).

The short lyric. It is in the short, one-stanza lyric that Sanskrit poetry is revealed most intimately in its real aims. As noted, almost all of high Sanskrit poetry is strophic in fact; in the lyric it is so in intention. It is eminently a genre of the poetic moment, making an aesthetic observation and placing it within the Sanskritic universe of discourse. It may be an observation of anything: a fish glintingly jumping from a pond, aboriginal tribesmen engaged in a bloody rite, love in all its manifestations, a glimpse of God perceived or remembered. But in the monumental lyric collections that have been pre-

served, and in the many stray verses still circulating among educated Hindus in India as so-called *subhāṣita*s ("well-turned" couplets), the more common topics are praise of the god of one's devotion and the vagaries of love.

In the short lyric it is hard to make a distinction that depends on the language in which it is composed; for, although the language may be different, the subject matter and forms are the same. Many love lyrics, especially when they describe feelings experienced by women, are composed not in Sanskrit but, instead, in one of the Prākrits, or Middle Indo-Aryan languages, among which the dialect called Māhārāṣṭrī is particularly popular. The collection of 700 poems in this language, compiled by Hāla under the name of *Sattasaī* ("The Seven Hundred"), tends to be simpler in imagery and in the emotion portrayed than their Sanskrit counterparts, but essential differences are difficult to pinpoint.

The devotional lyric, a short verse expressing the author's devotion to a god, is linked with both the hymnal poetry of the Ṛgveda—though far less determined by a desire for compelling magic—and the temple worship of Hinduism. Though by no means always, there is often a particularism about them: the deity is invoked as it appears in a specific iconic stance or in a local temple or in a manifestation especially pleasing to the poet. The number of such verses is countless; every major religious and philosophic leader is held to have added to their stock. Some are especially famous: the *Sūryāṣṭaka* ("Eight Strophes for the Sun"), by Mayūra; the collections attributed to the philosopher Śaṅkara, the Saundaryalaharī; and the *Kṛṣṇakarṇāmṛta* ("The Elixir of Hearing of Kṛṣṇa"), by Bilvamaṅgala, among others. These *stotra* ("lyrics of praise") quite often were set to music, and people continue to sing them today—without necessarily comprehending the full intention of the Sanskrit, much like Roman Catholic believers singing Latin hymns.

The love lyric

The entire erotic experience, from budding love to the aftermath of consummation, is represented brilliantly in lyric poetry. But among the many themes inspired by love, poets have been most attracted to the lament of separated lovers. It is mostly the sufferings of the woman that are portrayed, but the grief of the man is also depicted—in Kālidāsa's *Meghadūta*, for example. The love lyrics consist of single verses, many of which seek to suggest the mood of *śṛṅgāra* (physical love.) While often extremely erotic, they are very rarely obscene. Sanskrit norm banned all coarse expressions for sexual play; and, although much probably escapes the modern reader, blunt allusions to genital organs are rare and, where allusions occur, extremely veiled. Bodily parts with less overt sexual connotations, such as breasts and buttocks, are frankly mentioned and described—in fact, celebrated. In allusions to sexual intercourse the terminology of the *Kāmasūtra* of Vātsyāyana is frequently invoked, as though this ancient textbook of Indian erudition was a protection against possible opprobrium—not unlike Latin terms resorted to in the West for actions that most know by shorter, more colloquial names.

The erotic and the devotional lyric merge freely, and at times it is impossible to make out whether the free sexual imagery employed is to be taken literally or as an allegory of the human soul courting the love of its god. The task—not a very pressing one—is made more difficult by the fact that some *bhakti* (devotion) religions have developed the poetics of love poetry into a kind of theology, a phenomenon quite characteristic of Bengal Kṛṣṇaism (see below *Indo-Aryan literatures: 12th–18th century*).

Authors of *subhāṣita*s often collected them themselves, the favourite form being that of the *śataka* ("century" of verses), in which 100 short lyrics on a common theme were strung together. Mention has been made of Hāla's *Sattasaī* ("The Seven Hundred," consisting of lyrics in the Māhārāṣṭrī dialect). Four well-known Sanskrit collections, of the 7th century, are the famous "century" of Amaru, king of Kashmir, and the three "centuries" by the poet Bhartṛhari; one of the latter collections is devoted to love, another to worldly wisdom—a very popular theme in epigrammatic verse—and the third to dispassion. Of the same type but in a different vein is *Caurapañcāśikā* ("Fifty Poems on Secret Love"), in which the 12th-century poet Bilhaṇa fondly recalls the pleasure of his clandestine amours with a local princess.

The theatre. Of all the literary arts, the Indians esteemed the play most highly, and it is in this form that most of the other arts were wedded together. Its origins are obscure, but there is reason to assume that the play developed out of recitations of well-known epic stories by professional reciters. It is an extremely rich genre with a number of outstanding playwrights.

The style is extremely varied. Although it might be called a Sanskrit play, Sanskrit is by no means the only language used, for the less educated characters, including all women, speak Prākrits of different degrees of niceness. The action is carried by prose, but at the least provocation—indeed, at any of the poetic moments characteristic of the strophic lyric—the author reverts to verse, sometimes in mid-sentence. Two principal types of play are distinguished: the *nāṭaka*, which is based on epic material, and the *prakaraṇa*, which is of the author's invention, though often borrowed from narrative literature.

Principal types of drama

Characteristic of the Sanskrit theatre are elements of sacrality. The play begins and ends with a benediction, many of which consist of subject matter taken from sacred texts. It is also expressed in numerous taboos: the play must have a happy outcome in which harmony, interrupted by the drama of the play, is restored; improper scenes, such as eating, dressing and undressing, and sexual intercourse, are not to be portrayed; no violence among the higher characters is permitted; war, which often occurs, should simply be reported on, often by lower characters, not in any way staged.

Fragments of Buddhist plays prior to the flowering of Hindu theatre have survived, but no complete plays earlier than 13 ascribed to the playwright Bhāsa. There is considerable controversy over the authenticity of the Bhāsa plays, but at least some of them must be authentic, perhaps dating back to the 3rd century. The plays are based on the epic and on the *Bṛhat-kathā* narrative cycle (see below); among the latter, the *Svapnavāsavadattā* ("The Dream of Vāsavadattā") is the most famous. Of considerable interest also is the *Daridra-Cārudatta* ("The Poverty of Cārudatta"), which became the basis for the play *Mṛcchakaṭika* ("Little Clay Cart") of Śūdraka (see below).

It must be assumed that there was an efflorescence of poetry and theatre in the city of Ujjayinī, one of the capitals of the Gupta Empire, in the 5th century, for a number of authors can be placed there during this reign; among these were Viśākhadatta, Śūdraka, Śyāmilaka, the writer of one of the best farces, and Kālidāsa, who at the beginning of the development of the genre produced some of the greatest plays in the tradition.

Three plays by Kālidāsa remain, one of which is the *Mālavikāgnimitra* ("Agnimitra and Mālavikā"), a harem play of amorous intrigue at a royal court. The other two are based on old themes. *Vikramorvaśī* ("Urvaśī Won by Valour") is based on a story as old as the Ṛgveda, that of the nymph Urvaśī, who is loved by King Purūravas, whom she marries on the condition that she shall never see him nude. The accident happens, and the nymph returns to heaven, leaving her husband crazed with longing, until a final reunion. But the Indian tradition holds the *Abhijñāna śakuntalā* ("Śakuntalā and the Token of Recognition") to be the greatest of all Sanskrit plays. It recounts a *Mahābhārata* story—rather freely to be sure—of a hermit girl secretly married to a visiting king, who leaves with her a keepsake that will serve her as a token of recognition. She gives birth to a son, Bharata, and goes to the King's court; on the way she loses the ring in a river, where a fish swallows it. The King fails to recognize her and rejects her, and her mother, a nymph, carries her to heaven. When the ring is recovered by a fisherman and the King's memory is restored, he searches for Śakuntalā but does not find her. In the end he meets a boy who proves to be his son and is restored to him.

Kālidāsa's plays

Kālidāsa's great forte is the portrayal of emotions—ordinary enough in themselves (budding love, love consum-

mated, rejection, despair, a father's love for his son)—but Kālidāsa applies to them a mastery of expression and image that makes the play a work of perennial beauty.

Next to nothing is known of Śūdraka except that he must have hailed from Ujjayinī. His is the most charming of all *prakaraṇa* plays (those that are not based on epic material): the *Mṛcchakaṭikā* ("Little Clay Cart"), the story of an impoverished merchant and a courtesan who love each other but are thwarted by a powerful rival who tries to kill the woman and place the blame on the hero, Cārudatta. The play offers a fascinating view of the different layers of urban society. Viśākhadatta, the author of a rare semi-historical play called *Mudrārākṣasa* ("Minister Rākṣasa and his Signet Ring"), apparently was a courtier at the Gupta court. His play is a dramatization of the Machiavellian political principles expounded in the book *Artha-śāstra*, by Kauṭilya, who appears as the hero of the play.

To the 7th-century king Harṣa of Kanauj are attributed three charming plays: *Ratnāvalī* and *Priyadarśikā*, both of which are of the harem type; and *Nāgānanda* ("The Joy of the Serpents"), inspired by Buddhism and illustrating the generosity of the snake deity Jīmūtavāhana.

The plays of Bhavabhūti

Ranked by Indian tradition close to Kālidāsa himself, Bhavabhūti (early 8th century) was the author of three plays, two of which are based on the *Rāmāyaṇa* story. The *Mahāvīracarita* ("The Exploits of the Great Hero") treats of Rāma's battle with Rāvaṇa and the *Uttararāmacarita* ("The Later Deeds of Rāma") treats of the life of Rāma after he has abandoned Sītā. Bhavabhūti lacks the elegance and grace of Kālidāsa but is more pensive—even brooding—than his predecessor. His style is also very forceful. His *prakaraṇa Mālatī-Mādhava* ("Mālatī and Mādhava") is a complex love intrigue intermingled with sorcery and Tantric practices, including a human sacrifice and much violence.

This list by no means concludes that of the playwrights in the Sanskrit tradition. The writing of plays, mostly derivative from the great models, has continued until the present day.

Apart from the more seriously intended plays described above, the Sanskrit theatre also has a rich repertory of farces, which are usually in one act. Most interesting of these are the *bhāṇa*s, which may be monologues in which an actor addresses imaginary persons and is answered by them, as he paints a picture of town life full of personal and social satire. Among the best in this little-studied genre is Śyāmilaka's 5th-century *Pādataḍitaka* ("The Courtesan's Kick").

Narrative literature. Sanskrit narrative literature is extremely rich, so rich in fact that at one time it was believed that all folktales originally came from India. Many indeed have, and they have found a place in *The Arabian Nights' Entertainment*, Boccaccio's *Decameron*, and other such works down to the fairy tales of Hans Christian Andersen and the fables of Jean de La Fontaine. Certain collections of animal tales, some of which go back to the Buddhist *Jātaka* stories, had incredible histories. The most famous is the *Pañca-tantra* ("The Five Chapters"), which, within a framework of a lesson in the art of politics addressed to young princes, presents a number of animal characters who in their actions both admonish and exhort the reader to a life certain to lead to worldly success. A shorter version, partly drawn from the *Pañca-tantra*, is the *Hitopadeśa* ("Good Advice"). The *Pañca-tantra* found its way to the West through translations into Persian, Arabic, Syrian, Hebrew, and Latin, until most of the medieval literatures possessed their own versions of it. No less extensive were its migrations to Southeast and East Asia.

The *Pañca-tantra*

The principal work of the novelistic and picaresque tale is the *Bṛhat-kathā* ("Great Story") of Guṇāḍhya, written in Prākrit and now lost, save for Sanskrit retellings. The most important among these Sanskrit versions is the *Kathā-saritsāgara* ("Ocean of Rivers of Stories") of Somadeva (11th century), which includes so many subsidiary tales that the main story line is frequently lost. Perhaps more faithful to the original—in any case far less distracting—is the *Bṛhatkathāślokasaṃgraha* ("Summary in Verse of the Great Story"), by Budhasvāmin (probably 7th century), one of the most charming of Sanskrit texts. Other collections of tales include the *Vetāla-pañcaviṃśatikā* ("Twenty-five Tales of a Ghost"), *Śūkasaptati* ("The Seventy Stories of a Parrot"), and the *Siṃhāsanadvātrim-sātikā* ("Thirty-two Stories of a Royal Throne").

Related to the *Bṛhat-kathā* cycle, though the exact relationship is unclear, is the Jain Prākrit text of the *Vāsudevahiṇḍī*, "The Roamings of Vāsudeva" (before 6th century), describing the acquisition of numerous wives by Kṛṣṇa Vāsudeva.

Though the tales are often artless, sometimes they are elaborately embroidered in the Sanskrit *kāvya* style. A fine example is the *Daśakumāracarita* ("Tales of Ten Princes"), by Daṇḍin (6th/7th century), in which, within the framework of a boxing story, the picaresque adventures of 10 disinherited princes are described in prose. While tending overly to description, the work remains eminently readable for the modern reader, a quality that cannot be attributed to the prose novels of the 7th-century writer Bāṇa: the *Harṣacarita*, "The Life of Harṣa" (king of Kanauj and the author of three plays, discussed above), which is important for its information on culture and society; and the *Kādambarī* (the name of the heroine), which describes the affairs of two sets of lovers through a series of incarnations, in which they are constantly harassed by a cruel fate. (J.A.B.v.B.)

DRAVIDIAN LITERATURE: 1ST–19TH CENTURY

Of the four literary Dravidian languages, Tamil has been recorded earliest, followed by Kannada, Telugu, and Malayalam. Tamil literature has a classical tradition of its own, while the literatures of the other languages have been influenced by Sanskrit models.

Early Tamil literature (1st–10th century). *Caṅkam literature.* Early classical Tamil literature is represented by eight anthologies of lyrics, 10 long poems, and a grammar called the *Tolkāppiyam* ("Old Composition"). According to a fanciful Tamil tradition, this literature was produced by poets of three "academies," or *caṅkams*, that in the hoary past were centred in the southern Indian city of Madurai and supposedly lasted 4,400, 3,700, and 1,850 years, respectively. The *Tolkāppiyam* was ascribed to the second *caṅkam*, the eight anthologies and 10 long poems to the third; according to tradition, nothing is extant from the first *caṅkam*. The early literature, itself known as *caṅkam*, comprises 2,381 poems, ranging from four to nearly 800 lines each and assigned to 473 poets who are known by name or epithet; about 100 poems are anonymous. Though the literature does not go back as far as native tradition would have it, it is generally ascribed to the first three centuries of the Christian Era and represents the oldest non-Sanskrit literature to be found on the South Asian subcontinent.

The eight anthologies and their contents, excluding opening invocations that were added later, are as follows: *akam* anthologies consisting of (1) *Kuṟuntokai*, 400 love poems, (2) *Naṟṟiṇai*, 400 love poems, (3) *Akanāṉūṟu*, 400 love poems, (4) *Aiṅkuṟunūṟu*, 500 love poems, each 100 (assigned to a different poet) dealing with one of five phases of love, (5) *Kalittokai*, 150 love poems in a metre called *kali;* and *puṟam* anthologies consisting of (6) *Puṟanāṉūṟu*, 400 poems, (7) *Patiṟṟuppattu* ("The Ten Tens"), 100 poems on kings (the first and last decades are missing), and (8) *Paripāṭal*, a collection of 70 religious poems. *Paripāṭal* and *Kalittokai* appear to be the latest of the anthologies; *Kuṟuntokai* and *Puṟanāṉūṟu* probably contain the earliest compositions. The remarkable work of grammar and rhetoric, *Tolkāppiyam*, is the crucial text for an understanding of early Tamil language and literature. Divided into three sections (each consisting of *cūttiram*s, or aphorisms)—sounds, words, and meaning—the *Tolkāppiyam* details, in the third, the canons of *caṅkam* poetic traditions.

Characteristics of *caṅkam* poetry

In the *Tolkāppiyam* and the anthologies, poems are classified by theme into *akam* ("interior") and *puṟam* ("exterior"), the former highly structured love poems, the latter heroic poems on war, death, personal virtues,

the ferocity and glory of kings, and the poverty of poets. Both the *akam* and the *puṟam* had well-defined *tiṇai*s (genres) that paralleled one another: *e.g.*, the *kuṟiñci* genre, in love poetry, which dealt with the lovers' clandestine union on a hillside by night; and the *veṭci* genre, in heroic poetry, which dealt with the first onset of war, by nocturnal cattle stealing. Both *kuṟiñci* and *veṭci* are names of flowers that grow on the hillside, here symbolic of the poetic genre, the mood, and the theme. By such pairings across *akam* and *puṟam*, love and war become part of the same universe and metaphors for one another; the same poets—for example, Paraṇar and Kapilar—wrote great poems in both genres. The basic technique depended on a taxonomy of Tamil nature and culture, of culturally defined time, space, nature, and human nature. For example, matched in metaphor with five phases of *akam* love (union; infidelity; anxious waiting; patient waiting; and the lover or lovers eloping or journeying for wealth, knowledge, and so on) are six seasons, six parts (dawn, forenoon, noon, afternoon, evening, and night) of the day, and five landscapes (hill, seashore, forest, pasture, and wasteland, named after characteristic flowers —*kuṟiñci*, *neytal*, *mullai*, and *marutam*—and the evergreen tree, *pālai*) and their contents (including gods, foods, birds, beasts, drums, occupations, lutes, musical styles, flowers, and kinds of running or standing water). Each landscape becomes a repertoire of images—anything in it, bird or drum, tribal name or dance, may evoke a specific feeling. A favourite poetic device is *uḷḷuṟai* (*i.e.*, metonymy, a figure of speech consisting of the description of one thing used to evoke that of another with which it is associated). Thus, the natural scene implicitly evokes the human scene; for example, bees making honey out of *kuṟiñci* flowers evokes the lovers' union. Not only is the poet's language Tamil, but the landscapes, the personae, and the appropriate moods and situations formulate the realities of the Tamil world into a code of symbols. For some five or six generations, the *caṅkam* poets spoke this common language of symbols, creating a body of lyrical poetry probably unequalled in passion, maturity, and delicacy by anything in any Indian literature.

Eighteen Ethical Works. The *Patiṉeṇ-kīḻkkaṇakku* ("Eighteen Ethical Works"), usually dated as post-*caṅkam* (4th–7th centuries), are all affected by Jainism and Buddhism. Of these the *Tirukkuṟaḷ* ("Sacred Couplets"), ascribed to Tiruvaḷḷuvar, is the most celebrated. Its 1,330 hemistichs (half lines of verse) are probably the final distillation of different periods. There are many parallels in the work to the Sanskrit *Kāma-sūtra*, the treatise on erotic love, to *Manu-smṛti*, an ancient treatise on special obligation and religious law, and to *Artha-śāstra*, Kauṭilya's treatise on politics. The *Kuṟaḷ* has three sections: *aṟam*, or virtue (Sanskrit *dharma*); *poruḷ*, government and society (Sanskrit *artha*); and *kāmam*, love (Sanskrit *kāma*). There is no special treatment of *mokṣa*, or salvation, though *aṟam* seems to include it. In the *aṟam* (virtue) section, the *Kuṟaḷ* sums up a world-affirming wisdom, the wisdom of human sympathy, expanding from wife, children, and friends to clan, village, and country. In the *poruḷ* (government and society) section, the aphorisms project a vision of an ideal state, based on educated human nature, and relate the good citizen to the good man. Prostitution, disease, drink, and gambling are listed, with foreign enemies, as dangers to the state. In the *kāmam* (love) section, the *Kuṟaḷ* follows the *caṅkam's* love—eros, or sexual love—yet anticipates agape, the perfecting of love through many lives, which appears in religious poetry of the next age.

Epics. The age of the Pallavas (300?–900), a warrior dynasty of Hindu kings, is known for its epics, beginning with *Cilappatikāram* ("The Jewelled Anklet") and *Maṇimēkalai* ("The Girdle of Gems") and including an incomplete narrative, *Peruṅkatai* ("The Great Story"), the *Cīvakacintāmaṇi* ("The Amulet of Cīvakaṉ") by Tiruttakkatēvar, and *Cūḷāmaṇi* ("The Crest Jewel") by Tōlāmoḻittēvar. The last three works depict Jaina kings and their ideals of the good life, nonviolence, and the attainment of salvation through self-sacrifice. They are also characterized by excellent descriptions of city and country and by a mixture of supernatural and natural elements. In their episodic methods of narration and set descriptions of erotic, heroic, and religious themes, these Jaina epics became both models and sources for later epic works.

The *Cilappatikāram*

The Cilappatikāram, by Iḷaṅkō Aṭikaḷ, is in three books, set in the capitals of the three Tamil kingdoms: Pukār (the Cōḷa capital), Maturai (*i.e.*, Madurai, the Pāṇṭiya [Pāṇḍya] capital), and Vañci (the Cēra capital). The story is not about kings but about Kōvalaṉ, a young Pukār merchant, telling of his marriage to the virtuous Kaṇṇaki, his love for the courtesan Mātavi, and his consequent ruin and exile in Maturai, where he dies, unjustly executed when he tries to sell his wife's anklet to a wicked goldsmith who had stolen the Queen's similar anklet and charged Kōvalaṉ with the theft. Kaṇṇaki, the widow, comes running to the city and shows the King her other anklet, breaks it to prove it is not the Queen's—Kaṇṇaki's contains rubies, and the Queen's contains pearls—and thus proves Kōvalaṉ's innocence. Kaṇṇaki tears off one breast and throws it at the kingdom of Maturai, which goes up in flames. Such is the power of a faithful wife. The third book deals with the Cēra king's victorious expedition to the north to bring Himalayan stone for an image of Kaṇṇaki, now become a goddess of chastity (*pattiṉi*).

The *Cilappatikāram* is a fine synthesis of mood poetry in the ancient Tamil *caṅkam* tradition and the rhetoric of Sanskrit poetry—even the title is a blend of Tamil and Sanskrit—including in the epic frame *akam* lyrics, the dialogues of *Kalittokai* (poems of unrequited or mismatched love), chorus folk song, descriptions of city and village, lovingly technical accounts of dance and music, and strikingly dramatic scenes of love and tragic death. One of the great achievements of Tamil genius, the *Cilappatikāram* is a detailed poetic witness to Tamil culture, its varied religions, town plans and city types, the commingling of Greek, Arab, and Tamil peoples, and the arts of dance and music.

Maṇimēkalai (the heroine's name "Girdle of Gems"), the second, "twin," epic (the last part of which is missing), by Cātaṉār, continues the story of the *Cilappatikāram;* the heroine is Mātavi's daughter, Maṇimēkalai, a dancer and courtesan like her mother. Maṇimēkalai is torn between her passion for a princely lover and her spiritual yearnings, the first encouraged by her grandmother, the second by her mother. She flees the attentions of the prince, and, while he pursues her, she attains magical powers: she changes forms; survives prison, lecherous villains, and other dangers; converts the Queen; and finally goes to Pukār, which is being destroyed by oceanic erosion, worships Kaṇṇaki, and arrives in Vañcī to work in famine relief and to perform "penance." Unlike the *Cilappatikāram*, the *Maṇimēkalai* is partisan to Buddhism. It is known for its poetry and its lively discussions of religion and philosophy.

Bhakti poetry. From the 7th century onward, a movement with religious origins made itself heard in literature. The movement was that of *bhakti*, or intense personal devotion to the two principal gods of Hinduism, Śiva and Viṣṇu. The earliest *bhakti* poets were the followers of Śiva, the Nāyaṉārs (Śiva Devotees), whose first representative was the poetess Kāraikkāl Ammaiyār, who called herself a *pēy*, or ghostly minion of Śiva, and sang ecstatically of his dances. Tirumūlar was a mystic and reformer in the so-called Siddhānta (Perfected Man) school of Śaivism, which rejected caste and asceticism, and believed that the body is the true temple of Śiva. There were 12 early Nāyaṉār saints. Similar poets, in the tradition of devotion to the god Viṣṇu, also belonged to this early period. Called Āḻvārs (Immersed Ones), they had as their first representatives Poykai, Pūtaṉ, and Pēyār, who composed "centuries" (groups of 100) of linked verses (*antāti*), in which the final line of a verse is the beginning line of the next and the final line of the last verse is the beginning of the first, so that a "garland" is formed. To these Āḻvārs, God is the light of lights, lit in the heart.

The Nāyaṉārs

The most important Nāyaṉārs were Appar and Cam-

pantar, in the 7th century, and Cuntarar, in the 8th. Appar, a self-mortifying Jain ascetic before he became a Śaiva saint, sings of his conversion to a religion of love, surprised by the Lord stealing into his heart. After him, the term *tēvāram* ("private worship") came to mean "hymn." Campantar, too, wrote these personal, "bone-melting" songs for the common man. Cuntarar, however, who sees a vision of 63 Tamil saints—rich, poor, male, female, of every caste and trade, unified even with bird and beast in the love of God—epitomizes *bhakti*. To him and other Bhaktas, every act is worship, every word God's name. Unlike the ascetics, they return man to the world of men, bringing hope, joy, and beauty into religion and making worship an act of music. Their songs have become part of temple ritual. Further, in *bhakti*, erotic love (as seen in *akam*) in all its phases became a metaphor for man's love for God, the lover.

In the 9th century, Māṇikkavācakar, in his great, moving collection of hymns in *Tiruvācakam*, sees Śiva as lover, lord, master, and guru; the poet sings richly and intimately of all sensory joys merging in God. Minister and scholar, he had a child's love for God.

Āṇṭāḷ (8th century), a Vaiṣṇava poetess, is literally love-sick for Kṛṣṇa. Periyāḻvār, her father, sings of Kṛṣṇa in the aspect of a divine child, originating a new genre of celebrant poetry. Kulacēkarar, a Cēra prince, sings of both Rāma and Kṛṣṇa, identifying himself with several roles in the holy legends: a *gopī* in love with Kṛṣṇa or his mother, Devakī, who misses nursing him, or the exiled Rāma's father, Daśaratha. Tiruppāṇāḻvār, an untouchable poet (*pāṇaṉ*), sang 10 songs about the god in Śrīraṅgam, his eyes, mouth, chest, navel, his clothes, and feet. To these Bhaktas, God is not only love but beauty. His creation is his jewel; in separation he longs for union, as man longs for him. Tirumaṅkaiyāḻvār, religious philosopher, probably guru (personal religious teacher and spiritual guide in Hinduism) to the Pallava kings, and poet of more than 1,000 verses, was apparently responsible for the building of many Vaiṣṇava temples. The last of the Āḻvārs, Nammāḻvār (Our Āḻvār), writing in the 9th century, expresses poignantly both the pain and ecstasy of being in love with God, revivifying mythology into revelation.

Period of the Tamil Cōḻa Empire (10th–13th century). The next period, the time of the Tamil Cōḻa Empire (10th–13th centuries), saw an awakening of neighbouring literatures: Kannada, Telugu, and Malayalam. The first extant Kannada work is the 9th-century *Kavirājamārga* ("The Royal Road of Poets"), a work of rhetoric rather indebted to Sanskrit rhetoricians, containing the first descriptions of the Kannada country, people, and dialects, with references to earlier works. From the 10th century on, *campū* narratives (part prose, part verse) became popular both in Kannada and in Telugu, as did renderings of the Sanskrit epics *Rāmāyaṇa* and *Mahābhārata* and Jaina legends and biography.

In Kannada, this period was dominated by the "three gems" of Jaina literature, Pampa, Ponna, and Ranna, as well as by Nāgavarma I, a 10th-century Kannada grammarian. Pampa was the *ādikavi* ("first of poets"), having attained that stature with two great epics: *Vikramārjuna Vijaya* and *Ādipurāṇa*. The former is a rendering of the *Mahābhārata*, with the hero, Arjuna, identified with the poet's royal patron, Arikēsarī. This felicitous epic is known for its succinct, powerful characterizations, its rich descriptions of Kannada country and court, its moving sentiments, and its harmonious blend of Sanskrit and Kannada. While the *Vikramārjuna* is a secular work, Pampa's *Ādipurāṇa* tells the story of the Jaina hero-saint Purudēva, his previous lives, his life from birth to marriage to holy death, as well as the lives of his sons, Bharata and Bāhubali.

Telugu had its *ādikavi* ("first of poets"), in the Brahmin Nannaya Bhaṭṭa (1100–60), who, in *campū* style, wrote three books of a version of the *Mahābhārata*, later finished by Tikkana (13th century) and by Errāpraggaḍa. Like other regional versions of the *Mahābhārata*, the Telugu version is not a literal translation but an interpretation, with many local elements and differences of emphasis; for example, Nannaya emphasizes the importance of Vedic religion. Such works have made the Sanskrit epics and *Purāṇas* part of a live and growing tradition, both oral and literary, in the regional language.

The work of Kampaṉ

This period also saw the eminence of Kampaṉ's Tamil version of the *Rāmāyaṇa* (12th century). In him there is a climactic blend of earlier *caṅkam* poetry, Tamil epics, the Āḻvārs' fervour of personal *bhakti* (devotion) toward Rāma, folk motifs, and Sanskrit stories, metres, and poetic devices. Instead of a just king and a perfect man, Rāma is an incarnation of Viṣṇu and an intense object of devotion, dwarfing the Vedic gods; Kampaṉ called his work *Irāmāvatāram* ("Rāma's Incarnation"); yet the emphasis is not on Viṣṇu but on *dharma* ("the law"), localized and Tamilized. More like Sanskrit than *caṅkam* poets, Kampaṉ revels in elaborate metaphor, hyperbole, and fanciful descriptions of virtue and nature. The work is long, consisting of about 40,000 lines; the *Yuttakāṇṭam* ("War Canto") alone, with 14 battles, equals the *Iliad* in length. The poem is also justly known for its variety of style, its exploitation of the resources of Tamil and Sanskrit both in form and content, its humour, and its handling of the narrative, dramatic, and lyric modes.

Kampaṉ's popularity extended not only into all of Tamil country but apparently into the north, influencing some episodes of Tulsī's Hindi version of the *Rāmāyaṇa*, and into northern Kerala, where 32 plays based on Kampaṉ are enacted ritually with marionettes in Śiva temples.

Pre-15th-century Tamil influence on early Malayalam, the language of Kerala, was strong and led to the literature of *pāṭṭu* ("song"), in which only Dravidian, or Tamil, phonemes may occur and Tamil-like second-syllable rhymes are kept. The best known *pāṭṭu* is *Rāmacaritam* (*c.* 12th–13th century; "Deeds of Rāma"), probably the earliest Malayalam work written in a mixture of Tamil and Malayalam. Other *pāṭṭu*s in Tamilized Malayalam, written by a family of poets (14th–15th centuries) from Niraṇam in central Travancore, appear in *Kaṇṇassan Pāṭṭukaḷ*, in which Tamil conventions of metre and phonology are loosened and more Sanskrit is allowed. Similar in style is a version of the *Rāmāyaṇa* by Rāma Paṇikkar, an abridged *Bhagavadgītā* by his uncle Mādhava Paṇikkar, and a condensed *Mahābhārata* and the 10th book of the *Bhāgavata-Purāṇa* by another uncle, Śaṅkara Paṇikkar.

As strong as Tamil influence was, the predominant influence on Malayalam was Sanskrit, in language as well as literary form. The influence on language led early to a mixture of Sanskrit and Malayalam in a literary dialect called *maṇipravāḷa* (meaning "necklace of diamonds and coral"). The author of the *Līlātilakam*, a 14th-century treatise on grammar and poetics, describes both the Tamilizing and Sanskritizing trends and genres and insists on harmonious blendings. Many kinds of poems were composed in *maṇipravāḷa* styles: *kūḍyāṭṭam*s (dramatic presentations using Sanskrit *śloka*s, or epic metres, for hero and heroine, *maṇipravāḷa* for the clown, and Malayalam for explanations intended for the laity); didactic works such as the 11th-century *Vaiśikatantram* ("Advice to a Courtesan by Her Mother"); 13th- and 14th-century *campūs* (narratives combining prose and verse) on dancers, such as *Unniyati Caritam* by Dāmōdara Cākkiyār; and several short poems in praise of women and kings. *Maṇipravāḷa* poems like these are essentially artificial expressions of courtly high-caste poets, preoccupied with eroticism and harlots. The *Candrōtsavam* (*c.* 1500; "Moon Festival") is a satire on the voluptuary *maṇipravāḷa* tradition, jostling together all the famed courtesans of the period.

Coexisting with the Tamilized and Sanskritized Malayalam poems produced by scholars was a live *pacca* ("pure, fresh") Malayalam tradition represented mostly by folk songs and ballads—for example, *Vaḍukkan Pāṭṭukaḷ* (hero ballads of the northern Malabar Coast); songs sung during weddings, deaths, or festivals; and work songs. All three styles—the indigenous folk style, the Tamil, and the Sanskrit—began to converge and influence each other by the 15th century in works such as *Kṛṣṇa Pāṭṭu* ("Song of Kṛṣṇa") and *Gāthā* ("Song").

Such grafting reached its full flowering in the 16th-century poet Eḻuttaccan (Father [or Leader] of Letters), who popularized the *kiḷippāṭṭu* ("parrot song"), a genre in which the narrator is a parrot, a bee, a swan, and so on. His outstanding works are *Adhyātma Rāmāyaṇam*, *Bhāratam*, and *Bhāgavatam*, all based on Sanskrit originals yet powerfully re-created with masterly language craft.

While Vaiṣṇava works were proliferating in Malayalam, Śaiva movements swept the other three languages, Tamil, Kannada, and Telugu. In Tamil, the hymns of the Nāyaṉārs were arranged and anthologized for scriptural and recitative use by the 11th century. Another such consolidation of sacred materials was Cēkkiḻār's 12th-century *Toṇṭar Purāṇam*, or *Periyapurāṇam*, narrating in epic style the lives of the 63 great Śaiva saints and creating a tradition for all Śaivas, even in the Kannada and Telugu areas. The theology of the Siddhānta (Perfected Man) school of Śaivism was elaborated in Meykaṇṭār's *Civañāṉa-pōtam* (13th century).

The new Kannada genre of the 13th century

By the 12th century, a new Kannada genre, the *vacana* ("saying" or "prose poem"), had come into being with the Vīraśaiva saints. In the language of the people, the saints expressed their radical views on religion and society, rejected both Brahminical ritualism and Jaina ascetic world negation, called all men to the *anubhāva* ("experience") of God, and broke the bonds of caste, creed, and sexual difference. Five important poet-saints were Dāsimayya; Basava, a self-searching social reformer and a minister of the Jaina king Bijjaḷa; Allama Prabhu, the elder and metaphysical master of them all; Mahādēviyakka, a woman saint singing love poems to Śiva; and Cennabasava, a brilliant theologian of the movement, who elaborated the theory of "six stages" of mystic ascent for the devotee. The many-faceted lyrics written by the poet-saints were in the spoken dialects of Middle Kannada, yet they drew on archetypal human images as well as ancient pan-Indian symbology for their intense and searing expressions of *bhakti*. Inspired by these lyrics, Harihara, in the late 12th century, wrote some 120 *ragaḷe* (blank verse) biographies of the Śaiva saints, including the Vīraśaiva (or Liṅgāyat) and the earlier Tamil Nāyaṉārs. In the early 13th century, his disciple and nephew, Rāghavāṅka, wrote, in *ṣaṭpadi*s (six-line stanzas), of the lives of saints, in well-structured works such as *Sōmanātha Carite* and *Siddharāma Caritra;* his most mature work is *Hariścandrakāvya*, an unequalled reworking of an ancient Job-like story of Hariścandra, who suffered every ordeal for his love of truth. The Vīraśaiva saints' lives and the *vacana* ("saying" or "prose poem") literature were codified in a masterpiece called *Śūnya Sampādane* ("The Achievement of Nothing"), consisting of dialogues interweaving the saints' *vacana*s, with the poet Allama Prabhu as the central figure.

Contemporary with the 13th-century Vīraśaiva saints were Telugu Śaiva poets such as Pālkuriki Sōmanātha, who composed the *Basavapurāṇam* employing popular metres and idiomatic Telugu. His *Paṇḍitārādhya Caritra* is a life of the Śaiva devotee Paṇḍitārādhya as well as a book of general knowledge including social customs, arts, crafts, and particularly music. His *Vṛṣādhipa Śatakam* consists of verses in Tamil, Kannada, Marathi, Sanskrit, and Telugu. This work was probably the first of the genre of *śataka*s ("centuries" of verses) literature, particularly popular in Telugu but also written in the other three languages as well as in Sanskrit (see above *Sanskrit: formative period [1200–400 BC]*).

Also of the 13th century is Āṇḍayya's *Kabbigara Kāva* ("The Poet's Defender"), in Kannada, a linguistic tour de force, eschewing unmodified Sanskrit forms; and Mallikārjuna's *Sūktisudhārṇava*, an excellent Kannada anthology of lyrics and passages. From 1240 to 1326, poets of Telugu produced over 100 verse renderings of the Sanskrit epic *Rāmāyaṇa* and many more in prose, the earliest being *Raṅganātha Rāmāyaṇa*, assigned to Gōna Buddhā Reḍḍi.

14th–19th century. The next age, from the 14th to the 16th century, is the great age of the Vijayanagar Empire. In this period, Kannada and Telugu were under the aegis of one dynasty and were also hospitable to the influence of neighbouring Muslim Bahmanī kingdoms. Śrīnātha was a 15th-century poet honoured in many courts for his scholarship, poetry, and polemics. He rendered Sanskrit poems and wrote *Haravilāsam* (Four Śaiva Tales); *Krīḍābhirāmam*, a charming, often vulgar account of social life in Warangal; and *Palanāṭi Vīra Caritra*, a popular ballad on a fratricidal war. Many erotic *cāṭu*s, or stray epigrams, are also attributed to him. Bammera Pōtana, a great Śaiva devotee in life and poetry, unschooled yet a scholar, is widely known for his *Bhāgavatam*, a masterpiece that is said to excel the original Sanskrit *Bhāgavata-Purāṇa*. Tāḷḷapāka Annāmācārya, son of a great family of scholars, fathered an exciting new genre of devotional song, all addressed to the god Śrī Veṅkaṭeśvara of Tirupati (a form of Viṣṇu). His *Saṅkīrtana Lakṣaṇam* is a collection of 32,000 songs in Sanskrit and Telugu, which made a significant contribution to Karnatic (southern Indian) musical technique.

The 16th century was an age of patronage by Vijayanagar kings, beginning with Kṛṣṇa Dēva Rāya, himself a poet versed in Sanskrit, Kannada, and Telugu. The *rāyala yugam* ("age of kings") was known for its courtly *prabandha*s, virtuoso poetic narratives by and for pandits (learned men). Among the most famous court poets were Piṅgaḷi Sūranna, whose verse novel, *Kalāpurṇōdayam* (1550)—a story full of surprises, magic, and changes of identity—is justly celebrated for its artistry; and Tenāli Rāmakṛṣṇa, known for his clownish pranks and humour, whose writings are the centre of a very popular cycle of tales in all four Dravidian languages.

During the 16th century and for the next few centuries, Telugu poets also flourished outside the Telugu country, especially in Tanjore (Thanjavūr) and Madurai, in Tamil country, and Pudukkoṭṭa and Mysore, in Kannada country. Their most important contribution was to native Kannada and Telugu dance drama on mythological themes, called *yakṣagāna*. The form is comparable to *kathākali* in the Malayalam area and to *terukkūttu* ("street drama") and *kuṟavañci* ("gypsy drama") in the Tamil area. The earliest Telugu *yakṣagāna* text is *Sugrīva Vijayam* (*c.* 1570), by Kandukuru Rudra Kavi; the earliest in Kannada is probably Śāntavīra Dēśika's *Saundarēśvara* (1678). The most celebrated of Kannada *yakṣagāna* dramatists is the versatile Pārti Subba, who flourished around 1800 and is known for his moving *Rāmāyaṇa* episodes and songs.

Contribution of poets to native dance drama

The 15th and 16th centuries produced some of the most popular classics in Kannada. Of these the greatest is Gadugu's *Kumāra Vyāsa*, or *Nāraṇappa*'s, 10 cantos of the *Mahābhārata;* recited in assemblies as well as in households, these are a continual delight, abounding in humour, passion, and memorable poetry. In *Prabhuliṅgalīle*, Cāmarasa made poetry out of the life of the poet-saint Allama. The *Jaimini Bhārata* and the many versions of *Rāmāyaṇa* episodes (especially Sītā's abandonment in the forest) written by the distinguished Śaiva epic poet Lakṣmīśa are known for their melodious verses and moving scenes. Ratnākaravarṇi's *Bharateśa Vaibhava* is a great Jaina story, tersely told in a Kannada song metre and celebrated for its depiction of many *rasa*s ("moods"), especially the erotic.

Kannada Vaiṣṇava *dāsa*s ("servants [of God]") wrote in a song genre called *pada*, parallel and often indebted to the Vīraśaiva *vacana*s ("sayings" or "prose poems"). Purandaradāsa, a rich 16th-century merchant turned mendicant, saint, and poet, composed *bhakti* (devotional) songs on Viṭṭhala (a manifestation of the god Viṣṇu), criticizing divisions of caste and class and calling on the mercy of God. His *pada*s and *kīrtanas* ("lauds") are also landmarks in Karnatic music. Kanakadāsa, his contemporary and a shepherd by birth, wrote *pada*s and longer popular works. *Dāsa* songs are part of the repertory of all South Indian musicians.

The folk *tripadi* ("three-line verse") of Sarvajña (1700?) is a household word for wit and wisdom, like the *Kuṟaḷ* in Tamil (see above *Eighteen Ethical Works*) and the "century" of four-line verses in Telugu by Vēmana (15th century). The moral, social, satiric, and wise proverb-like aphorisms of Vēmana and Sar-

vajña are widely quoted by pundit and layman alike. Equally popular in the Malayalam region is the 18th-century folk poet of *tuḷḷal*s (a song–dance form), Kuñcan Nampiyār, unparalleled for his wit and exuberance, his satiric sketches of caste types, his versions of Sanskrit *Purāṇa* narratives projected on the backdrop of Kerala, and his humorous renderings even of mythic characters.

The 17th and 18th centuries also saw Tamil court poetry—*Purāṇas*, translations from the Sanskrit, and praise poems, known more for their learning and imitative character than for their genius. This was also a period of many schisms and the founding of monasteries in Śaivism and Vaiṣṇavism, which led to many sectarian and polemic works. Muslims and Christians also wrote epics in the Hindu *Purāṇa* style; for example, Umaṟu-p-pulavar's 17th-century *Cīṟā-p-purāṇam*, on the life of the prophet Muḥammad, and Father Beschi's *Tēmpāvaṇi*, on the life of St. Joseph, with echoes from both Kampaṉ and the 16th-century Italian poet Torquato Tasso.

Probably the most impressive Tamil poetry of this period is that of Arunakiriv's learned and melodious *Tiruppukaḻ* (praise of Munikaṉ) and of the Cittars, eclectic mystics known for their radical, fierce folk songs and common-speech style. Tāyumāṉavar (18th century) and Paṭṭiṉattār (and later, in the 19th century, Rāmaliṅkar) are poets of unconditioned love, self-search, and rejection of corrupt society.

The 17th and 18th centuries are also periods of datable folk expression, which include many *tiruviḷaiyāṭal* ("stories of God's sport") *purāṇas;* temple tales (about miracles that took place in the temple); *kuṟavañci* (*i.e.*, "gypsy," a kind of musical dance drama); *paḷḷu*s (plays about village agricultural life); realistic *noṇṭi-nāṭakam*s ("dramas of the lame"), in which a Hindu temple god cures lameness; *kummi* songs sung by young girls, clapping as they dance round and round; and *ammāṉai* ballads. Noteworthy historical ballads are *Kaṭṭa Pommaṉ*, about a chieftain who revolted against the British, and *Tēciṅku-rācaṉ Katai*, about the prince of Gingi and his Muslim friend. Malayalam *āṭṭakkatha*, the literature associated with *kathākali*, the complex traditional dance drama, was also written during this period. Royal poets such as Kōṭṭayattu Tampurān, in the 17th century, and Kārttika Tiruṉal, in the 18th, wrote *āṭṭakkathās*.

(A.K.R.)

INDO-ARYAN LITERATURES: 12TH–18TH CENTURY

It is difficult to pinpoint the time when the Indo-Aryan dialects first became identifiable as languages. Around the 10th century AD, Sanskrit was still the language of high culture and serious literature, as well as the language of ritual. The spoken language, however, had continued to develop, and at the turn of the millennium there began to appear, at different times during the subsequent two or three centuries, the languages now known as the regional languages of the subcontinent: Hindi, Bengali, Kashmiri, Punjabi, Rajasthani, Marathi, Gujarati, Oriya, Sindhi (which did not develop an appreciable literature), and Assamese; Urdu did not develop until much later (see below *Islāmic literatures: 11th–19th century*).

Regional languages

The literatures in their early stages show three characteristics: first, a debt to Sanskrit that can be seen in their use of Sanskrit lexicon and imagery, in their use of myth and story preserved in that refined language, and frequently in their conformity to ideals and values put forward in Sanskrit texts of poetics and philosophy; second, a less obvious debt to their immediate Apabhraṃśa past (dialects that are immediate predecessors of the modern Indo-Aryan vernaculars); third, regional peculiarities.

The narratives in the early stages of the development of the languages are most often mythological tales drawn from the epics and *Purāṇas* of classical Hindu tradition (see above *Sanskrit, Pāli, and Prākrit literatures: 1200 BC–AD 1200*), though in later times, in the 17th and 18th centuries, secular romances and heroic tales were also treated in narrative poems. Although the themes of the narratives are based on *Purāṇa* tales, often they include materials peculiar to the area in which the narrative was written.

In addition to themes, regional literatures frequently borrowed forms from the Sanskrit; for example, the *Rāmāyaṇa* appears in a 16th-century Hindi version by Tulsīdās, called the *Rāmcaritmānas* ("Lake of Rāma's Deeds"), which has the same form, though a different emphasis, as the Sanskrit poem. The stylized conventions and imagery of Sanskrit court poetry also appear, though here, too, with different emphasis; for example, in the work of the 15th-century Maithili (Eastern Hindi) lyric poet Vidyāpati. Even the somewhat abstruse rhetorical speculations of the Sanskritic poetic schools of analysis were used as formulas for the production of 17th-century Hindi court poetry; the *Rasika-priyā* of Keśavadāsa is a good example of this kind of tour de force.

There are other characteristics common to the regional literatures, some of which come not from Sanskrit but most likely from the Apabhraṃśa. There are two poetic forms, for example, that are found in many northern Indian languages: the *bārah-māsā* ("twelve months"), in which 12 beauties of a girl or 12 attributes of a deity might be extolled by relating them to the characteristics of each month of the year; and the *caūtīs* ("thirty-four"), in which the 34 consonants of the northern Indian Devanāgarī alphabet are used as the initial letters of a poem of 34 lines or stanzas, describing 34 joys of love, 34 attributes, and so on.

Finally, there are common characteristics that may have come either through Apabhraṃśa or through the transmission of stories and texts from one language to another. The stories of Gopi-candra, the cult hero of the Nātha Yogī sect, a school of mendicant *samnyasin*s, were known from Bengal to the Punjab even in the early period. And the story of the Rājput heroine Padmāvatī, originally a romance, was beautifully recorded, with a Ṣūfī (mystic) twist, by the 16th-century Muslim Hindi poet Malik Muḥammad Jāyasī and later by the 17th-century Bengali Muslim poet Ālāol. From the late 13th through the 17th century, *bhakti* (devotional) poetry took hold in one region after another in northern and eastern India. Beginning with the *Jñāneśvarī*, a Marathi verse commentary on the *Bhagavadgītā* written by Jñāneśvara (Jñānadeva) in the late 13th century, the devotional movement spread through Mahārāshtra, in the works of the poet-saints Nāmdev and Tukārām; through Rājasthān, where it is represented by the works of Mīrā Bāī; through northern India, in the poetry of Tulsīdās, Sūrdās, Kabīr, and others; through Mithilā, in the work of the great poet Vidyāpati; and into Bengal, where Caṇḍīdās and others sang of their love of God. Because of the *bhakti* movement, beautiful lyric poetry and passionate devotional song were created; and in some cases, as in Bengal, serious philosophical works and biographies were written for the first time in a regional language rather than in Sanskrit. The languages and their literatures gained strength as mediums of self-expression as well as exposition. And, although there is much Sanskrit imagery and expression in the poetry and song, as well as similarities to Sanskrit textual models, its basic character is not Sanskritic: true to the nature of any spoken, everyday language, it is more vital than polished, more vivid than refined.

Spread of *bhakti* poetry

One more historical generality can be stated regarding regional Indian literature before considering the characteristics peculiar to the several "Indian literatures." In all of the early literatures, writing was lyrical, narrative, or didactic, entirely in verse, and all in some way related to religion or love or both. In the 16th century, prose texts, such as the Assamese histories known as the *buranji* texts, began to appear.

Hindi. What is commonly spoken of as Hindu is actually a range of languages, from Maithili in the east to Rajasthani in the west. The first major work in Hindi is the 12th-century epic poem *Pṛthvīrāj Rāsau*, by Chand Bardaī of Lahore, which recounts the feats of Pṛthvīrāj, the last Hindu king of Delhi before the Islāmic invasions. The work evolved from the bardic tradition maintained at the courts of the Rājputs. Noteworthy also is the poetry of the Persian poet Amīr Khosrow, who wrote in the Awadhi dialect. Most of the literature in Hindi is religious in inspiration; in the late 15th and early 16th centu-

ries, the reform-minded Kabīr, for example, wrote sturdy short poems in which he sought to reconcile Islām and Hinduism.

The most celebrated author in Hindi is Tulsīdās of Rājāpur (died 1623), a Brahmin who renounced the world early in life and spent his days in Benares (Vārānasi) as a religious devotee. He wrote much, mostly in Awadhi, and focussed Hinduism on the worship of Rāma. His most important work is the *Rāmacaritmānas* ("Mountain Lake of Rāma's Deeds"), which is based on the Sanskrit *Rāmāyaṇa*. More than any other work it has become a Hindu sacred text for the Hindi-speaking area and annually has been staged in the popular Rām Līlā festival.

Outstanding among the followers of Vallabha, philosopher and *bhakti* ("devotion") advocate of the Middle Ages, is the blind poet Sūrdās (died 1563), who composed countless *bhajan*s (chants) in praise of Kṛṣṇa and Rādhā, which are collected in the *Sūrsāgar* ("Ocean of Sūrdās"). While many of the *bhakti* poets were of modest origin, an exception was Mīrā Bāī, a princess of Jodhpur, who wrote her famous lyrics both in Hindi and Gujarati; the quality of her poetry, still very popular, is not as high, however, as that of Sūrdās. Significant also is the religious epic *Padmāvatī* by Jāyasī, a Muslim from former Oudh state. Written in Awadhi (*c.* 1540), the epic is composed according to the conventions of Sanskrit poetics.

The 18th century saw the beginning of a gradual transformation from the older forms of religious lyric and epic to new literary forms influenced by Western models that began to be known. The new trends reached their pinnacle in the work of Prem Chand (died 1936), whose novels (especially *Godān*) and short stories depict common rural life; and in the work of Harishchandra of Benares (died 1885), honoured as Bhāratendu (Moon of India), who wrote in the Braj Bhasa dialect.

Bengali. While developments in Bengali literature began somewhat earlier, they followed the same general course as those in Hindi. The oldest documents are Buddhist didactic texts, called *caryā-pada*s ("lines on proper practice"), which have been dated to the 10th and 11th centuries and are the oldest testimony to literature in any Indo-Aryan language.

The three genres of Bengali poetry

Bengali poetry, including poetry by Bengalis in other dialects, is largely written in three distinct genres. It is certain that well before the 15th century there existed texts in a typically Bengali genre called *maṅgal-kāvya* ("poetry of an auspicious happening"), which consists of eulogies of gods and goddesses; such poetry is likely to have had a considerable history in oral transmission before it was committed to writing. A good example of an orally transmitted *maṅgal* poem is the *Caṇḍī-mangal* ("Poem of the Goddess Caṇḍī"), by Mukundarāma, which was put into written form in the latter part of the 15th century. *Maṅgal* poetry remained a favourite genre well into the 18th century, when Bhārat-candra wrote the *Annadā-maṅgal* ("Maṅgal of the Goddess Annadā [the Giver of Food]"), a witty and sophisticated poem that bears little resemblance to its more rustic forebears. Despite this popularity, it is the devotional lyrics to the divine pair Kṛṣṇa and Rādhā that are still known and sung today in Bengal, and these lyrics are the gems of medieval Bengali literature.

Poems of the second genre, the *mahākāvya* ("great poem," but not to be confused with the Sanskrit *mahākāvya* genre), are based mainly on the Sanskrit models of the *Mahābhārata*, *Rāmāyaṇa*, and *Purāṇas*. Kṛttibās Ojhā (late 14th century) stands at the beginning of this literature; he wrote a version of the *Rāmāyaṇa* that often differs from the Sanskrit original, for he includes many local legends and places the setting in Bengal. Kavīndrā (died 1525) wrote on the *Mahābhārata* theme, as did Kāsiram Dās in the 17th century.

The third genre, *padāvalī* ("string of verse") songs, is also found elsewhere; inspired by the religious *bhakti* movement, the songs resemble the devotional poetry of the Nāyaṇārs and Āḻvārs in Tamil. It was such poetry that established Bengali as a significant literary language. The earliest work in what may be considered a distinctively Bengali style is the *Śrīkṛṣṇa-kīrtana* ("Praise of the Lord Kṛṣṇa"), a long *padāvalī* poem by Baṛu-Caṇḍīdās, which is dated to the early 15th century. In it the poet praises the virtues and celebrates the loves of Kṛṣṇa, a theme that had remained popular in Bengal ever since its first glorification by the Bengali Sanskrit poet Jayadeva, who composed his *Gītagovinda* ("The Cowherd Song") in the 12th century. *Padāvalī* songs describe and glorify all phases of Kṛṣṇa's love for the cowherds' wives (especially Rādhā, who later became a goddess), and it is love poetry before it is religious poetry. After the great Bengali mystic and saint Caitanya (died 1533), love *is* religion, and the erotic is inspirited with religious fervour. The great flowering of this poetry occurred in the 16th and 17th centuries.

Religious edification took the forms not only of *maṅgal*s and *padāvalī*s but also of biography (more like hagiography) and philosophy. Important in that style is the long hagiography *Caitanya-caritāmṛta* ("Elixir of the Life of Caitanya"), by the 16th-century author Kṛṣṇadās.

While most of the literature is Hindu in theme and inspiration, there arose a secular Bengali literature among Bengali Muslims. One of the outstanding Muslim poets is Ālāol, author of the *Padmāvatī* (*c.* 1648), which was written after the poem of the same name by the Hindi poet Jāyasī.

Assamese. The earliest text in a language that is incontestably Assamese is the *Prahlāda-caritra* of Mena Sarasvati (13th century); in a heavily Sanskritized style it tells the story, from the *Viṣṇu-Purāṇa*, of how the mythical king Prahlāda's faith and devotion to Viṣṇu saved him from destruction and restored the moral order. The first great Assamese poet was Kavirāja Mādhava Kandalī (14th century), who translated the Sanskrit *Rāmāyaṇa* and wrote *Devajit*, a narrative on the god Kṛṣṇa. In Assamese, too, the *bhakti* movement brought with it a great literary upsurge; the most famous Assamese poet of the period was Śaṅkaradeva (died 1568), whose 27 works of poetry and devotion are alive today and who inspired such poets as Mādhava-devi to write lyrics of great beauty. Peculiar to Assamese literature are the *buranji*s, chronicles written in a prose tradition brought to Assam by the Ahoms of Burma. These date in Assamese from the 16th century, while in the Ahom language they are much earlier.

Oriya. *Mādaḷā-pāñji* ("The Drum Chronicle") texts in Oriya, the chronicles of the great temples of Jagannātha in Puri, date from the 12th century. They are in prose, and as such they represent the earliest prose in a regional Indo-Aryan language, although they cannot be said to be literary texts. The 14th century was productive for Oriya literature. Dating from this period are the anonymous *Kalasa-cautīśa*, which tells in 34 verses the story of the marriage of the god Śiva and the mountain goddess Pārvatī, and the famous *Caṇḍī-purāṇa* of Saraladāsa. But the *bhakti* period was once again the most stimulating one; the best known medieval Oriya poet is Jagannātha Dās (whose name means Servant of Jagannātha), a 16th-century disciple of the Bengali Vaiṣṇava saint Caitanya, who spent the better part of his life in Puri. Among the many works of Jagannātha Dās is a version of the Sanskrit *Bhāgavata-Purāṇa* that is still popular in Orissa today.

Marathi. With Bengali, Marathi is the oldest of the regional literatures in Indo-Aryan, dating from about AD 1000. In the 13th century, two Brahminical sects arose, the Mahānubhāva and the Varakari Panth, both of which put forth vast quantities of literature. The latter sect was perhaps the more productive, for it became associated with *bhakti*, when that movement stirred Mahārāshtra in the early 14th century, and particularly with the popular cult of Viṭṭhoba at Pandharpur. It was out of this tradition that the great names of early Marathi literature came: Jñāneśvara, in the 13th century; Nāmdev, his younger contemporary, some of whose devotional songs are included in the holy book of the Sikhs, the *Ādi Granth;* and the 16th-century writer Eknāth, whose most famous work is a Marathi version of the 11th book of the *Bhāgavata-Purāṇa*. Among the *bhakti* poets of Mahārāshtra the most famous is Tukārām, who wrote in the

16th century. A unique contribution of Marathi is the tradition of *povāḍās*, heroic stories popular among a martial people. There is no way of dating the earliest of these; but the literary tradition is particularly vital at the time of Śivajī, the great military leader of Mahārāshtra (born 1630), who led his armies against the might of the Mughal emperor Aurangzeb.

Tradition of *povāḍās*

Gujarati. The oldest examples of Gujarati date from the writings of the 12th-century Jaina scholar and saint Hemacandra. The language had fully developed by the late 12th century. There are works extant from the middle of the 14th century, didactic texts written in prose by Jaina monks; one such is the *Bālāvabodha* ("Instructions to the Young"), by Taruṇa-prabha. A non-Jaina text from the same period is the *Vasanta-vilāsa* ("The Joys of Spring"). The two Gujarati *bhakti* poets, both of the 15th century, are Narasiṃha Mahatā (or Mehtā) and Bhālaṇa (or Puruṣottama Mahārāja); the latter cast the 10th book of the *Bhāgavata-Purāṇa* into short lyrics. By far the most famous of the *bhakti* poets is the woman saint Mīrā Bāī, who lived in the first half of the 16th century. Mīrā, though married, thought of Kṛṣṇa as her true husband, and the lyrics telling of her relationship with her god and lover are among the warmest and most movingly personal in any Indian literature. One of the best known of the non-*bhakti* Gujarati poets is Premānanda Bhaṭṭa (16th century), who wrote narrative poems based on *Purāṇa*-like tales; although his themes were conventional, his characters were real and vital, and he infused new life into the literature of his language.

Punjabi. Punjabi developed a literature later than most of the other regional languages of the subcontinent; and some of the early writings, such as those of the first Sikh Gurū, Nānak (late 15th and early 16th centuries), are in Old Hindi rather than true Punjabi. The first work identifiable as Punjabi is the *Janam-sākhī*, a 16th-century biography of Gurū Nānak by Bala. In 1604, Arjun, the fifth Gurū of the Sikhs, collected the poems of Nānak and others into what is certainly the most famous book to originate in the Punjab (though its language is not entirely Punjabi), the *Adi Granth* ("First Book"). Writing that is not merely incidentally Punjabi began in the 17th century and is almost entirely by Muslims. Between 1616 and 1666, a writer named ʿAbdullāh, for example, composed a major work called *Bāra Anva* ("Twelve Topics"), which is a treatise on Islām in 9,000 couplets. Muslim Ṣūfīs, such as Bullhē Shāh (died 1758), also contributed many devotional lyrics, and Ṣūfī Islām can be said to have been the main stimulus to Punjabi literature in the medieval period. There are also many romances in the language (as in Rajasthani) which, being oral literature, are undatable.

Kashmiri. The hitherto commonly accepted period of Old Kashmiri is 1200–1500; but in fact the earliest example of the language is found in 94 four-line stanzas embedded in the Sanskrit philosophical work *Mahānaya-prakāśa* ("Illumination of the Highest Attainment"), which some scholars now date as late as the 15th century. As is true for Gujarati, the most famous poets of Kashmiri in this period are women. Lallā (14th century) wrote poems about the god Śiva; and Hubb Khātun (16th century) and especially Arani-mal (18th century) are famous for their hauntingly beautiful love lyrics. Despite these outstanding poets in Kashmiri, the great literary language of Kashmir in the medieval period was Persian, which was encouraged by many rulers of the country, such as Zayn-ul-ʿĀbidīn, in whose 15th-century court were many scholars and poets writing in both the Kashmiri and Persian languages. (E.C.D.)

ISLAMIC LITERATURES: 11TH–19TH CENTURY

The adventure of Islām in India began in the 8th century with the conquest of Sind (the extreme western province), but it was only in the 11th and 12th centuries that Muslim literary and cultural traditions reached the Indian heartland. Then, in the 13th century, refugee noblemen, soldiers, and men of letters from Iran and Central Asia came pouring into India. Although the causes changed, the attraction of India as a place of refuge and gracious patronage did not decline for several subsequent centuries. At the same time Muslim soldier-adventurers continued with their conquests, joining hands with their non-Muslim Indian counterparts in many instances, establishing minor or major kingdoms all over the subcontinent. The political map of India remained very much in flux—except for a brief period during the reign of Akbar—until Queen Victoria declared herself empress of India in 1858. The period of Muslim influence thus extends over 800 years.

At the time of the spread of Muslim power and culture in India, Sanskrit was the chief language of Hindu cultural, learned, and religious expression, while Buddhism and Jainism had lent their prestige and patronage to various Prākrits. The progress of and developments in these literatures remained unaffected by the advent of Islām in India. The emergence of the new Indo-Aryan languages out of the Prākrit and Apabhraṃśa stages of Sanskrit, however, was furthered by the newcomers, who preferred these regional languages over Sanskrit and encouraged the development of popular regional literatures. The conversion to Islām of a large number of indigenous people enhanced these developments. Thus, the vehicles of literary expression used by those professing Islām in India were regional dialects and languages, both Indo-Aryan and Indo-Iranian, such as Braj, Awadhi, Sindhi, Baluchi, Urdu, Dakhini, and Bengali, as well as the foreign Arabic, Turkish, and Persian spoken by the Muslim immigrants and conquerors. Of these, only Persian and Urdu require detailed consideration; the others will be discussed only briefly.

Effect of Islām on Sanskrit and Prākrit literatures

Arabic. Arabic was the language of the conquerors of Sind. But it enjoyed more permanent prestige as the language of the Qur'ān, the sacred book of Islām; as such it was extensively used for religious scholarship during the medieval period. Even as late as the 18th century, Shāh Walī Allāh, the greatest theologian to have lived in India, wrote his most important treatises in Arabic. Arabic was also used early for historiography and for making Indian scientific books available to the Middle East in translation. One does not find, however, much in the way of significant Arabic belles lettres in India.

Turkish. Although the earliest Muslim conquerors in northern India were Turks, their language was Persian. It was only during the reigns of Bābur and his son Humāyūn (1526–56) that Turkish flourished for a while as a medium of learned expression. Bābur himself was the foremost contributor. Although his memoirs are better known, he also left a volume of verses of considerable merit.

Regional languages. The literatures of the Indo-Iranian languages of Baluchi and Pashto are exclusively creations of Muslim writers. In the Indo-Aryan languages of Kashmiri, Sindhi, and Punjabi, Muslims were the most influential contributors; the names of Lallā (14th century) for Kashmiri, Shāh ʿAbd-ul-Laṭīf (17th–18th century) for Sindhi, and Wārīs Shāh (18th century) for Punjabi exemplify that fact. Muslim chieftains gave impetus to the growth of Bengali literature through their patronage of writers and through their efforts to have Sanskrit classics translated into Bengali. There are also many famous Muslim names during the medieval period of Bengali literature, such as Dawlat Qāzī and Ālāol in the 17th century. In the heartland of northern India, notable contributions were made by Muslims to Hindi literatures in the Braj and Awadhi dialects. Malik Muḥammad Jāyasī, Raḥīm, and Manjhan (all 16th century) and ʿUsman (17th century) are some of the important names. In the 16th, 17th, and 18th centuries in India there was a tremendous production of mystic (Ṣūfī and *bhakti*) poetry in all of the important dialects and languages. It was a period of great mystic, syncretic movements, and the Muslim contribution in the form of love narratives and lyrics was considerable. Quite often metres, motifs, and assorted rhetorical features of Persian *maṣnavī*s and *ghazal*s (see below *Urdu*) were used in a new medium. Moreover, interaction and assimilation took place between the Muslim Ṣūfī traditions, thought, and practices and the Indian *bhakti* schools. Muslim

bhakti poets either expressed Ṣūfī ideas, which were close to monotheistic orthodoxy as well as to the doctrines of Indian saints Kabīr and Nānak, in the Indian dialects through narrative poems modelled on Persian *maṣnavī*s or chose the path of ecstasy and became devotees of Kṛṣṇa (which was still close to the more orthodox forms of Ṣūfīsm). None of them followed the devotional style of Tulsīdās, their contemporary and a devotee of Rāma.

It was, however, in Persian and Urdu that Muslim men of letters made the greatest contributions—contributions that led in the former case to the establishment of an "Indian" school of Persian poetry and influenced profoundly the development of poetry in Afghanistan and Tadzhikistan and, in the latter case, led to the emergence of a unique pan-Indian language and literature in Urdu.

Persian. Maḥmūd of Ghazna, with whom the chain of Muslim conquests in northern India began, was also the patron of Ferdowsī, one of the greatest of Persian poets. The later conquerors admired literature no less. Since the language of all of them was Persian, the growth of Persian literature in India kept pace with its conquest by the Muslims.

The poems of Amīr Khosrow

Masʿūd Saʿd Salmān (born 1046 in Lahore), who later became the governor of Jullundhur, was the first noteworthy person of Indian origin to have written poetry in Persian. The first truly great poet was Amīr Khosrow, who wrote in the 13th and 14th centuries. Of Turkish descent, born in the district of Etah in northern India, Khosrow was connected with royal courts all his life, even after 1272, when he became a disciple of the great mystic Niẓām-ud-Dīn Awliyā. He wrote five books of poems, or *dīvān*s, composed of *ghazal*s (see below *Urdu*), panegyrics and several *maṣnavī*s—altogether some 200,000 couplets. In poetry, his innovative spirit displayed itself in *waṣf-nigārī*—that is, descriptions of natural objects in short poems, which Khosrow incorporated within longer ones. His keenness of observation is also evident in his use of local fauna and flora as poetic images. Khosrow's distinction lies not so much in the fact that he is an innovator, however, as in the fact that he is equally superb in narrative poetry, panegyrics, and lyrics. The range of his popularity and influence can best be gauged by the fact that, in northern Indian folk literature, one comes across numerous songs and riddles consistently attributed to Amīr Khosrow.

In the centuries that followed Khosrow, until the end of the Islāmic period, India contributed to Persian literature in two ways: first, through the production of dictionaries that helped to standardize the language and consolidate its vocabulary; second, through the development of the so-called Indian style of Persian poetry.

It is generally agreed that this Indian style, *sabk-e hindī*, did not originate within the geographic confines of India, though it reached its most sublime form there at the hands of poets who either were born in India or spent their most productive years at various Indian courts. Some of the characteristics of the style are (in the words of one modern critic) the emphasis on

> parallel statement . . . ; on complex conceit like that of the seventeenth century English "metaphysical" poets, arising out of economy of expression and telescoping into a single image a variety of emotional states; on "cerebral" artifice in pushing familiar images to unfamiliar and unexpected lengths; and on the creation of a synthetic poetic diction in which a whole phrase constitutes a single image.

The keen observation of daily life that is also characteristic of Indian Persian poetry could have been inspired by the traditions of classical Sanskrit poetry, with which these poets must have been familiar through the extensive translations done during the reign of the Mughals.

The century (1556–1657) of the reigns of Akbar, Jahāngīr, and Shāh Jahān was the most glorious period for Persian poetry in India, though, except for Fayẓī, all of the important poets were immigrants from Persia who found relief from religious and political persecution as well as generous patronage at the hands of the great Mughals and the lesser kings of southern India. The great men of letters of that period were ʿUrfī, Ṭālib Āmulī, Naẓīrī, Ẓuhūrī, Kalīm, and Ṣāʾib.

The greatest poet of the Indian style, however, was ʿAbdul Qādir Bēdil, born in 1644 in Patna, of Uzbek descent. He came early under the influence of the Ṣūfīs, refused to be attached to any court, and travelled widely throughout India during his long life. Bēdil's 16 books of poetry contain nearly 147,000 verses and include several *maṣnavī*s. Though ignored by the Iranians, Bēdil's poetry had an impact on Tadzhik and Uzbek literatures, and its influence is still evident in Afghanistan. A poet of great virtuosity and philosophic bent, he was well acquainted with Indian religions and philosophy. His anti-feudal views and his critical and skeptical attitude toward all kinds of dogma make his poetry relevant even today. His style is difficult, his metaphors and syntax quite complex (though the language itself is quite simple); and yet, as a modern critic puts it, "the intensity of his subjective assessment is so acute and factual, and his metaphysical experience so intense, that genuine poetry emerges in all its splendour."

The works of ʿAbdul Qādir Bēdil

Urdu. Earlier varieties of Urdu, variously known as Gujari, Hindawi, and Dakhani, show more affinity with eastern Punjabi and Haryani than with Khari Boli, which provides the grammatical structure of standard modern Urdu. The reasons for putting together the literary products of these dialects, forming a continuous tradition with those in Urdu, are as follows: first, they share a common milieu, consisting of Ṣūfī and Muslim court culture, increasingly dominated by the life and values of the urban elite; second, they display wholesale acceptance of Perso-Arabic literary traditions, including genres, metres, and rhetoric; third, they show an increasing acceptance of Perso-Arabic grammatical devices and vocabulary; and fourth, they tend to prefer Perso-Arabic forms over indigenous forms for learned usage.

Apart from themes and metaphysics, the influence of Ṣūfī hospices and royal courts can be seen in two practices that were essential to the development of Urdu poetry (and also unique to the Urdu milieu in the medieval period) and that still exist in modified forms. First, Urdu poets generally chose an *ustād*, or master, just as a Ṣūfī novice chose a *murshid*, or preceptor, and one's poetic genealogy was always a matter of much pride. Second, poets read poetry in private or semiprivate gatherings, called *mushāʿirah*, which displayed hierarchies, status consciousness, and rivalries reminiscent of royal courts.

Urdu literature began to develop in the 16th century, in and around the courts of the Quṭb Shāhī and ʿĀdil Shāhī, kings of Golconda and Bījāpur in the Deccan (central India). In the later part of the 17th century, Aurangābād became the centre of Urdu literary activities. There was much movement of the literati and the elite between Delhi and Aurangābād, and it needed only the genius of Walī Aurangābādī, in the early 18th century, to bridge the linguistic gap between Delhi and the Deccan and to persuade the poets of Delhi to take writing in Urdu seriously. In the 18th century, with the migration of poets from Delhi, Lucknow became another important centre of Urdu poetry, though Delhi never lost its prominence.

The first three centuries are dominated by poetry. Urdu prose truly began only in the 19th century, with translations of Persian *dāstān*s, books prepared at the Delhi College and the Fort William College at Calcutta, and later with the writers of the Aligarh movement.

To focus on essential matters, the discussion that follows forgoes a chronological account of the poetry, concentrating instead on characteristics of particular genres and the achievements of the most significant of their practitioners up to 1857. There is one poet, however, who cannot be described as a practitioner of the classical Perso-Arabic traditions adopted by his fellow poets. Naẓīr Akbarābādī, who wrote in the late 18th and early 19th centuries, was a poet of consummate skill who chose to display it in short poems (in various forms) written in the language of popular speech as well as of literature. His themes show similar eclecticism. In his voluminous body of work, there are poems on such diverse topics as popular festivals, the seasons, the vanities of life, erotic pleasures and pursuits, dancing bears, and niggardly merchants. He is a master of the telling detail that immedi-

The works of Naẓīr Akbarābādī

ately brings any event to life. Generally ignored by elitist poets and literary chroniclers of his time, Nazīr has gained increasing respect and recognition as the first and best poet of the people.

Qasīdahs. *Qasīdah*s are poems written with a "purpose"—the purpose being worldly gain, in the case of poems praising kings and noblemen, or benefit in the afterworld, in the case of poems praising God, the prophet Muḥammad, and other holy personages. These panegyrics are generally overly long and are written in a highly ornate and hyperbolic style, the poets vying to display their prowess by using as many rhymes and discovering as many associative themes as possible. Because of their style and language they are of special interest to lexicographers. Not much scholarly work has been done on the *qasidah*s written in the Deccan, but in northern India a number of poets are regarded highly for their achievements in this genre: in the 18th century, Sawdā and Inshā', and in the 19th, Żawq and Ghālib.

Ḥaju and shahr-āshūb. Less ornate, if not less elaborate, and more edifying are the *ḥaju* (derogatory verses, personal and otherwise) and the *shahr-āshūb* (poems lamenting the decline or destruction of a city). They provide useful information about the mores and morals of the period from the 18th to mid-19th century and truly depict the problems facing the society at large. The poems are not formally restricted to any particular metre or stanza pattern. Sawdā again is one of the more famous names.

Urdu lamentatory elegies

Marsiyeh. *Marsiyeh* means "elegy," but in Urdu literature it generally means an elegy on the travails of the family and kinsmen of Ḥusayn (grandson of Muḥammad) and their martyrdom in the field of Karbalā, Iraq. These elegies and other lamentatory verses were read at public gatherings, especially during the month of Muḥarram. Although a large number of *marsiyeh*s were written in the Deccan and at Delhi, it was in Lucknow, with the patronage of Shī'ite elite and royalty, that *marsiyeh*s gained the tenor and magnitude of epic poetry. The two great masters of that 19th-century period were Mīr Anīs and Mīrzā Dabīr, who together established *musadda*s (a six-line stanza with an *aaaa bb* rhyme scheme) as the preferred form for *marsiyeh*s and added several new topics and details to the ranks of associated themes, thus carrying the form beyond a simple lament. An interesting aspect of these elegies is that, although the scene and personae are Arab, there is no attempt at verisimilitude: Arab gallants and maidens speak and gesture like the elites of Lucknow. Perhaps this added to the pathos and effectiveness of the poems at public readings.

Mašnavī. *Mašnavī* was the preferred genre for all descriptive and narrative purposes, for it allows the most freedom (only the lines of each couplet must rhyme). In the Deccan, all major poets wrote at least one long *mašnavī*, but lack of knowledge of the dialect has prevented their full appreciation. Thus, the more famous *mašnavī*s are by later poets of Delhi and Lucknow, such as Mīr, Mīr Ḥasan, Dayā Shankar Nasīm, and Mīrzā Shawq. The topics of descriptive *mašnavī*s range from mundane events of life, hunting trips of kings, and the vagaries of nature's seasons to autobiographical discourses. Narrative *mašnavī*s are considerably longer, running into hundreds of couplets. In the Deccan several poets wrote abridged versions of Persian *mašnavī*s, but many others wrote original compositions utilizing Indian romances as well as the better known Persian and Arabic ones. Apart from the names of the protagonists in the *mašnavī*s inspired by Persian and Arabic poems, all else is always local; the landscape, cityscape, processions, customs and rituals, social values and taboos, even the physical characteristics of the people are totally Indian, though dominantly Muslim and feudal. Despite their length, these narratives gained much popularity and, at least in northern India, were often read in public places, in much the same way as storytellers told stories. The *mašnavī* form was also used by some of the Hindi Ṣūfī poets.

Ghazal. For the most part, the history of Urdu poetry in India is the story of Urdu *ghazal*, which has been the favourite of both poets and their audiences in every period. A short lyric, with prosodic requirements of both metre and rhyme, *ghazal* demands great skill and thought from the poet, for its couplet must be a complete semantic entity and fully express a whole, well-integrated poetic experience. Favourite themes are erotic love, Ṣūfī love, and metaphysics. Naturally, Urdu poets began by closely imitating, often even plagiarizing, Persian masters, but later on they spoke in a more authentic voice. They continued, however, to employ a vocabulary of love that owed almost everything to Persian and shared very little with the traditions of lyrical poetry in other Indian languages. For example, with few exceptions, the lover is always masculine; expression of love is never made by a woman. Unique, too, is the use of masculine grammatical forms and imagery for the beloved, even when, in every other way, the poem is clearly celebrating heterosexual love. This peculiarity, as well as other traditions borrowed from Persian masters, gives a *ghazal* couplet a tremendously wide range of interpretations. It is amazing indeed what a master poet can condense into one terse couplet.

Characteristics of the *ghazal*

The two greatest *ghazal* writers in Urdu are Mīr Taqī Mīr, in the 18th century, and Mīrzā Asadullāh Khān Ghālib, in the 19th. They are in some ways diametrical opposites. The first prefers either very long metres or very short, employs a simple, non-Persianized language, and restricts himself to affairs of the heart. The other writes in metres of moderate length, uses a highly Persianized vocabulary, and ranges wide in ideas. Mīr speaks of passion and pathos; Ghālib betrays a skeptic's mind and leaves nothing unquestioned, not even his feelings. But both have left indelible marks on the ideas and emotions of succeeding generations. Ghālib wrote poetry in Persian as well as Urdu and also published a couple of volumes of letters in Urdu that helped usher in modern prose. In many ways he bridges the gap separating the medieval sensibility from the modern. The contemporary mind, however, is also moved by the authentic passion of Mīr, idolizing him for the sublimity of his concept of love and for his personal integrity. The poems of Ghālib and Mīr represent the best of the Urdu *ghazal;* and the Urdu *ghazal*, as an anonymous wit has remarked, is the Muslims' greatest gift to India, after the Tāj Mahal. (C.M.N.)

SINHALESE LITERATURE: 10TH CENTURY AD TO 19TH CENTURY

The island nation of Ceylon (now called Sri Lanka), formally a part of South Asia, has been little noticed by the subcontinent, apart from the fact that according to an uncertain tradition it is celebrated in the *Rāmāyaṇa* as the island called Laṅkā. Buddhist sway was introduced there early, during the reign of Aśoka Maurya (*c.* 269–232 BC); and, while on the subcontinent Buddhism prospered, declined, and finally disappeared, in Ceylon it has continued until today. Although there are obvious borrowings in Ceylon from subcontinental literature, notably Sanskrit, and there was rather precarious communication with India through the island's Hindu community of Tamils, Ceylon never became culturally continuous with the mainland. The language itself, although of Indo-Aryan stock, is strongly mixed with a substratum of Dravidian. Also, it was Ceylon's fate early to fall victim to European colonialism, first to the Portuguese, then to the Dutch, and finally to the British, before it regained nationhood in 1948.

While there are inscriptions that antedate the Christian Era, no texts appear to survive from before the 10th century AD. The first texts that emerged were aids in Sinhalese—glossaries, paraphrases, and the like—to the study of the Pāli texts of Buddhism. More interesting are Sinhalese renderings of the life and virtues of the Buddha. Important in this genre, hagiographical rather than literary, is the *Amāvatura* ("Flood of the Ambrosia"), by Guruḷugōmī, which in 18 chapters purports to narrate the life of the Buddha, with specific emphasis on one of his nine virtues—his capacity to tame recalcitrant people or forces. In a similar vein is the literature of devotion and counsel, in which Buddhist virtues are celebrated.

Exceptional in the context of the South Asian subconti-

nent is the early and persistent interest in historical records. Such interest had begun in Pāli with the *Dīpavaṃsa* ("Chronicle of the Island") and had continued with the *Mahāvaṃsa* ("Great Chronicle") and *Cūlavaṃsa* ("Lesser Chronicle"), but it had a life of its own in Sinhalese. The most important, and possibly the oldest, of such chronicles is the *Thūpavaṃsaya* ("Chronicle of the Great Stupa"), by Pārakrama Paṇḍita. Subsequent chronicles, or genealogies of place, comprise the history of all of the major Buddhist monuments. Several chronicles were also inspired by the Tooth Relic, received from Kaliṅga in the 4th century by King Kīrtiśrīmēghavarṇa and worshipped ever since. Such chronicling included that of the kings who protected the relic.

Historical records

All of this literature was mostly in prose, but poetry as a literary form no doubt antedated it, as evidenced by early inscriptions. Much poetry was occasioned by Pāli *Jātakas* (stories of the Buddha's previous births) and other Buddhist stories, though Hindu stories were not lacking; for example, a version of the Sanskrit *Mahābhārata* (received through a Tamil source) was cast in the style of a *Jātaka* in the *Mahāpadaraṅga-Jatakaya*.

Likewise of Hindu Indian origin was a genre that took off from the Sanskrit poet Kālidāsa's "Meghadūta" (see above *Classical Sanskrit kāvya* [*200–1200*]), in which an exiled lover sends a message to his beloved by way of a monsoon cloud, thus giving the poet the opportunity to dwell on the description of landmarks in a poetic travelogue. This genre, so-called *saṃdeśa* literature, by no means unknown on the mainland, proliferated widely on Ceylon.

Of a different style are panegyrics and war poems, the earliest of which is the *Parakumbasirita* ("History of Parakramabahu VI," king in Jayavardhanapura from 1410 to 1468). Again reminiscent of the mainland and the religious tradition are the plentiful eulogies of the Buddha. Popular, too, was didactic verse, among the most notable of which is the *Kusajataka*, 687 stanzas of epigrams and exempla by the 17th-century poet Alagiyavanna Mohoṭṭāla.

MODERN PERIOD: 19TH AND 20TH CENTURIES

The modern period was ushered in by the arrival of the British, the influence of Western models becoming discernible in the early 19th century. Reform-minded Hindus, led by Rammohan Ray, took a positive attitude to Western literature and urged on their countrymen a Western type of education. Newly formed literary clubs spread the influence of predominantly British works, thereby opening the Indian educated elite to Western culture and literature in general. After a period of translation, authors sought to imitate Western models and eventually to be independently creative in the new styles.

Effects of Westernization

The most striking result of Westernization was the introduction of prose on a major scale. Vernacular prose, rarely looked upon previously as a medium for art, was now used as a literary vehicle, and such hitherto unknown forms as the novel, novella, and short story began to emerge. In poetry the thrall of tradition was stronger, and verse in the older forms continued to be written. With modernity, realism appeared, as well as symbolism in some quarters, and there was new psychological and social interest.

From Bengal spread a new sense of national purpose, which became the principal motivation for much English as well as vernacular literature. Three trends can be distinguished in the products of this increasing literary activity. The old traditionalism was transformed into romanticism, which looked to the past, to Indian history, for inspiration and sought to preserve what was considered valuable in the past; a tendency to mysticism went hand in hand with the romantic mood (a mood that was also widespread in 19th-century Europe). Greater social awareness in European literature was reflected in the literature of Indian progressives, in whose works a somewhat romantic Marxism prevailed. Finally, there was a humanistic trend. The teachings of Mahatma Gandhi, combining social concerns with traditional ethics, later exerted a very great influence on literature.

In the years preceding and following India's independence (1947) and control of the princely states, the fervour of writers sometimes turned to an increasingly articulate progressivism of various Marxist schools; sometimes to disappointment and bitterness; and most recently, it appears, to a mood of introspection. These developments, which occurred with a different pace in different regions, are described briefly below. A complete coverage of the most modern literature has not been attempted, but an endeavour has been made to mention persons who are considered to be representative.

Bengali. Except for the iconoclastic poet Michael Madhusudan Datta, poetic activity in the mid-19th century was giving ground to experimentation with the new prose style learned from English. During this period, Bengali literature produced a spate of novels—satiric, social, and picaresque. While Michael's work *Mēghanādavadh* (1861; a long poem on the Rāma theme in which Rāma and Lakṣmaṇa become the villains and Rāvaṇa the hero) caused a stir, the literary event of the period was the appearance on the scene of Bankim Chandra Chatterjee, whose first novel, *Durgeśanandinī* ("Daughter of the Lord of the Fort"), appeared in 1865. While not at first overtly nationalist, Bankim Chandra became more and more an apologist for the Hindu position. In *Kṛṣṇacaritra*, Christ suffers in comparison with Kṛṣṇa, and in his best known work, *Ānanda-maṭh* (1892; "The Abbey of Bliss"), the motherland in the person of the goddess Durgā is extolled.

Chatterjee and Tagore

Perhaps first among novelists of the late 19th and early 20th centuries is Saratchandra Chatterjee, whose social concerns with the family and other homely issues made his work popular. But the early 20th century is certainly best known for the poet who towers head and shoulders above the rest, Rabindranath Tagore. Poet, playwright, novelist, painter, essayist, musician, social reformer, Rabindranath produced works, still not completely collected, that fill 26 substantial volumes. The winner of the Nobel Prize for Literature in 1913, primarily for his little books of songs called *Gītāñjali*, which were much praised by Ezra Pound and William Butler Yeats, Tagore is more known for these devotional poems than for the wit and clear thought with which his later work is filled. He was the last of an era, looking back as he did to the religious and political history of Bengal for his inspiration. Those who followed him were more concerned with introspection and dramatic imagery.

If Tagore was the last poet in the Bengali tradition, Jibanananda Das was the first of a new breed. Musing and melancholy, yet known for vivid and unusual imagery Jibanānada is a poet who has much influence on younger writers in Bengal. There have been many other poets in the 20th century who are equally powerful but stand somewhat apart from the mainstream. One of these was Sudhindranath Datta, a poet much like Pound in careful and etymological use of language; another is the poet and prose writer Buddhadeva Bose.

Bose has been termed a progressive, and indeed he consciously turned away from the tradition orientation of Tagore and sought inspiration in schools foreign to Bengal—for example, the French Symbolists. He is the leader of an artistic faction, the Kallol school, and editor of an influential literary magazine, *Kavitā*. Unjustifiably called obscene, his writing has been experimental, probing into social and psychological realities of Bengali life.

While there have been, and still are, literary factions associated with political positions, they have been less definitive than some in other parts of India. Bengali writers in the 20th century have tended to be personal and individual rather than propagandists for political positions.

Assamese. Assamese literature began with Hemchandra Baruwa, a satirist and playwright, author of the play *Bahiri-Rang-Chang Bhitare Kowabhaturi* (1861; "All That Glitters Is Not Gold"). The most outstanding among the early modern writers was Lakshminath Bezbaruwa, who founded a literary monthly, *Jōnāki* ("Moonlight"), in 1889, and was responsible for infusing Assamese letters with 19th-century Romanticism. Later

20th-century writers have tried to remain faithful to the ideals of *Jōnāki*. The short story in particular has flourished in the language; notable practitioners are Mahichandra Bora and Holiram Deka.

The year 1940 marked a shift toward psychology, but World War II effectively put an end to literary development. When writers resumed after the war, there was a clear break with the past, in experimental verse and the growth of the novel form.

Hindi. Modern Hindi literature began with Harischandra Bharatendu in poetry and drama, Mahavir Prasad Dvivedi in criticism and other prose writings, and Prem Chand in fiction. This period, the second half of the 19th century, saw mainly translations from Sanskrit, Bengali, and English. The growth of nationalism and social reform movements of the Arya Samaj led to the composition of long narrative poems, exemplified by those of Maithili Sharan Gupta; dramas, by those of Jayashankar Prasad; and historical novels, by those of Prasad, Chatureen Shastri, and Vrindavan Lal Varma. The novels drew mainly on the periods of the Maurya, Gupta, and Mughal empires.

Gandhi's influence on poetry

This period was followed by the Non-cooperation and *satyāgraha* movements of Mahatma Gandhi, which inspired poets such as Makhan Lal Chaturvedi, Gupta, and Subhadra Chauhan and novelists such as Prem Chand and Jainendra Kumar. Eventual disillusionment with Gandhian experiments and the increasing influence of Marxism on European literature influenced writers such as Yashpal, Rangeya Raghava, and Nagarjuna.

S.N. Pant, Prasad, Nirala, and Mahadevi Varma, the most creative poets of the 1930s, drew inspiration from the Romantic tradition in English and Bengali poetry and the mystic tradition of medieval Hindi poetry. Reacting against them were the Marxist poets Ram Vilas Sharma and Nagarjuna and experimentalists such as H.S. Vatsyayan "Agyeya" and Bharat Bhuti Agarwal. Nirala, who developed from a mystic–romantic into a realist and experimentalist, was the most outstanding poet of the 1950s; and Sarweshwar Dayal Saxena, Kailash Vajpeyi, and Raghubir Sahay were the most creative poets of the 1960s.

Two trends, represented by the work of Prem Chand and Jainendra Kumar, led Hindi fiction in two different directions: while social realists like Yashpal, Upendranath Ashk, Amritlal Nagar, Mohan Rakesh, Rajendra Yadav, Kamleshwar, Nagarjuna, and Renu faithfully analyzed the changing patterns of Indian society, writers such as Ila Chandra Joshi, "Agyeya," Dharm Vir Bharati, and Shrikant Varma explored the psychology of the individual, not necessarily within the Indian context.

Among the dramatists of the 1930s and 1940s were Govind Ballabh Pant and Seth Govind Das; because of their highly Sanskritized language, their plays have had a limited audience. Plays by minor writers such as Ramesh Mehta, however, are repeatedly staged by professional theatres. In between these extremes there are some notable playwrights.

Gujarati. In Gujarāt, too, the advent of British rule deeply influenced the literary scene. The year 1886 saw the *Kusumamālā* ("Garland of Flowers"), a collection of lyrics by Narsingh Rao. Other poets include Kalapi, Kant, and especially Nanalal, who experimented in free verse and was the first poet to eulogize Gandhi. Gandhi, himself a Gujarati, admonished poets to write for the masses and thus inaugurated a period of poetic concern with changes in the social order. Many incidents in Gandhi's life inspired the songs of poets. The Gandhi period in Gujarāt as elsewhere gave way to a period of progressivism in the class-conflict poetry of R.L. Meghani and Bhogilal Gandhi. In post-independence India, poetry has tended to become subjectivist and alienated without, however, fully superseding the traditional verse of devotion to God and love of nature.

Among novelists, Govardhanram stands out; his *Sarasvatīchandra* is a classic, the first social novel. In the novel form, too, the influence of Gandhiism is clearly felt, though not in the person of Kanaiyalal Munshi, who was critical of Gandhian ideology but still, in several *Purāṇa*-inspired works, tended to preach much the same message. In the period after independence the modernists embraced existentialistic, surrealistic, and symbolistic trends and gave voice to the same kind of alienation as the poets.

Marathi. The modern period in Marathi poetry began with Kesavasut and was influenced by 19th-century British Romanticism and liberalism, European nationalism, and the greatness of the history of Mahārāshtra. Kesavasut declared a revolt against traditional Marathi poetry and started a school, lasting until 1920, that emphasized home and nature, the glorious past, and pure lyricism. After that, the period was dominated by a group of poets called the Ravikiran Maṇḍal, who proclaimed that poetry was not for the erudite and sensitive but was instead a part of everyday life. Contemporary poetry, after 1945, seeks to explore man and his life in all its variety; it is subjective and personal and tries to speak colloquially.

Among modern dramatists, S.K. Kolhatkar and R.G. Gadkari are notable. Realism was first brought to the stage in the 20th century, by Mama Varerkar, who tried to interpret many social problems.

Novelists of the 20th century

The *Madhalī Sthiti* (1885; "Middle State"), of Hari Narayan Apte, began the novel tradition in Marathi; the work's message was one of social reform. A high place is held by V.M. Joshi, who explored the education and evolution of a woman (*Suśīlā-cha Diva*, 1930) and the relation between art and morals (*Indu Kāḷe va Saralā Bhoḷe*, 1935). Important after 1925 were N.S. Phadke, who advocated art for art's sake, and V.S. Khandehar, who countered the former with an idealistic art for life's sake. Noteworthy contemporary novelists are S.N. Pendse, V.V. Shirwadkar, G.N. Dandekar, and Ranjit Desai.

Punjabi. Modern Punjabi literature began around 1860. A number of trends in modern poetry can be discerned. To the more traditional genres of narrative poetry, mystic verse, and love poems was added nationalist poetry in a humorous or satiric mood and experimental verse. Among the more important Punjabi poets are Bhai Vir Singh, in the 19th century, and Purana Singh, Amrita Pritam, and Baba Balwanta, in the 20th century.

Modern prose is represented by Bhai Vira Singha, Charana Singha, and Nanaka Singha, all of whom wrote novels; the same writers, as well as Gurbhaksh Singh and Devendra Satyarathi, also wrote short stories. Among playwrights mention may be made of I.C. Nanda, Harcharan Singh, and Santa Singh Sekhon.

Rajasthani. It is generally agreed that modern Rajasthani literature began with the works of Suryamal Misrama. His most important works are the *Vamsa Bhaskara* and the *Vira satsaī*. The *Vamsa Bhaskara* contains accounts of the Rājput princes who ruled in what was then Rājputāna (at present the state of Rājasthān), during the lifetime of the poet (1872–1952). The *Vira satsaī* is a collection of couplets dealing with historical heroes. Two other important poets in this traditional style are Bakhtavara Ji and Kaviraja Muraridana.

The period of nationalist strife against the British inspired a number of poets to verse that was both nationalist and in the traditional heroic vein; among them are Hiralala Sastri, Manikyalala Varma, and Jayanarayana Vyasa. This period was followed by one in which progressive social ideals inspired such poets as Ganeshilala Vyasa, Murlidhara Vyasa, and Satyaprakasha Jodhi.

Primarily known for their lyrics are Kanhaya Lal Sethiya and Megharaja Mukula, among others, and known for their narrative poems are Manohara Sharma, Shrimanta Kumara, and Naraina Singha Bhati.

Modern prose is represented in the novel, short story, and play. Among the novelists are Shiva Candra Bharatiya, Shri Lal Jodhi, Vijaya Dana Detha, and Yadavendra Sharma Chandra; the short-story writers are Rani Lakshmi Kumari Chandavata, Narasingh Rajapurohita, Dinadayala Ojha, and Purushottama Lala Menariya. Vijaya Dana Detha and Rani Lakshmi Kumari Chandavata are also known for their retelling of Rajasthani folktales. Among the playwrights is Shivachandra Bharatiya.

Tamil. In the second half of the 19th century two tendencies were present in Tamil literature. One was the

old traditional prose style of the *Patiṉeṇ-kīḻkkaṇakku*, or "Eighteen Ethical Works" (see above *Dravidian literature: 1st–19th century*), learned and severely scholastic; among others, V.V. Svaminatha Iyer and Arumuga Navalar wrote in this style. Another tendency, begun by Aruṇācala Kavirāyar in the 18th century, sought to bring the spoken and written languages together. This tendency developed on one side into such works as the operatic play *Nantaṉār Carittarak Kīrttaṉai* by Gopalakrishna, and on the other into ballads, often based on the lore of the Sanskrit *Purāṇas*. Despite attempts to effect a synthesis between the two languages, however, the scholastic style has continued to have a profound influence on modern Tamil literature; the normal spoken language, in fact, never became a literary medium.

First Tamil novel

The first novel in Tamil appeared in 1879, the *Piratāpamutaliyār Carittiram*, by Vetanayakam Pillai, who was inspired by English and French novels. In important respects Pillai's work is typical of all early modern Tamil fiction: his subject matter is Tamil life as he observed it, the language is scholastic, and the inspiration comes from foreign sources. Not strictly a novel, his work, which has a predominantly moral tone, is a loosely gathered string of narratives centred around an innocent hero.

Quite different is the *Kamalāmpāḷ Carittiram* ("The Fatal Rumor"), by Rajam Aiyar, whom many judge to be the most important prose writer of 19th-century Tamil literature. In this work, the author created a series of characters that appear to have become classics; the story is a romance, yet life in rural Tamil country is treated very realistically, with humour, irony, and social satire. In language Aiyar follows the classical style, which he intermixes with informal conversation, a style that has been imitated by modern authors.

The turn of the century saw the development of the *centamiḻ* style, which in many respects is a continuation of the medieval commentatorial style. The best representative is V.V. Swaminathan, who also is responsible for the rediscovery of the Tamil classical legacy, usually called "Tamil Renaissance," which tended to direct the mood of writers back to the glorious past. The pride in Tamil subsequently gave rise to a purist tradition and a second style, called *tuyattamiḻ*, or "pure Tamil." With exaggerated Tamilian self-consciousness, the language was purged of all non-Tamil loanwords, particularly Sanskrit, which removed the literary language even further from the spoken one. This style was not ineffective in verse but led easily to rhetoric.

The purist trend brought forth a reaction in *putumaṇipravāla naṭai*, "the new *maṇipravāla*" (see above *Dravidian literature: 1st–19th century*), which was Sanskritized with a vengeance and is of little literary interest.

The scholastic and formalist character of Tamil prose was predominant in the literature until the advent, in the early 20th century, of the poet and prose writer Subrahmanya Bharati. Bharati sought to synthesize the popular and the scholastic traditions of Tamil literature, and he created thereby a Tamil that was amenable to all literary expression. This synthesis, however, did not extend to the literary language itself, which in grammar continued the formal language, though for syntax, vocabulary, etc., he drew upon colloquial speech. In doing so he saved the language from the Sanskrit tradition of *Purāṇa* writing. His style is the *maṟumaḻarcci naṭai*, the "renaissance style."

In the first half of the 20th century, R. Krishnamurthy was an immensely popular writer. Under the pseudonym Kalki, he was an influential journalist who wrote voluminous historical romances.

In the 1930s there was a literary movement inspired by a journal called *Manikkoti*. Writers in this movement contributed extremely important new works, both in verse and prose, to Tamil letters. Among them was Putumaippittan, who wrote realistically, critically, and even bitterly about the failings of society.

Contemporary literature is represented by T. Janakiraman, who writes novels, short stories, and plays with themes from urban Tamil middle-class family life; Jayakanthan, a sharp and passionate writer, with a tendency to shock his readers; and L.S. Ramatirthan, probably the finest stylist at work in Tamil today, who started by writing in English.

Malayalam. In Malayalam the modern movement began in the late 19th century with Asan, who was temperamentally a pessimist—a disposition reinforced by his metaphysics—yet all his life was active in promoting his downtrodden Ezhava community. Ullor wrote in the classical tradition, on the basis of which he appealed for universal love, while Vallathol (died 1958) responded to the human significance of social progress.

Contemporary poetry records the encounter with problems of social, political, and economic life. The tendency is toward political radicalism.

Drama, native in Malayalam tradition, emerged in the modern period as farce, comedy, and satire but turned in the 1920s to a more sombre appraisal of outdated social conventions. The novel dates back to the late 1880s and was early concerned with social realism. At present the general tendency is introspective.

Emergence of modern Kannada poetry

Kannada. Modern Kannada poetry emerged about the beginning of the 20th century and showed a spirit of national purpose that pervaded other literature as well. By 1920, after major translations from Western models had been published, new literary forms such as the lyric and the short story came to the fore in the works of Panje Mangesh Rao and B.M. Srikantiah. Other prominent Kannada writers were D.V.G. Masti, Govinda Pai, and Kuvempe. In recent years a modernist movement has influenced the literature.

Urdu. The modern period in Urdu literature coincides with the mid-19th-century emergence of a middle class that saw in Western thought and science a means of needed social reform. Naẓīr Aḥmad wrote novels about the conflicts of Muslim middle class people. Shiblī, a poet and critic, wrote on the lives of great Muslims. The more famous novelists of the later period are Ratan-Nāth Sharshār, ʿAbd-ul-Ḥalīm Sharar, and Mīrzā Ruswā. The fathers of modern Urdu poetry were Ḥālī and Muḥammad Ḥusayn Āzād, the latter particularly characterized by a fine sensitivity for the past.

The greatest modern poet is Iqbāl. Writing in the early 20th century, he was influenced by the general sense of national purpose and the freedom movement. His poetic imagery, the power of his expression, and his philosophical outlook won the admiration of his fellow Muslims. In prose the most important writer of short stories was Prem Chand, who late in his career took to writing in Hindi. The 1930s saw the influence of progressivism, which attempted to make literature an arm of social revolution. Among the representative writers of this period are Sajjad Zahir, Upendranath Ashk, and Ismat Chughtai, the last a woman who is considered among the best.

English. There has been Indian literary activity in English for the last 200 years. It began with the insistence of the reformist Rammohan Ray (who was himself a forceful writer) and other like-minded Hindus that, for India to take its rightful place among nations, a knowledge of and education in English were essential. English literary activity took on a new aspect with the independence movement, whose leaders and followers found in English the one language that united them.

Among the first poets were Henry Derozio, Kashiprasad Ghose, and Michael Madhusudan Datta, all of whom wrote narrative verse. In the following generation there was Toru Dutt, important among women poets in this genre. Carrying on her work was Sarojini Naidu, judged by many the greatest of women poets; among her writings are *The Golden Threshold* (1905), *The Bird of Time* (1912), and *The Broken Wing* (1917). Best known of the Indian poets in English was the Bengali Rabindranath Tagore (see above *Bengali*), who, however, wrote most of his verse first in Bengali, and then translated it. A very different figure from Tagore is Sri Aurobindo, who started out as an ardent nationalist and was jailed by the British. After his conversion from activism to introspection, which took place in jail, he established a hermitage in Pondicherry. He left behind a rich *oeuvre* of verse that has inspired a contemporary school of mystic poets.

Other modern poets show the influence of T.S. Eliot and Ezra Pound.

The independence movement gave strong impetus to expository prose. Important contributors to this genre were Bal Gangadhar Tilak, who edited the English journal *Mahratta*, Lala Lajpat Rai, Kasturi Ranga Iyengar, and T. Prakasam. Mahatma Gandhi, too, wrote widely in English and edited *Young India* and the *Harijan*. He also wrote the autobiography *My Experiments with Truth* (originally published in Gujarati, 1927–29), now an Indian classic. In this he was followed by Jawaharlal Nehru, whose *Discovery of India* is justly popular.

Prose fiction

Prose fiction in English began in 1902 with the novel *The Lake of Palms*, by Romesh Chunder Dutt. The next important novelist is Mulk Raj Anand, who fulminated against class and caste distinction in a series of novels, *The Coolie* (1936), *Untouchable* (1935), *Two Leaves and a Bud* (1937), and *The Big Heart* (1945). Less fierce, though a better craftsman, is R.K. Narayan, who has published nine novels (as well as many short stories), among them *The Guide* (1958), *The Man-Eater of Malgudi* (1961), and *The Vendor of Sweets* (1967); his work has a wider circle of readers outside India than within. Other Indian novelists in the English medium are Santha Rama Rau, Manohar Malgonkar, Kamala Markandeya, and Khushwant Singh. The most popular is Raja Rao, whose novels *Kanthapura* (1938), *The Cow of the Barricades* (1947), and *The Serpent and the Rope* have attracted a wide following.

Sinhalese. Traditional contemporary poetry continues to be Buddhist in subject matter and sentiment. A more modern literature arose under the influence of Western models; notable among the contemporary representatives of Sinhalese literature are Kumaranatunga, a critic, Matin Wickremasinghe, a novelist, and Tennakoon, a poet.

(J.A.B.v.B.)

II. Music

FOLK, CLASSICAL, AND POPULAR MUSIC

Rural areas. The wide field of musical phenomena in South Asia ranges from the relatively simple two- or three-tone melodies of some of the hill tribes in central India to the highly refined art music heard in concert halls in the large cities. This variety reflects the heterogeneous population of the subcontinent in terms of race, religion, language, and social status. In the villages, music is not just a form of entertainment but is an essential element in many of the activities of daily life and plays a prominent part in most of the rituals. These include life-cycle events, such as birth, initiation, marriage, and death; events of the agricultural cycle, such as planting, transplanting, harvesting, and threshing; and a variety of work songs. Much of this music could be described as functional, for it serves a utilitarian purpose; for instance, a harvest song might well give thanks to God for a bountiful harvest, but underlying this is the idea that singing this song in its traditional manner will help to ensure that the next harvest will be equally fruitful. These songs are usually sung by all of the members participating in the activity and are not sung for a human audience. They are often sung in the form of leader and chorus, and the musical accompaniment, if any, is generally provided by drone instruments (those sustaining or reiterating a given note or notes), usually of the lute family, or percussion instruments, such as drums, clappers, and pairs of cymbals. Occasionally, a fiddle or flute might also accompany the singers, who often dance while they sing.

In each area and even within a single area, different social groups have their own individual songs whose origins are lost in antiquity. The songs are passed on from one generation to another, and in most cases the composers are unknown. Apart from folk songs, one also hears outdoor instrumental music in villages. The music is provided by an ensemble of varying size, which consists basically of an oboe type of instrument (usually called *shahnāʾī* in North India and *nāgasvaram* in the south) and a variety of drums. Sometimes straight, curved, or S-shaped horns may be added. These groups play at weddings, funerals, and religious processions. The musicians are professional or semiprofessional and usually belong to a very low caste. Such ensembles are found in tribal as well as folk societies and in villages as well as in cities.

Itinerant musicians

Other professional music is also found in the rural regions. Most areas are visited by religious mendicants, many of whom travel around the countryside singing devotional songs, accompanying themselves either with a one-, two-, or three-stringed lute that generally provides only a drone or with a frame (tamborine-like) drum. They carry with them a small begging bowl and maintain themselves entirely on what they receive in alms. There are also itinerant magicians, snake charmers, acrobats, and storytellers who travel in the rural areas, often providing the only entertainment available in the villages. Music is often involved in their acts, and the storyteller generally sings his tales, which may be taken from the two epics, the *Mahābhārata* and the *Rāmāyaṇa*, or from the *Purāṇas*, the legends that describe the adventures of the incarnations of God as they rid the world of evil. Sometimes the narrative songs are concerned with historical characters and describe the wars and the heroic deeds of the regional rulers. Some storytellers specialize in generally tragic stories of romance and of lovers.

During certain religious festivals, the villages might be visited by a travelling band of players who enact some of the mythological episodes connected with the festival. Such performances are accompanied by music and may also include dances. During the festivals villagers may visit neighbourhood shrines or temples, there encountering religious mendicants singing devotional songs and perhaps watching elaborate enactments of the episodes connected with the festival. Thus, the villagers become familiar with the mythological and philosophical aspects of their religion, in spite of the low level of literacy in the rural areas and the difficulties of communication, often limited to a narrow dirt road traversed by bullock carts.

In modern times, rural areas are being influenced to a greater extent by urban culture. The principal impact has come through the introduction of relatively inexpensive transistorized radios, which have found their way into fairly remote villages. In addition, travelling cinemas, set up quickly and easily in tents, have visited the rural areas for some years. As a result, the traditional rural forms of music and dance are in the process of change.

Classical music. In the cities many different forms of music can be heard. Of these the best known in the West are the classical music of North India, including Pakistan, sometimes called Hindustani music, and that of South India, or Karnatic music. Both classical systems are supported by an extensive body of literature and elaborate musical theory. Until modern times, classical music was patronized by the princely courts and to some extent also by the wealthy noblemen. Since India gained independence in 1947, and with the abolition of the princely kingdoms, the emphasis has shifted to the milieu of large concert halls. The concertgoer, radio, and the cinema are now the main patrons of the classical musicians. In recent times the growth of university music programs, particularly involving classical music, has placed greater emphasis on music history and theory and has provided a further source of income for musicologists and musicians. The traditional system of private instruction, however, still continues to this day.

Rāga and *tāla*

Classical music is based on two main elements, *rāga* and *tāla*. The word *rāga* is derived from a Sanskrit root meaning "to colour," the underlying idea being that certain melodic shapes, involving specific intervals of the scale, produce a continuity of emotional experience and "colour" the mind. Since neither the melodic shapes nor their sequence are fixed precisely, a *rāga* serves as a basis for composition and improvisation. Indian music has neither modulation (change of key) nor changing harmonies; instead, the music is invariably accompanied by a drone that establishes the tonic, or ground note, of the *rāga* and usually its fifth (*i.e.*, five notes above). These are chosen to suit the convenience of the main performer, as there is no concept of fixed pitch. While a *rāga* is primarily a musical concept, specific *rāga*s have acquired, particularly in North Indian music, a number of extramusical ele-

ments and are associated with particular periods of the day, seasons of the year, colours, deities, and specific moods.

The second element of Indian music, *tāla*, is best described as time measure and has two main constituents; the duration of the time measure in terms of time units that vary according to the tempo chosen; and the distribution of stress within the time measure. *Tāla*, like *rāga*, serves as a basis for composition and improvisation.

Indian classical music is generally performed by small ensembles of not more than five or six musicians. Improvisation plays a major part in a performance, and great emphasis is placed on the creativity and sensitivity of the soloist. A performance of a *rāga* usually goes through well-defined stages, beginning with an improvised melodic prelude that is followed by a composed piece set in a particular time measure. The composition is generally quite short and serves as a frame of reference to which the soloist returns at the conclusion of his improvisation. There is no set duration for the performance of a *rāga*. A characteristic feature of North Indian classical music is the gradual acceleration of tempo, which leads to a final climax.

Nonclassical music of the cities. Classical music interests only a small proportion of the peoples of South Asia, even in the cities. Since about the 1930s a new genre, associated with the cinema, has achieved extraordinary popularity. Most Indian films are very much like Western musicals and generally include six or more songs.

Film music

Film music derives its inspiration from a number of sources, both Indian and Western; classical, folk, and devotional music are the main Indian sources, while Western influence is seen most obviously in the use of large orchestras that employ both Western and Indian instruments. The influence of Western popular music, too, is very evident. In spite of the eclectic nature of Indian film music, most of the songs maintain an Indian feeling that arises largely from the vocal technique of the singers and the ornamentation of the melody line. This music is an experimental and developing form, and there have been attempts to add harmony and counterpoint, some of which may seem rather naïve to the Western ear. But the film music differs from typical Western music in that the melody line is generally not dictated by harmonic progressions and in that the harmonies used are incidental additions.

Aside from classical and film music, there are several other forms of urban music, some of which closely resemble the music of the rural areas. In city streets one is likely to encounter an outdoor band of oboes and drums announcing a wedding or a funeral. Street musicians, religious mendicants, snake charmers, storytellers, and magicians perform at every available opportunity, and work songs are sung by construction workers and other labourers. In private homes still other forms of music are performed, ranging from religious chanting to traditional folk and devotional songs. In public places of entertainment, the listener may encounter, apart from classical and film music, theatrical music from one of the relatively modern forms of regional theatre; and in the low-brow places of entertainment courtesans still sing and dance in traditional fashion. In the larger cities there are performances of Western chamber music and occasionally symphony concerts, as well as popular dance music, rock, and jazz in the night clubs.

ANTIQUITY

In a musical tradition in which improvisation predominates, and written notation, when used, is skeletal and more a tool of the theorist than of the practicing musician, the music of past generations is irrevocably lost. References to music in ancient texts, aesthetic formulations, and depictions and written discussions of musical instruments can offer clues. In rare instances an ancient musical style may be preserved in unbroken oral tradition. For most historical eras and styles, surviving treatises explaining musical scales and modes—the framework of melody—provide a particularly important means of recapturing at least a suggestion of the music of former times, and tracing the musical theory of the past makes clear the position of the present musical system.

Little is known of the musical culture of the Indus Valley civilization of the 3rd and 2nd millennia BC. Some musical instruments, such as the arched, or bow-shaped, harp and more than one variety of drum, have been identified from the small terra-cotta figures and among the pictographs on the seals that were probably used by merchants. Further, it has been suggested that a bronze statuette of a dancing girl represents a class of temple dancers similar to those found much later in Hindu culture. It is known that the Indus civilization had established trade connections with the Mesopotamian civilizations, so that it is possible that the bow harp found in Sumeria would also have been known in the Indus Valley.

Vedic chant. *Compilation of hymns.* It is generally thought among scholars that the Indus Valley civilization was terminated by the arrival of bands of semi-nomadic tribesmen, the Aryans, who descended into India from the northwest, probably in the first half of the 2nd millennium BC. An important aspect of Aryan religious life was the bard-priest who composed hymns in praise of gods, to be sung or chanted at sacrifices. This tradition was continued in the invaders' new home in northern India until a sizable body of oral religious poetry had been composed. By about 1000 BC this body of chanted poetry had apparently grown to unmanageable proportions, and the best of the poems were formed into an anthology called Ṛgveda, which was then canonized. It was not committed to writing, but text and chanting formula were carefully handed down by word of mouth from one generation to the next, up to the present period. The poems in the Ṛgveda are arranged according to the priestly families who used and, presumably, had composed the hymns. Shortly after this a new Veda, called the Yajurveda, basically a methodical rearrangement of the verses of the Ṛgveda with certain additions in prose, was created to serve as a kind of manual for the priest officiating at the sacrifices. At approximately the same time, a third Veda, the Sāmaveda, was created for liturgical purposes. The Sāmaveda was also derived from the hymns of the Ṛgveda, but the words were distorted by the repetition of syllables, pauses, prolongations, and phonetic changes, as well as the insertion of certain meaningless syllables believed to have magical significance. A fourth Veda, the Atharvaveda, was accepted as a Veda considerably later and is quite unrelated to the other three. It represents the more popular aspects of the Aryan religion and consists mostly of magic spells and incantations.

Each of these Vedas has several ancillary texts, called the *Brāhmaṇas*, *Āraṇyakas*, and *Upaniṣads*, which are also regarded as part of the Vedas. These ancillary texts are concerned primarily with mystical speculations, symbolism, and the cosmological significance of the sacrifice. The Vedic literature was oral and not written down until very much later, the first reference to a written Vedic text being in the 10th century AD. In order to ensure the purity of the Vedas, the slightest change was forbidden, and the priests devised systems of checks and counterchecks, so that there has been virtually no change in these texts for about 3,000 years. Underlying this was the belief that the correct recitation of the Vedas was "the pivot of the universe" and that the slightest mistake would have disastrous cosmic consequence unless expiated by sacrifice and prayer. The Vedas are still chanted by the Brahmin priests at weddings, initiations, funerals, and the like, in the daily devotions of the priests, and at the now rarely held so-called public sacrifices.

Ritual accuracy of recitation

From the Vedic literature it is apparent that music played an important part in the lives of the Aryan peoples, and there are references to stringed instruments, wind instruments, and several types of drums and cymbals. Songs, instrumental music, and dance are mentioned as being an integral part of some of the sacrificial ceremonies. The bow harp (*vīṇā*), a stringed instrument (probably a board zither) with 100 strings, and the bamboo flute were the most prominent melody instruments. Little is known of the music, however, apart from the Vedic chanting, which can still be heard today.

Chant intonation. The chanting of the Ṛgveda and Yajurveda shows, with some exceptions, a direct correlation with the grammar of the Vedic language. As in ancient Greek, the original Vedic language was accented, with the location of the accent often having a bearing on the meaning of the word. In the development of the Vedic language to Classical Sanskrit, the original accent was replaced by an automatic stress accent, whose location was determined by the length of the word and had no bearing on its meaning. It was thus imperative that the location of the original accent be inviolate if the Vedic texts were to be preserved accurately. The original Vedic accent occurs as a three-syllable pattern: the central syllable, called *udātta,* receives the main accent; the preceding syllable, *anudātta,* is a kind of preparation for the accent; and the following syllable, *svarita,* is a kind of return from accentuation to accentlessness. There is some difference of opinion among scholars as to the nature of the original Vedic accent; some have suggested that it was based on pitch, others on stress; and one theory proposes that it referred to the relative height of the tongue.

Verbal accent and musical pitch

In the most common style of Ṛgvedic and Yajurvedic chanting found today, that of the Tamil Aiyar Brahmins, it is clear that the accent is differentiated in terms of pitch. This chanting is based on three tones; the *udātta* and the nonaccented syllables (called *pracaya*) are recited at a middle tone, the preceding *anudātta* syllable at a low tone, and the following *svarita* syllable either at the high tone (when the syllable is short) or as a combination of middle tone and high tone. The intonation of these tones is not precise, but the lower interval is very often about a whole tone, while the upper interval tends to be slightly smaller than a whole tone but slightly larger than a semitone. In this style of chanting the duration of the tones is also relative to the length of the syllables, the short syllables generally being half the duration of the long.

The more musical chanting of the Sāmaveda employs five, six, or seven tones and is said to be the source of the later secular and classical music. From some of the phonetic texts that follow the Vedic literature, it is apparent that certain elements of musical theory were known in Vedic circles, and there are references to three octave registers (*sthāna*), each containing seven notes (*yama*). An auxiliary text of the Sāmaveda, the *Nāradīśikṣā,* correlates the Vedic tones with the accents described above, suggesting that the Samavedic tones possibly derived from the accents. The Samavedic hymns as chanted by the Tamil Aiyar Brahmins are based on a mode similar to the D mode (D–d on the white notes of the piano; *i.e.,* the ecclesiastical Dorian mode). But the hymns seem to use three different-sized intervals, in contrast to the two sizes found in the Western church modes. They are approximately a whole tone, a semitone, and an intermediate tone. Once again, the intervals are not consistent and vary both from one chanter to another and within the framework of a single chant. The chants are entirely unaccompanied by instruments, and this may account for some of the extreme variation of intonation.

The changes brought by the 20th century have weakened the traditional prominent position of the Vedic chant. The Atharvaveda is seldom heard in India now. Samavedic chant, associated primarily with the large public sacrifices, also appears to be dying out. Even the Ṛgveda and Yajurveda are virtually extinct in some places, and South India is now the main stronghold of Vedic chant.

The classical period. The ritual of the Vedas involves only the three upper classes, or castes, of Aryan society: the Brahmin, or priestly class; the Kṣatriya, or prince-warriors; and the Vaiśya, or merchants. The fourth caste, the Śūdras, or labourers, were excluded from Vedic rites. The primary sources of religious education and inspiration for the Śūdra were derived from what is sometimes called the fifth Veda: the epic poems *Rāmāyaṇa* and *Mahābhārata,* as well as the collections of legends, called the *Purāṇas,* depicting the lives of the various incarnations of the Hindu deities. The *Rāmāyaṇa* and the *Mahābhārata* were originally secular in character, describing the heroic deeds of kings and noblemen, many of whom are not recorded in history. Subsequently, religious matter was added, including the very famous sermon *Bhagavadgītā* ("Song of the Lord"), which has been referred to as the most important document of Hinduism; and many of the heroes of the epics were identified as incarnations of the Hindu deities. The legends were probably sung and recited by wandering minstrels and bards even before the advent of the Christian Era, in much the same way as they still are. The stories were also enacted on the stage, particularly at the time of the religious festivals. The earliest extant account of drama is to be found in the *Nāṭya-śāstra* ("Treatise on the Dramatic Arts"), a text that has been dated variously from the 2nd century BC to the 5th century AD and even later. It is virtually a handbook for the producer of stage plays and deals with all aspects of drama, including dance and music.

Music and drama

Theatrical music of the period apparently included songs sung on stage by the actors, as well as background music provided by an orchestra (which included singers) located offstage, in what was very like an orchestra pit. Melodies were composed on a system of modes, or *jāti*s, each of which was thought to evoke one or more particular sentiments (*rasa*) by its emphasis on specific notes. The modes were derived in turn from the 14 *mūrchanā*s —seven pairs of ascending seven-note series beginning on each of the notes of two closely related heptatonic (seven-note) parent scales, called *ṣaḍjagrāma* and *madhyamagrāma.* The *mūrchanā*s were thus more or less analogous to the European modal scales that begin progressively on D, E, F, G, etc. A third parent scale, *gāndhāragrāma,* was mentioned in several texts of the period and some even earlier but is not included in the system laid out in the *Nāṭya-śāstra.*

Qualities of the scales. The two parent scales differed in the positioning of just one note, which was microtonally flatter in one of the scales. The microtonal difference, referred to as *pramāṇa* ("measuring") *śruti,* presumably served as a standard of measurement. In terms of this standard it was determined that the intervals of the *mūrchanā*s were of three different sizes, consisting of two, three, or four *śruti*s, and that the octave comprised 22 *śruti*s. An interval of one *śruti* was not used. Several modern scholars have suggested that the *śruti*s were of unequal size; from the evidence in the *Nāṭya-śāstra,* it would appear, however, that they were thought to be equal. There has been no attempt to determine the exact size of the *śruti*s in any of the traditional Indian musical treatises until relatively modern times (18th century). The term *śruti* was also used to define consonance and dissonance, as these terms were understood in the period. In this connection, four terms are mentioned: *vādī,* comparable to the Western term sonant, meaning "having sound"; *saṃvādī,* to the Western consonant (concordant; reposeful); *vivādī,* to dissonant (discordant; lacking repose); and *anuvādī,* to assonant (neither consonant nor dissonant). As in the ancient Greek Pythagorean system, which influenced Western music, only fourths and fifths (intervals of four or five tones in a Western scale) were considered consonant. In the Indian system of measurement, tones separated by either nine or 13 *śruti*s correspond in size to Western fourths and fifths and are described as being consonant to each other. "Dissonant" in this system referred only to the minor second, an interval of two *śruti*s, and to its inversion (complementary interval), the major seventh (20 *śruti*s). All other tones, including the major third, were thought to be assonant.

Melodically important microtonal differences

The musical difference between the two parent scales is best seen not in terms of the microtonal deviation mentioned earlier but rather in terms of a musically influential consonance found in one but lacking in the other and vice versa. In each of the parent scales there are two nonconsonances, one of which is the tritone (interval of three Western whole tones, such as F–B) of 11 *śruti*s inevitable in all diatonic scales (seven-note scales of the major scale and *mūrchanā* type) and which in medieval Europe was described as *diabolus in musica* ("the devil in music").

The second is a microtonal nonconsonance unique to

Table 1: Intervals of Ṣaḍjagrāma and Madhyamagrāma Parent Scales*

Ṣaḍjagrāma

				10 *śrutis*								
Indian notes:	(*ni*)		*ṣa*	*ṛi*	*ga*		*ma*	*pa*	*dha*	*ni*		(*ṣa*)
Śruti intervals:		4	3	2	4		4	3	2		(4)	
Comparable Western notes†:	(C)		D	E⁻	F		G	A	B⁻	c		(d)
					11 *śrutis*							

Madhyamagrāma

				12 *śrutis*								
Indian notes:	(*ni*)		*ṣa*	*ṛi*	*ga*		*ma*	*pa*	*dha*	*ni*		(*ṣa*)
Śruti intervals:		4	3	2	4		3	4	2		(4)	
Comparable Western notes†:	(C)		D	E	F		G	A⁻	B⁻	c		(d)
					11 *śrutis*							

*Minus signs indicate slightly lower pitch. †Without reference to precise pitch.

this ancient Indian system. It can be illustrated by referring in the subsequent explanation to Table 1, in which the seven Indian notes *ṣaḍja, ṛṣabha, gāndhāra, madhyama, pañcama, dhaivata,* and *niṣāda* are given in their commonly abbreviated forms, *ṣa, ṛi, ga, ma, pa, dha,* and *ni.*

The nonconsonance arises from variances of one *śruti* from the fundamental consonances of the fourth and the fifth—a variance of about a quarter tone. In the *ṣaḍjagrāma* scale the interval *ṛi–pa* (E⁻ to A) contains 10 *śrutis*; *i.e.*, one more than the nine of the consonant fourth. Comparably, in the *madhyamagrāma* scale the interval *ṣa–pa* (D to A⁻) contains 12 *śrutis*, or one fewer than the consonant fifth. These variances involve the consonant relationships of two melodically prominent notes, the first and the fifth. In the *madhyamagrāma* the first note, *ṣa,* has no consonant fifth, and perhaps for this reason this scale is said to begin not on the *ṣa* (D) but on its fourth, the note *ma* (G); hence, it resembles the G mode—*i.e.*, the ecclesiastical Mixolydian mode—whereas the *ṣaḍjagrāma* resembles the D mode, the ecclesiastical Dorian.

There is a striking resemblance of the *ṣaḍjagrāma* scale to the intervals used by the Tamil Aiyar Brahmins in their chanting of the Sāmaveda. Not only are their hymns set in a mode similar to the D mode, but they seem to use three different-sized intervals, the intermediate one corresponding to the three-*śruti* interval. The *Nāṭya-śāstra* claims to have derived song (*gīta*) from the chanting of the Sāmaveda, and the resemblances between the two may not be entirely fortuitous.

The two parent scales are complementary and between them supply all the consonances found in the ancient Greek Pythagorean scale. Thus, if in a mode the consonance *ṛi–pa* (E–A) were needed, one would tune to the *madhyamagrāma* scale. But, if the consonance *ṣa–pa* (D–A) were important, it could be obtained with the *ṣaḍjagrāma* tuning. There was a further development in this system caused by the introduction of two additional notes, called *antara ga* (F♯) and *kākalī ni* (C♯), which could be substituted for the usual *ga* (F) and *ni* (C). The *antara ga* eliminates the 11-*śruti* tritone between *ga* and *dha* (F–B), but its use creates a further tritone between F♯ and C. The second additional note, *kākalī ni* (C♯), eliminates this tritone but once again creates a new one, this time between C♯ and G. This process of adding notes, if carried further, would eventually lead to the circle, or, rather, the spiral, of fourths or fifths found in Western music (whereby a sequence of fifths, such as C–G, G–D, D–A, etc., leads eventually back to a microtonally out-of-tune C); there is no evidence that such a circle or spiral was known in ancient India.

Introduction of new notes

Mode, or jāti. From each of the two parent scales were derived seven modal sequences (the *mūrchanās* described above), based on each of the seven notes. The two *mūrchanās* of a corresponding pair differed from each other only in the tuning of the note *pa* (A), the crucial distinction in the tunings of the two parent scales. One of each pair was selected as the basis for a "pure" mode, or *śuddha-jāti*; in the groups of seven pure modes, four used the tuning of the *ṣaḍjagrāma* and three that of the *madhyamagrāma.* In addition to these seven pure modes, a further 11 "mixed" modes, or *vikṛta-jātis*, are also mentioned in the *Nāṭya-śāstra.* These were derived by a combination of two or more pure modes, but the text does not explain just how this was done.

The *jātis* were similar to the modern concept of *rāga* in that they provided the melodic basis for composition and presumably, improvisation. They were not merely scales, but were also assigned 10 melodic characteristics: *graha,* the initial note; *aṃśa,* the predominant note; *tāra,* the note that forms the upper limit; *mandra,* the note that forms the lower limit; *nyāsa,* the final note; *apanyāsa,* the secondary final note; *alpatva,* the notes to be used infrequently; *bahutva,* the notes to be used frequently; *ṣāḍavita,* the note that must be omitted in order to create the hexatonic (six-note) version of the mode; and *auḍavita,* the two notes that must be omitted to create the pentatonic (five-note) version of the mode.

No written music survives from this early period. It is not clear from the description whether or not the music was like that of the present period. There is no mention of a drone, nor do the instruments of the orchestra—consisting of the *vipañcī* and *vīṇā* (bow harps?), bamboo flute, a variety of drums, and singers—appear to include any specifically drone instrument, such as the modern *tambūrā.* The evidence tends rather to suggest, from the emphasis on consonance and some of the playing techniques, that some form of organum (two or more parts paralleling the same melody at distinct pitch levels) and even some type of rudimentary harmony may have been characteristic.

MEDIEVAL PERIOD

Precursors of the medieval system. It is not clear just when the *jāti* system fell into disuse, for later writers refer to *jātis* merely out of reverence for Bharata, the author of the *Nāṭya-śāstra.* Later developments are based on musical entities called *grāma-rāgas,* of which seven are mentioned in the 7th-century Kuṭimiyāmalai rock inscription in Tamil Nadu state. Although the word *grāma-rāga* does not occur in the *Nāṭya-śāstra,* the names applied to the individual *grāma-rāgas* are all mentioned. Two of them, *ṣaḍjagrāma-rāga* and *madhyamagrāma-rāga,* are obviously related to the parent scales of the *jāti* system. The other five seem to be variants of these two *grāma-rāgas* in which either or both the altered forms of the notes *ga* and *ni* (F♯ and C♯) are used. In the *Nāṭya-śāstra,* the reference to the various *grāma-rāgas* is far removed from the main section in which the *jāti* system is discussed, and there is no obvious connection between the two. Each of the *grāma-rāgas* is said to be used in one of the seven formal stages of Sanskrit drama. They have been reconstructed as shown in Table 2.

Further development of the grāma-rāgas. In the next significant text on Indian music, the *Bṛhaddeśī,* written by the theorist Mātaṅga about the 10th century AD, the *grāma-rāgas* are said to derive from the *jātis.* In some respects, at least, the *grāma-rāgas* resemble not the *jātis* but their parent scales. The author of the *Bṛhaddeśī* claims to be the first to discuss the term *rāga* in any detail. It is clear that *rāga* was only one of several kinds of musical entities in this period and is described as having "varied and graceful ornaments, with emphasis on clear, even and

Character and possible origin of the *rāga*

Table 2: Grāma-Rāgas

	scale											
Madhyama-grāma-rāga												
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*	*ṣa*
Śruti values	3		2		4		3	4		2		4
Comparable Western notes	D	E⁻		F		G	A⁻		B⁻		c	d
Ṣaḍja-grāma-rāga												
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*	*ṣa*
Śruti values	3		2		4		4	3		2		4
Comparable Western notes	D	E⁻		F		G	A		B⁻		c	d
Sādava-grāma-rāga				(a)*								
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*	*ṣa*
Śruti values	3		4		2		3	4		2		4
Comparable Western notes	D	E⁻		F♯		G	A⁻		B⁻		c	d
Pañcama-grāma-rāga				(a)*								
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*	*ṣa*
Śruti values	3		4		2		4	3		2		4
Comparable Western notes	D	E⁻		F♯		G	A		B⁻		c	d
Kaiśika-grāma-rāga				(a)*							(k)*	
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*	*ṣa*
Śruti values	3		4		2		3	4		4		2
Comparable Western notes	D	E⁻		F♯		G	A⁻		B⁻		c♯	d
Sādhārita-grāma-rāga				(a)*							(k)*	
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*	*pa*		*dha*		*ni*	*ṣa*
Śruti values	3		4		2		4	3		4		2
Comparable Western notes	D	E⁻		F♯		G	A		B⁻		c♯	d
Kaiśika-madhyama-grāma-rāga				(a)*							(k)*	
Indian notes	*ṣa*	*ṛi*		*ga*		*ma*			*dha*		*ni*	*ṣa*
Śruti values	3		4		2		7			4		2
Comparable Western notes	D	E⁻		F♯		G			B⁻		c♯	d

*(a) and (k) refer to *antara* and *kakali*, the variant forms of the notes *ga* and *ni*.

deep tones and having a charming elegance." The *rāga*s of this period seem to have been named after the different peoples living in the various parts of the country, suggesting that their origin might lie in folk music. Mātaṅga appears to contrast the two terms *mārga* and *deśī*. *Mārga* (literally "the path") apparently refers to the ancient traditional musical material, whereas *deśī* (literally "the vulgar dialect spoken in the provinces") designates the musical practice evolving in the provinces, which may have had a more secular basis. Although the title *Bṛhaddeśī* ("The Great Deśī") suggests that the latter music might have been the focus of the treatise and that the *grāma-rāga*s were possibly out of date by that time, the surviving portion of the text does not support such a theory.

The mammoth 13th-century text *Saṅgītaratnākara* ("Ocean of Music and Dance"), composed by the theorist Śārṅgadeva, is often said to be one of the most important landmarks in Indian music history. It was composed in the Deccan (south central India) shortly before the conquest of this region by the Muslim invaders and thus gives an account of Indian music before the full impact of Muslim influence. A large part of this work is devoted to *mārga*—that is, the ancient music that includes the system of *jāti*s and *grāma-rāga*s—but Śārṅgadeva mentions a total of 264 *rāga*s. Despite the use in both the *Bṛhaddeśī* and the *Saṅgītaratnākara* of a notation equivalent to the Western tonic sol–fa (*i.e.*, with syllables, as do–re–mi . . .) to illustrate the *rāga*s, modern scholars have not yet been able to reconstruct them with assurance.

The basic difficulty scholars face lies in determining the intervals used in each of the *rāga*s. In the ancient system, the *jāti*s were something like the ancient Greek and medieval church modes in that each was derived from a parent scale by altering the ground note and the tessitura (range). In modern Indian music, however, the *rāga*s are all transposed to a common ground note. This change may well be connected with the introduction of the drone and the evolution of the long-necked-lute family on which the drone is usually played. In the old system, with the changing ground note, it would have been necessary to retune drone instruments from one *rāga* to another, a cumbersome and impractical operation to carry out during a recital. It may have been this factor that provided the impetus for the change to the standard-ground-note system. There is no conclusive evidence to show just when this change might have taken place, and it is not clear whether the *Bṛhaddeśī* and the *Saṅgītaratnākara* are using the old system or one similar to that used in modern times.

THE ISLAMIC PERIOD

Impact on musical genres and aesthetics. The Muslim conquest of India can be said to begin in the 12th century, although Sind (now in Pakistan) had been conquered by the Arabs as early as the 8th century. Muslim writers such as al-Jāḥiẓ and al-Masʿūdī had already commented favourably on Indian music in the 9th and 10th centuries, and the Muslims in India seem to have been very much attracted by it.

In the beginning of the 14th century the great poet Amīr Khosrow, who was considered to be extremely proficient in both Persian and Indian music, wrote that Indian music was superior to the music of any other country. Further, it is stated that, after the Muslim conquest of the Deccan under Malik Kāfūr (*c.* 1310), a large number of Hindu musicians were taken with the royal armies and settled in the North. Although orthodox Islām considered music illegal, the acceptance of the Ṣūfī doctrines, in which music was an accepted means to the realization of God, enabled Muslim rulers and noblemen to extend their patronage to this art. At the courts of the Mughal emperors Akbar, Jahāngīr, and Shāh Jahān, music flourished on a grand scale. Apart from Indian musicians, there were also musicians from Persia, Afghanistan, and Kashmir in the employ of these rulers; nevertheless, it appears that it was Indian music that was most favoured. Famous Indian musicians, such as Svāmī Haridās and Tānsēn, are legendary performers and innovators of this period. After the example set by Amīr Khosrow, Muslim musicians took an active interest in the performance of Indian music and added to the repertoire by inventing new *rāga*s, *tāla*s, and musical forms, as well as new instruments.

The Muslim patronage of music was largely effective in the north of India and has had a profound influence on North Indian music. Perhaps the main result of this influence was to de-emphasize the importance of the words of the songs, which were mostly based on Hindu devotional themes. In addition, the songs had been generally composed in Sanskrit, a language that had ceased to be a medium of communication except among scholars and priests. Sanskrit songs were gradually replaced by compositions in the various dialects of Hindi, Braj Bhasa, Bhojpuri, and Dakhani, as well as in Urdu and Persian; nevertheless, the problems of communication, in terms of both language and subject matter, were not easily recon-

ciled. A new approach to religion was, in any case, sweeping through India at about this time. This emphasized devotion (*bhakti*) as a primary means to achieving union with God, bypassing the traditional Hindu beliefs of the transmigration of the soul from body to body in the lengthy process of purification before it could achieve the Godhead. The Islāmic Ṣūfī movement was based on an approach similar to that of the *bhakti* movements and also gained many converts in India. A manifestation of these devotional cults was the growth of a new form of mystic–devotional poetry composed by wandering mendicants who had dedicated their lives to the realization of God. Many of these mendicants have been sanctified and are referred to as poet-saints or singer-saints, since their poems were invariably set to music. A number of devotional sects sprang up all over the country, some Muslim, some Hindu, and others merging elements from both. These sects emphasized the individual's personal relationship with God. In their poetry, man's love for God was often represented as a woman's love for man and, specifically, the love of the milkmaid Rādhā for Kṛṣṇa, a popular incarnation of the Hindu god Viṣṇu. In the environment of the royal courts, there was a less idealistic interpretation of the word love, and much of the poetry, as well as the miniature painting, of the period depicts the states of experience of the lover and the beloved.

Influence of religious movements

This attitude is also reflected in the musical literature of the period. From early times, both *jāti*s and *rāga*s in their connection with dramatic performance were described as evoking specific sentiments (*rasa*) and being suitable for accompanying particular dramatic events. It was this connotational aspect, rather than the technical one, that gained precedence in this period. The most popular method of classification was in terms of *rāga*s (masculine) and their wives, called *rāgiṇī*s, which was extended to include *putra*s, their sons, and *bhāryā*s, the wives of the sons. The *rāga*s were personified and associated with particular scenes, some of which were taken from Hindu mythology, while others represented aspects of the relationship between two lovers. The climax of this personification is found in the *rāgamāla* paintings, usually in a series of 36, which depict the *rāga*s and *rāgiṇī*s in their emotive settings.

Theoretical developments. From the middle of the 16th century, a new method of describing *rāga*s is found in musical literature. It was also at about this time that the distinction between North and South Indian music became clearly evident. In the literature, *rāga*s are described in terms of scales having a common ground note. These scales were called *mela* in the South and *mela* or *thāṭa* in the North.

Emergence of modern theoretical concepts

It was in the South that a complete theoretical system of *mela*s was introduced, in the *Caturdaṇḍi-prakāśika* ("The Illuminator of the Four Pillars of Music"), a text written in the middle of the 17th century. This system was based on the permutations of the tones and semitones, which had by this time been reduced to a basic 12 in the octave. The octave was divided into two tetrachords, or four-note sequences, C–F and G–c, and six possible tetrachord species were arranged in an order showing their relationship with each other. It will be noted in the sequence of tetrachords shown below that each lower tetrachord has an analogous upper tetrachord and that the outer notes of each are constant, whereas the inner notes change their pitch.

1.	C	D♭	E𝄫	F	and	G	A♭	B𝄫	c
2.	C	D♭	E♭	F		G	A♭	B♭	c
3.	C	D♭	E	F		G	A♭	B	c
4.	C	D	E♭	F		G	A	B♭	c
5.	C	D	E	F		G	A	B	c
6.	C	D♯	E	F		G	A♯	B	c

The list could have extended further, except that apparently no pitch distinction was made between the enharmonic pairs D–E𝄫, D♯–E♭, A–B𝄫, and A♯–B♭. (Enharmonic notes have different pitch names but sound either the same pitch or, in some tuning systems, have very slight differences in pitch.)

By utilizing all possible combinations of a lower with an upper tetrachord, 36 *mela*s, or *rāga* scales, were derived; a further 36 were formed by using F♯ in place of the F in the lower tetrachord. The *mela*s were named in such a way that the first two syllables of the name, when applied in a code, gave the number of that *mela* in the sequence. The musician, given the number, could easily reconstruct the scale of the *mela*. The names of the *mela*s were often derived from prominent *rāga*s in those *mela*s, with a two-syllable prefix that supplied the code numbers; for instance, the name of the *mela Dhīra-śaṅkarābharaṇa* is derived from the *rāga Śaṅkarābharaṇa*, the two syllables *dhīra* giving the code number 29, which indicates a scale similar to the Western major scale, or C mode. The *Caturdaṇḍiprakāśika* acknowledges the theoretical nature of its analytical system and mentions clearly that only 19 of the possible 72 *mela*s were in use at the time it was written.

Although North Indian texts also describe *rāga*s in terms of *mela*s or *thāṭa*s, there is no attempt to arrange them systematically. In the *Rāgataraṅgiṇi* ("The River of Rāga"), probably of the 16th century, 12 *mela*s are mentioned:

bhairavī	C	D	E♭	F	G	A	B♭	c
toṛī	C	D♭	E♭	F	G	A♭	B♭	c
gaurī	C	D♭	E	F	G	A♭	B	c
karṇāṭa	C	D	E	F	G	A	B♭	c
kedāra	C	D	E	F	G	A	B	c
imana	C	D	E	F♯	G	A	B	c
sāraṅga	C	D	E♯	F♯	G	A♯	B	c
megha	C	D	E	F	G	A♯	B	c
dhanāśrī	C	D♭	E	F♯	G	A♭	B	c
pūravā	C	D	E	F♯	G	A⁺	B	c
mukhārī	C	D	E♭	F	G	A♭	B♭	c
dīpaka	no description							

Although it appears from the description of *sāraṅga* and *megha mela*s that enharmonic intervals were used, there is good reason to believe that the E♯ and A♯ in the two *mela*s really represent their chromatic counterparts, F and B♭, and that F and F♯ (and B and B♭) do not appear in sequence. The A⁺ in the *mela pūravā* is said to be raised by one *śruti*. The description of the *rāga*s in these *mela*s shows that the North Indian system was by this time also based on 12 semitones.

THE MODERN PERIOD

With the collapse of the Mughal Empire in the 18th century and the emergence of the British as a dominant power in India, the subcontinent was divided into many princely states. Music continued to be patronized by the rulers, although the courts were never again to be so opulent.

Musically, there has been a continuous evolution from the Islāmic period to the present, and both North and South Indian classical music have continued to expand. South Indian music has clearly been influenced more by theory than has the North. The 72-*mela* system continues to be the basis of classifying the *rāga*s in South India, but it has had more than a classificatory significance. Many new *rāga*s have been composed in the past few centuries, some of them inspired by the theoretical scales of the *mela* system. As a result, there are now *rāga*s in all of the 72 *mela*s.

Influence of theory on composition

In North Indian music, theory has had little influence on performance practice. This can be ascribed to the language problem, an especially significant influence on the many Muslim musicians in North India, who were not able to cope with the Sanskrit musical literature. Thus, there had been no attempt to systematize the music, and there was a considerable gap between performance and theory until the present century. Vishnu Narayana Bhatkande, one of the leading Indian musicologists of this century, contributed a great deal toward diminishing the gap. Being both a scholar and a performer, he devoted much effort to collecting and notating representative versions of a number of *rāga*s from musicians belonging to different family traditions, or *gharānā*s. Based on this collection, he concluded that most of the *rāga*s of North Indian

music can be grouped into the following scales, called *thāṭa*s (compare the South Indian *mela*s shown above):

kalyāṇa	C	D	E	F♯	G	A	B	c
bilāvala	C	D	E	F	G	A	B	c
khamāja	C	D	E	F	G	A	B♭	c
bhairava	C	D♭	E	F	G	A♭	B	c
pūrvī	C	D♭	E	F♯	G	A♭	B	c
mārvā	C	D♭	E	F♯	G	A	B	c
kāfī	C	D	E♭	F	G	A	B♭	c
āsāvarī	C	D	E♭	F	G	A♭	B♭	c
bhairavī	C	D♭	E♭	F	G	A♭	B♭	c
toṛī	C	D♭	E♭	F♯	G	A♭	B	c

The *thāṭa*s do not cover all the *rāga*s used in North Indian music, but there is reason to believe that most of the *rāga*s having scales other than the above are relatively modern innovations. New *rāga*s are constantly being created, and some North Indian musicians are using the vast potential of the South Indian *mela* system as their source of inspiration.

Scale classifications as opposed to *rāga*

Mela and *thāṭa* are theoretical devices for the classification of *rāga*s. *Rāga*s have scalar elements, such as specified ascending and descending movements, that might or might not employ adjacent steps. They may also employ oblique or zigzag movements. *Rāga*s can be heptatonic, hexatonic, or pentatonic and may also have accidentals (sharpened or flattened notes) that occur only in specific melodic contexts. A further distinction between scale and *rāga* is found in the varying emphasis placed on different notes in a *rāga*. *Rāga*s, furthermore, also have melodic elements, such as certain recurrent nuclear motives (brief melodic fragments) that enable the *rāga* to be identified more easily. One scale type can be the basis for perhaps 20 or 30 *rāga*s, in which case it is the nonscalar elements that provide the distinguishing features of each *rāga*.

Rhythmic organization. *South India.* Just as the system of classifying *rāga* is better organized in South Indian music, so too is the system of classifying *tāla*, or time measure. The main group is composed of 35 *tāla*s, called the *sūlādi-tāla*s. Each *tāla* is composed of one, two, or three different units: short, medium, and long. The medium unit is twice the duration of the short; the long unit is, however, a variable and may be three, four, five, seven, or nine times the duration of the short. There are seven basic *tāla* patterns, and, because the long unit of these *tāla*s can be of five different durations, the total number of *tāla*s in this system is 35. The basic *tāla* patterns are:

dhruva-tāla—long, medium, long, long
maṭhya-tāla—long, medium, long
rūpaka-tāla—medium, long
jhampā-tāla—long, short, medium
tripuṭa-tāla—long, medium, medium
āṭa-tāla—long, long, medium, medium
eka-tāla—only a single long.

The total duration of each pattern is controlled by the duration of the variable long; thus, if the long unit is five times the short, a *tāla* pattern such as *dhruva-tāla* will be 5+2+5+5, or 17 units. Several of these *tāla*s have the same total duration but are distinguished from each other by their internal subdivisions. In the course of a performance, the vocalist, as well as the audience, may mark the time by clapping, hand waving, and finger counting.

In addition to the *sūlādi-tāla*s, there are four *cāpu-tāla*s that are used in South Indian classical music. Said to derive from folk music, they consist of two sections of unequal length, 1+2, 2+3, 3+4, and 4+5. Of these, the 3+4 combination is the most prominent. On rare occasions a performer may use one of the "classical" *tāla*s referred to in Sanskrit texts. These generally involve long time cycles composed of as many as 100 short units. The most frequently heard time measures, however, are *ādi-tāla*, a modified eight-beat version of *tripuṭa-tāla* (4+2+2); *miśra-cāpu-tāla* (3+4); and *rūpaka-tāla* (4+2). The difficult and long *tāla*s are used primarily as a tour de force. Each *tāla* may be performed in either slow, medium, or quick tempo; there is no gradual acceleration as in North Indian music.

North India. In North Indian music the *tāla*s are fewer and not organized in any systematic manner. As in South Indian music, the two main factors are the duration of the time cycle and the subdivisions within the cycle. Each of these subdivisions is marked by a clap or a wave, with the greatest emphasis falling on beat 1 of the cycle, which is called *sam*. North Indian *tāla*s have a further feature, the *khālī* ("empty"), a conscious negation of stress occurring at one or more points in each *tāla* where one would expect a beat. It often falls at the halfway point in the time cycle and is marked by a wave of the hand. There is nothing comparable to the *khālī* in the South Indian system. A further distinguishing feature found only in North Indian *tāla*s is the emphasis placed on the characteristic drum pattern of each *tāla*, called *ṭhekā*. Two *tāla*s might have the same duration and subdivisions but might, nevertheless, be differentiated from each other by different characteristic drum patterns. In addition, the *tāla*s are also associated with different forms of song and even particular tempi. The usual North Indian *tāla*s range from six to 16 time units in duration. The most popular are *tīn-tāla* (4+4+4+4), *eka-tāla* (2+2+2+2+2+2), *jhap-tāla* (2+3+2+3), *kaharavā* (4+4), *rūpaka-tāla* (3+2+2), and *dādrā* (3+3). *Tīn-tāla* should not be confused with Western $\frac{4}{4}$, or common time, for the time cycle repeats only after 16 units and is more like four bars of common time.

Characteristic North Indian rhythmic features

Musical forms and instruments. *South India.* Both *rāga* and *tāla* provide bases for composition and improvisation in Indian classical music. A performance usually begins with an improvised section, called *ālāpa*, played in free time without accompaniment of drums. It may have various sections and might on occasion last half an hour or longer. It is followed by a composed piece in the same *rāga*, set in a particular *tāla*. In South Indian music all composed pieces are primarily for the voice and have lyrics. In North India, however, there are also some purely instrumental compositions, called *gat* and *dhun*. The emphasis on the composition varies in the different forms of song and, to some extent, in the interpretation of the performer. In South Indian music the composed piece is generally emphasized more than in the North. Much of the South Indian repertoire of compositions stems from three composers, Tyagaraja, Muthuswami Dikshitar, and Syama Sastri, contemporaries who lived in the second half of the 18th and the beginning of the 19th centuries. The devotional songs that they composed, called *kṛti*, are a delicate blend of text, melody, and rhythm and are the most popular items of a South Indian concert. The composed elements in these songs sometimes include sections such as *niraval*, melodic variations with the same text, and *svara-kalpana*, passages using the Indian equivalent of the sol–fa syllables, which are otherwise improvised.

Interaction of composition and improvisation

The longest item in the South Indian concert, called *rāgam-tānam-pallavi*, is, on the other hand, mostly improvised. It begins with a long *ālāpa*, called *rāgam* in this context, presumably because this elaborate, gradually developing *ālāpa* is intended to display the *rāga* being performed in as complete a manner as possible, without the limitations imposed by a fixed time measure. This is followed by another improvised section, *tānam*, in which the singer uses meaningless words to produce more or less regular rhythms, but still without reference to time measure. This section, too, is without drum accompaniment. The final section, *pallavi*, is a composition of words and melody set in a particular *tāla*, usually a long or complex one. The *pallavi* may have been composed by the performer himself and be unfamiliar to his accompanists, usually a violinist who echoes the singer's phrases and a drummer who plays the *mṛdaṅgam*, a double-ended drum. The statement of the composition is followed by elaborate rhythmic and melodic variations that the accompanists are expected to follow. It is customary to have a drum solo at the end of the *pallavi*, and the performance concludes with a brief restatement of the *pallavi*.

Other forms used in South Indian classical music derive largely from the musical repertoire of *bhārata-nāṭyam*, the classical South Indian dance. The *varṇam*, a completely composed piece, serves mainly as a warming up and is performed at the beginning of a concert. *Pada* and *jāvali* are two kinds of love songs using the poetic imag-

ery characteristic of the romantic–devotional movement mentioned earlier. *Tillānā* has a text composed mostly of meaningless syllables, which may include the onomatopoeic syllables used to represent the different drum sounds. This is a very rhythmic piece and is usually sung in fast tempo.

Melodic, drone, and rhythmic instruments

The ensemble used in present-day South Indian classical music consists of a singer or a main melody instrument, a secondary melody instrument, one or more rhythmic percussion instruments, and one or more drone instruments. The most commonly heard main melody instruments are the *vīṇā*, a long-necked, fretted, plucked lute with seven strings; the *veṇu*, a side-blown bamboo flute; the *nāgasvaram*, a long, oboe-like, double-reed instrument with finger holes; the violin, imported from the West about 200 years ago, played while seated on the floor with the scroll resting on the player's left foot; and the *goṭṭuvādyam*, a long-necked lute without frets, played like the Hawaiian guitar, with a sliding stop in the left hand.

The violin is by far the most common secondary melody instrument in South India. It plays in unison where the passage is composed but imitates the voice or main melody instrument in the improvised passages. Of the rhythm instruments, the *mṛdaṅgam*, a double-conical, two-headed drum, is the most common. Others include the *kañjīrā*, a tambourine; the *ghaṭam*, an earthenware pot without skin covering; the *morsiṅg*, a metallic jew's harp; and the *tavil*, a slightly barrel-shaped, double-ended drum, which accompanies the *nāgasvaram*. The most prominent drone instrument is the four-stringed *tambūrā*, a long-necked lute without frets. It accompanies the voice and all melody instruments, except the *nāgasvaram*, which is usually accompanied by the *oṭṭu*, a longer version of the *nāgasvaram* but without finger holes. A hand-pumped harmonium drone, called *śruti* or *śruti* box, sometimes replaces the *oṭṭu* or the *tambūrā*.

North India. The most common vocal form in North Indian classical music at the present time is the *khyāl*, a Muslim word meaning "imagination." The *khyāl* is contrasted with the *dhruvapada* (now known as *dhrupad*), which means "fixed words." The two forms existed side by side in the Islāmic period, and it is only in the last century or two that *khyāl* has achieved ascendancy. There are two types of *khyāl*. The first is sung in extremely slow tempo, with each syllable of the text having extensive melisma (prolongation of a syllable over many notes), so that the words are virtually unrecognizable. It is not usually preceded by a lengthy *ālāpa;* instead, *ālāpa*-like phrases are generally sung against the very slow time measure to the accompaniment of the drums. Also characteristic of the *khyāl* are the *sargam tānas*, passages using the Indian equivalent of the sol–fa syllables, and the *ā-kār tanas*, which are rapid runs sung to the syllable *aah*. The second type of *khyāl*, which may be as much as eight times faster than the slow and is generally set in a different *tāla*, follows the slow. Its composed portion is usually quite short, and the main features of the improvisation are the *ā-kār tānas*. Occasionally, a composition called *tarānā*, made up of meaningless syllables, may replace the fast-tempo *khyāl*.

The *ṭhumrī* is another North Indian vocal form and is based on the romantic–devotional literature inspired by the *bhakti* movement. The text is usually derived from the Rādha-Kṛṣṇa theme and is of primary importance. The words are strictly adhered to, and the singer attempts to interpret them with his melodic improvisations. It is quite usual for a singer to deviate momentarily from the *rāga* in which the composition is set, by using accidentals and evoking other *rāgas* that might be suggested by the words, but he always returns to the original *rāga*.

Some of the North Indian musical forms are very like the South Indian. The vocal forms *dhrupad* and *dhamār* resemble the *rāgam-tānam-pallavi*. They begin with an elaborate *ālāpa* followed by the more rhythmic but unmeasured *non-tom* using meaningless syllables such as *te, re, na, nom,* and *tom*. Then follow the four composed sections of the *dhrupad* or *dhamār*, the latter being named after *dhamār-tāla* of 14 units (5+5+4) in which it is composed, the former name derived from *dhruvapada*. The song, usually in slow or medium tempo, is first sung as composed; then the performer introduces variations, the words often being distorted and serving merely as a vehicle for the melodic and rhythmic improvisations. Although the *dhrupad-dhamār* form has been out of favour for over a century, it is now apparently being revived.

Instrumental forms

Instrumental music has gained considerable prominence in North India in recent times. The most common instrumental form is the *gat*, which seems to have derived its elements from both *dhrupad* and *khyāl*. It is usually preceded by *ālāpa* and *joṛ*, which resemble the *ālāpa* and *non-tom* sections of the *dhrupad*. On plucked stringed instruments these two movements are often followed by *jhālā*, a fast section in which the rhythmic plucking of the drone strings is used to achieve a climax. The performer usually pauses before the composed *gat* is introduced. Like the *khyāl*, the *gat* can be in slow or fast tempo. The composition is generally short, and the emphasis is on the improvisations of the melody instrumentalist and the drummer, who for the most part alternate in their extemporizing. The final climax may once again be achieved by a *jhālā* section, in which the tempo is accelerated quite considerably. Other forms played on instruments are the *ṭhumrī*, basically an instrumental rendering of a vocal *ṭhumrī*, and *dhun*, which is derived from a folk tune and does not usually follow a conventional *rāga*. One may also hear a piece called *rāga-māla* (literally, "a garland of *rāgas*"), in which the musician modulates from one *rāga* to another, finally concluding with a return to the original *rāga*.

Musical instruments of North India

The most prominent melody instruments used in North Indian classical music are the *sitār*, a long-necked fretted lute; *surbahār*, a larger version of the *sitār*; the *sarod*, a plucked lute without frets and a shorter neck than that of the *sitār*; the *sāraṅgī*, a short-necked bowed lute; the *bāṃsrī*, a side-blown bamboo flute with six or seven finger holes; the *shahnā'ī*, a double-reed wind instrument similar to the oboe, but without keys; and the violin, played in the same manner as in South India. Secondary melody instruments are used only in vocal music, the two most common being the *sāraṅgī* and the keyboard harmonium, an import from the West. The violin and the *surmaṇḍal*, a plucked board zither, are also used in this context. In recent times, instrumental duets, in which the musicians improvise alternately, have grown in popularity. In these duets the musicians may imitate each other's phrases, temporarily creating something of the effect of a secondary melody instrument.

As with South Indian music, the drone is usually provided by a *tambūrā* or a hand-pumped reed drone similar to the harmonium but without a keyboard, called *sur-peṭī* in North India. The *shahnā'ī* is usually accompanied by one or more drone *shahnā'ī*s, called *sur*.

The rhythmic accompaniment is usually provided on the *ṭablah*, a pair of small drums played with the fingers. As accompaniment to the somewhat archaic *dhrupad*, however, the *pakhāvaj*, a double-conical drum, similar to the South Indian *mṛdaṅgam*, is generally used. The *shahnā'ī* in classical music is usually accompanied by a small pair of kettledrums, called *ḍukaṛ-ṭikaṛ*.

Interaction with Western music. It is in the sphere of musical instruments that the influence of Western music is most obvious. In addition to the violin and the harmonium, many other Western instruments are occasionally used. These include the clarinet, saxophone, trumpet, guitar, mandolin, and organ. Scholars have criticized the use of some of these instruments on the ground that their tuning, being based on the Western tempered scale (having 12 equal semitones), is not suitable for the performance of Indian music, and All-India Radio has forbidden the use of the harmonium in its programs. Most of the leading North Indian singers, however, have been using the harmonium as a secondary melody instrument for many years and have continued to do so in concerts and on recordings.

Apart from the area of musical instruments, Indian music appears to have absorbed very little of Western music. It is possible that some modern developments in classical music might have been inspired by Western music. These

include the slightly increased use of chromaticism (using a succession of semitones) and some of the new drone tunings in which the major third is added (making for example, the drone on the first, third, and fifth notes of the scale, rather than on the first and fifth only). But the evidence is not conclusive, and it could equally be argued that these are natural developments within the system. Western technology has, of course, had a profound influence on Indian music. Sound-amplification devices have made concerts available to large audiences, and the intimate atmosphere in which the music was traditionally performed is now seldom encountered. The Indian musician has been obliged to adapt his music, once played before a select and musically educated group of listeners, to new circumstances involving a mass of people, many of whom are unable to appreciate the finer points of the music. The use of microphones during concerts has had a marked effect on voice production, and, since the voice no longer needs to project over distances, many modern singers now sing with a relaxed throat and produce a more mellow tone.

Impact outside of India

Since the mid-1950s, Indian classical music has been performed fairly regularly in the West. Initially, the audiences were composed mainly of South Asians, but gradually an increasing number of Westerners have been attending the concerts. Perhaps the music would not have reached beyond a very limited audience were it not for the interest shown by the American violinist Yehudi Menuhin, who sponsored a number of programs in the West, and the British popular-music group the Beatles, who attempted to incorporate the sound of the *sitār* and other elements of Indian culture into the world of Western popular music. At the same time, several North Indian instrumentalists, such as Ravi Shankar, Ali Akbar Khan, Vilayat Khan, Imrat Khan, and Nikhil Banerjee, were received with overwhelming enthusiasm by Western audiences. By the end of the 1960s the *sitār* and *ṭablah* were heard frequently in Western pop music, jazz, cinema, and television programs, as well as in radio and television advertisements.

Within three or four years, the mass Western involvement with Indian music was over. It is perhaps too early to assess the full impact of this period. It would seem that Indian music has not as yet had any significant influence on Western music. A few modern composers have attempted to incorporate elements of Indian music into their compositions, but their works remain as experiments. The fusing of Indian music and jazz would seem to have more possibilities, since improvisation is an important factor in both, but present attempts have not fulfilled this expectation. Most of these attempts seem to be premature and based on an inadequate understanding of one or the other musical system. (N.A.J.)

III. Dance and theatre

Theatre and dance in South Asia stem principally from Indian tradition. The principles of aesthetics and gesture language in the *Nāṭya-śāstra*, a 2,000-year-old Sanskrit treatise on dramaturgy, have been the mainstay of all the traditional dancers and actors in India. Even folk performers follow some of its conventions; *e.g.*, the *Kandyan* dancers of Ceylon (now Sri Lanka), who preserve some of the whirls and spins described in this ancient Indian text. Despite the influence of the different religious waves that swept the subcontinent through the centuries, the forms of South Asian dance and theatre were always able to preserve their ancient core.

Traditionally, dance and acting are inseparable. The classical South Asian dancer, equipped with a repertoire of gesture language, alternates between *nṛta*, pure dance; *nṛtya*, interpretive dance; and *nāṭya*, dance with a dramatic element. (The Sanskrit word *naṭa* means a dancer-actor.) Traditional theatre throughout both South and Southeast Asia is a combination of music, dance, mime, stylized speech, and spectacle. The classical and folk actor must be a dancer, a singer, and a mime in one.

Influence on Southeast Asia. Between the 2nd century BC and the 8th century AD, South Indian kings sent overseas trade missions, priests, court dancers, and sometimes armies to Southeast Asia. During these years of cultural expansion, Indian dance forms, mythological lore, and the language of gesture flourished in Burma, Cambodia, Java, Sumatra, and Bali. Later, when India's economic and political power shrank, its cultural empire remained intact. Even when these Southeast Asian countries embraced Buddhism or Islām, they continued performing dance dramas with Hindu gods and goddesses, adding to these their own local myths, costumes, and masks. The two Hindu epics, the *Rāmāyaṇa* and the *Mahābhārata*, storehouses of dramatic personae of traditional dramas, have been absorbed by these countries as part of their own cultural heritage. Some dance forms and gesture vocabulary that died out in their land of birth have been preserved in Bali. For a discussion of the dance and theatre of Southeast Asia see the article SOUTHEAST ASIAN PEOPLES, ARTS OF.

THE PERFORMING ARTS IN INDIA

The royal courts and temples of India traditionally have been the chief centres of the performing arts. In ancient times, Sanskrit dramas were staged at seasonal festivals or to celebrate special events. Some kings were themselves playwrights; the most notable of the playwright-kings was Śūdraka, the supposed 4th-century author of *Mṛcchakaṭika* (The Little Clay Cart). Other well-known royal dramatists include Harṣa, who wrote *Ratnāvalī* in the 7th century; Mahendravikramavarman, author of the 7th-century play *Bhagavad-Ajjukiya*; and Viśākhadatta, creator of the 9th-century drama *Mudrārākṣasa*.

In the 4th century BC, Kauṭilya, the chief minister of Emperor Candragupta, referred in his book on the art of government, the *Artha-śāstra*, to the low morals of players and advised the municipal authorities not to build houses in the midst of their villages for actors, acrobats, and mummers. But, in the glorious era of the Hindu kings during the first eight centuries after Christ, actors and dancers were given special places of distinction. This tradition continued in the princely courts of India even under British rule. *Kathākali* dance-drama, for instance, was created by the Raja of Kottarakkara, ruler of one of the states of South India in the 17th century. The powerful peshwas of the Marāthā kingdom in the 18th century patronized the *tamasha* folk theatre. Nawab Wajid Ali Shah (flourished mid-19th-century) was an expert *kathak* dancer and producer of Krishnalore plays in which his palace maids danced as the *gopīs* (milkmaids who were devotees of Kṛṣṇa). Maharajas of Travancore and Mysore competed with each other for the excellence of their dance troupes. In the 20th century, the Maharaja of Banaras (Vārānasi) carried on this tradition by being patron and producer of the spectacular *rāmlīlā*, a 31-day cycle play on Rāma's life that he witnessed every night while sitting on his royal elephant. On special nights the spectators numbered more than 30,000.

Royal patronage

Dance is a part of all Hindu rituals. Farmers dance for a plentiful harvest, hunters for a rich bag, fishermen for a good catch. Seasonal festivals, religious fairs, marriages, and births are celebrated by community dancing. A warrior dances before the image of his goddess and receives her blessings before he leaves for battle. A temple girl dances to please her god. The gods dance in joy, in anger, in triumph. The world itself was created by the Cosmic Dance of Lord Śiva, who is called Naṭarāja, the king of dancers, and worshipped by actors and dancers as their patron.

Religious festivals

Religious festivals are still the most important occasions for dance and theatrical activity. The *rāmlīlā kṛṣṇalīlā* and *rāslīlā* in North India (Uttar Pradesh, Delhi, Rājasthān, Haryana, and Punjab), the *chhau* masked dance-drama in Saraikela region in Bihār, and the *bhagavatha mela* in Melatur village in Tamil Nadu are performed annually to celebrate the glory of their particular deities. During the Daśaharā festival every village in North India enacts for a fortnight the story of Rāma's life, with songs, dances and pageants. The *jātrā* in West Bengal is a year-round dramatic activity, but the number of troupes swells to many thousands in Calcutta during the Pūjā festival. The hill and tribal people dance all night to celebrate

their community festivals and weddings rich in masks, pageants, and carnivals. In far-flung areas of South Asia, people may not have seen a drama, but there will be hardly a person who has not witnessed or taken part in a community dance.

Audience participation

In folk theatre, traditional dance, classical music, and poetical symposia (especially the Urdu *mushāʿirah*), performances are held in the open air or in a well-lit canopied courtyard so that the players can see the spectators and be motivated by their reactions.

For the usually all-night folk dramas, people come with their children, straw mats, and snacks, making themselves at home. There is a constant inflow and outflow of spectators. Some go to sleep, asking their neighbours to awaken them for favourite scenes. Stalls selling betel leaves, peanuts, and spicy fried things, adorned with flowers and incense and lighted by oil lamps, surround the open-air arena. The clown, an essential character in every folk play, comments on the audience and contemporary events. Zealous spectators offer donations and gifts in appreciation of their favourite actor or dancer, who receives them in the middle of the performance and thanks the donor by singing or dancing a particular piece of his choice. The audience constantly throws sparks to the performer, who throws them back. People laugh, weep, sigh, or suddenly fall silent during a moving scene.

In both folk and classical forms of drama, the performer may lengthen or shorten his piece according to audience response. During a *kathak* dance, the drummer, in order to test the perfection of the dancer, disguises the main beat of his drum by slurs and offbeats, a secret he shares with the audience and announces by a loud thump synchronized with the dancer's stamping of the foot. At this point the spectators shout, swaying their heads in admiration. They show their approval and disapproval through delighted groans or sullen headshakes as the performance goes on. In the *rāslīlā*, the audience joins in singing the refrain and marks the beat by hand clapping. At a climactic point the people rock and sway, rhythmically clapping and singing. These practices bind the performers, chanters, and spectators together in a sense of aesthetic pleasure.

Instrumental music and singing

Instrumental music and singing are integral parts of Indian dance and theatre. Musicians, chanters, and drummers sit on the stage in view, a tradition observed throughout almost all of Asia. They watch the dancer and play on their instruments following his movements, whereas in the West the movements of a ballerina are timed and controlled by the already-written music. An Indian dancer is constantly reacting to the accompanying musician, and vice versa. He may signal the chanters and drummers and even instruct them during the performance without spoiling its aesthetic effect.

In some classical dance forms, such as *kuchipudi*, the dancer sings in voiceless whispers as she dances. In *bhārata-nāṭyam* the dance movements are like sculpted music in space, and the accompanying musician is invariably a dance guru (teacher). In *kathak* the rhythmic syllables beaten out by the dancer with her feet are vocalized by the singer and then chirped out by the drummer. No folk dancing is complete without the use of drum and vocal singing. Women's folk singing such as the *giddha* in the Punjab and the men's *kīrtan* in West Bengal takes the form of dance when the rhythm becomes fast.

In folk theatre this relationship is even more apparent. *Rāslīlā* dance sequences are interspersed with the singing as a decorative frill, to accentuate emotional appeal, or to mark the climax of a song. The *yakṣagāna* hero gives a brisk dance number to announce his entry. In many folk forms of opera (*bhavai, terukkūttu*, and *nautanki*), the characters sing and dance at the same time or alternate. Ballad singers from the states of Orissa and Andhra Pradesh dramatize their singing by strong facial gestures and rhythmic ankle bells and execute dance phrases between the narrative singing. On the other hand, no one can imagine a dancer who is not at the same time a musician. This double aesthetic discipline enriches both of these arts, and the Indian audience is conditioned to this tradition.

INDIAN DANCE

Dance in India can be organized into three categories: classical, folk, and modern. Classical dance forms are among the best preserved and oldest practiced in the 20th century. The royal courts, the temples, and the guru to pupil teaching tradition have kept this art alive and unchanged. Folk dancing has remained in rural areas as an expression of the daily work and rituals of village communities. Modern Indian dance, a product of the 20th century, is a creative mixture of the first two forms, with freely improvised movements and rhythms to express the new themes and impulses of contemporary India.

The popularity of dance in 20th-century India can be judged from the fact that there is hardly any Indian film that does not have half a dozen dances in it. In the typical "boy meets girl" film the heroine dances everywhere and anywhere. A film company may not have a script writer (in some cases the financier writes the story himself), but it must have a dance director. To provide ample dance opportunities, films have been made on the lives of poets, courtesans, and temple dancers and on mythological themes. For these the services of expert dancers are sought.

In the 20th century, classical dance has left the temples and royal courts and is now presented regularly on the stage in large cities. Rich industrialists, international hotels, and the wealthy families of the upper class are the chief patrons. It is not uncommon to have a classical-dance recital by a major performer at a business dinner or for the annual function of a club. Some universities have dance as a regular subject in their curricula. Women learn it as a social grace, and young girls learn a few classical dances for greater eligibility in marriage. Folk dancing has also become more common as a contemporary cultural event in the cities. Most colleges have their folk-dance troupes, and even the police of the Punjab have their folk-dance groups to perform the *bhangra*. Folk dance, cut off from its rural settings, has lost much of its original vigour and beauty, but that is the inevitable result of cross-fertilization of regional cultures through folk-troupe exchanges at an interstate level.

Classical dance. *The dance-drama.* India has evolved through its classical and folk traditions a type of dance-drama that is a form of total theatre. The actor dances out the story through a complex gesture language, a form that, in its universal appeal, cuts across the multilanguage barrier of the subcontinent. Some of the classical dance-drama forms (*e.g., kathākali, kuchipudi, bhagavatha mela*) enact well-known stories derived from Hindu mythology. The 20th-century dancers Uday Shankar and Shanti Bardhan have created ballets inspired by such traditional dance-dramas. Contemporary Indian directors and writers are re-examining traditional dance forms and using these in their current works for greater psychological appeal and deeper artistic impact. Millions in villages are still entertained by dance-dramas. In spite of the popularity of straight prose plays in the cities, the appeal of dance-drama is unquestionably deeper and more satisfying to the rural Indian, whose aesthetics is still rooted in tradition.

The *Nāṭya-śāstra*

The chief source of classical dance is Bharata Muni's *Nāṭya-śāstra* (1st century BC to 1st century AD), a comprehensive treatise on the origin and function of *nāṭya* (dramatic art that is also dance), on types of plays, gesture language, acting, miming, theatre architecture, production, makeup, costumes, masks, and various *bhāvas* ("emotions") and *rasas* ("sentiments"). No other book of ancient times contains such an exhaustive study of dramaturgy.

Techniques and types of classical dance. According to the *Nāṭya-śāstra*, the dancer-actor communicates the meaning of a play through four kinds of *abhinaya* (histrionic representations): *angika*, transmitting emotion through the stylized movements of parts of the body; *vācika*, speech, song, pitch of vowels, and intonation; *āhārya*, costumes and makeup; and *sāttvika*, the entire psychological resources of the dancer-actor.

The actor is equipped with a complicated repertoire of stylized gestures. Conventionalized movements are pre-

scribed for every part of the body, the eyes and hands being the most important. There are 13 movements of the head, seven of the eyebrows, six for the nose, six for the cheek, seven for the chin, nine for the neck, five for the breasts, and 36 for the eyes. There are 32 movements of feet, 16 on the ground and 16 in the air. Various positions of the feet (strutting, mincing, tromping, splaying, beating, etc.) are carefully worked out. There are 24 single-hand gestures (*asaṃyuta-hasta*) and 13 for combined hands (*saṃyuta-hasta*). One gesture (*hasta*) may mean over 30 different things quite unrelated to each other. The *patāka* gesture of the hand, for example, in which all the fingers are extended and held close together with the thumb bent, can represent heat, rain, a crowd of men, the night, a forest, a horse, or a flight of birds. The *patāka* hand with the third finger bent (*tripatāka*) can mean a crown, a tree, marriage, fire, a door, or a king. In *karkaṭa* ("crab"), one of the combined hand gestures, the fingers of the hands are interlocked, and this may indicate a honeycomb, yawning after sleep, or a conch shell. Of course, for each of these different meanings a *hasta* is given a different body posture or action.

The male or female classical dancer portraying a story in a solo performance simultaneously plays two or three principal characters by alternating facial expressions, gestures, and moods. Kṛṣṇa, his jealous wife Satyabhāmā, and his gentle wife Rukmiṇī, for example, may be played by one person.

Bhāva and rasa

The aesthetic pleasure of Hindu dance and theatre is determined by how successful the artist is in expressing a particular emotion (*bhāva*) and evoking the *rasa*. Literally *rasa* means "taste" or "flavour" and is that exalted sentiment or mood that the spectator experiences after witnessing a performance. The critics do not concern themselves so much about plot construction or technical perfection of a poem or play as about the *rasa* of a particular work. There are nine *rasa*: erotic, comic, pathetic, furious, heroic, terrible, odious, marvellous, and spiritually peaceful. There are nine corresponding *bhāvas*: love, laughter, pathos, anger, energy, fear, disgust, wonder, and quietude.

Four distinct schools of classical Indian dance—*bhārata-nāṭyam, kathākali, kathak, and manipuri*—exist in the 20th century, along with two types of temperament—*tāṇḍava*, representing the fearful male energy of Śiva, and *lāsya*, representing the lyrical grace of Śiva's wife Pārvatī. *Bhārata-nāṭyam*, which takes its name from Bharata's *Nāṭya-śāstra*, has the *lāsya* character, and its home is Tamil Nadu, in South India. *Kathākali*, a pantomimic dance-drama in the *tāṇḍava* mood with towering headgear and elaborate facial makeup, originated in Kerala. *Kathak* is a mixture of *lāsya* and *tāṇḍava* characterized by intricate footwork and mathematical precision of rhythmic patterns; it flourishes in the north. *Manipuri*, with its swaying and gliding movements, is *lāsya*, and it has been preserved in Manipur state in the Assam Hills. In 1958 the Sangeet Natak Akademi (National Academy of Music, Dance and Drama) in New Delhi bestowed classical status on two other schools of dance—*kuchipudi*, from Andhra Pradesh, and *Odissi*, from Orissa. These two styles overlap the *bhārata-nāṭyam* school and therefore are not as distinctly different in temperament and style as other forms.

The bhārata-nāṭyam school. *Bhārata-nāṭyam* (also called *dasi attam*) has survived to the present through the *devadāsīs*, temple dancing girls who devoted their lives to their gods through this medium. Muslim invasions from the north destroyed the powerful Hindu kingdoms in the south but could not disrupt their arts, which took shelter in the temples. After the 16th century the Muslims overpowered the south completely until the British came, thus giving a setback to Hindu dance. Slowly the institution of *devadāsī* fell into disrepute, and temple dancing girls became synonymous with prostitutes. In the latter half of the 19th century in Tanjore, Chinniyah Punniah, Vadivelu, and Shivanandam, four talented dancers who were brothers, revived the original purity of *dasi attam* by studying and following the ancient texts and temple friezes, with missing links supplied by the socially spurned *devadāsīs*. Their popularized form of *dasi attam* was called *bhārata nāṭyam*.

A performance of *bhārata nāṭyam* lasts for about two hours and consists of six parts, beginning with *allarippu* (Telegu language, "to decorate with flowers"), a devotional prologue that shows off the elegance and grace of the dancer. The second part is *jātisvaram*, a brilliant blaze of *jāti*s ("dance phrases") with *svara*s ("musical sounds"). This is followed by *shabdam*, the singing words that prepare the dancer to interpret through *abhinaya* (gesture language) interspersed with pure dance. The fourth part is *varṇam*, a combination of expressive and pure dance. Then follow the *padam*s, songs in Telegu, Tamil, or Kanarese that the dancer dramatizes by facial expressions and hand gestures. The accompanying singer chants the line again and again, and the dancer enacts the clashing and contrasting meanings. Her virtuosity consists of exhausting all possible shades of suggestion. The performance ends with *tillānā*, a pure dance accompanied by meaningless musical syllables chanted to punctuate the rhythm. The dancer explodes into leaps and jumps forward and backward, from right and left, in a state of ecstasy. *Tillānā* ends with three clangs of the cymbals while the dancer executes a triple blaze of *jāti*s, thumping her feet with a jingling flourish of ankle bells.

During the last 20 years *bhārata-nāṭyam* has attained world recognition as the most exquisite form of classical dance. Its aspirants go to Tamil Nadu to learn from gurus who still live in villages. Because of its *lāsya* character, performing artists have always been women. But their teachers have invariably been old men who chant the lines to tiny cymbals, controlling the complex rhythm without dancing themselves.

The major 20th-century performers associated with the *bhārata-nāṭyam* school of dance are T. Balasaraswathi, especially known for her *abhinaya* (expressive interpretation) of *padam*s; Rukmini Devi, who popularized *bhārata-nāṭyam* among the upper classes in the 1930s; Yamini Krishnamurthi; and Shanta Rao. Two of the most important gurus are Minakshisundaram Pillai, who injected vigour into *bhārata-nāṭyam* by his choreography, and his son-in-law, Chokkalingam Pillai.

The kathākali school. *Kathākali* (*kathā*, "story"; *kali*, "performance") originated in the 17th century in Kerala, the lush tropical coastal strip of South India washed by the Arabian Sea. It was devised by the Raja of Kottarakkara, who, angry over the refusal of a neighbouring prince to allow his dancers to perform a Sanskrit dance-drama in his court, decided to create his own dance troupe using Malayalam, the spoken language of the people. This school of dance has its own *hasta*s, based on a regional text influenced by the *Nāṭya-śāstra* and later treatises. It also has marked elements of energetic ritualistic dances. The makeup has its roots in the grotesque Dravidian demon masks of the pre-Hindu period. Themes are taken mainly from the *Rāmāyaṇa*, the *Śiva-Purāṇa*, the *Bhāgavata-Purāṇa*, the *Mahābhārata*, and other religious texts. The superhuman characters represent primal forces of good and evil at war. Because of its terrifying vigour, men play all the roles.

Characters and costumes

Most *kathākali* characters (except those of women, Brahmins, and sages) wear towering headgear and billowing skirts and have their fingers fitted with long silver nails to accentuate hand gestures. The principal characters are classified into seven types. (1) *Pacca* ("green") is the noble hero whose face is painted bright green and framed in a white bow-shaped sweep from ears to chin. Heroes such as Rāma, Lakṣmaṇa, Kṛṣṇa, Arjuna, and Yudhiṣṭhira fall into this category. (2) *Katti* ("knife"), haughty and arrogant but learned and of exalted character, has a fiery upcurled moustache with silver piping and a white mushroom knob at the tip of his nose. Two walrus tusks protrude from the corners of his mouth, his headgear is opulent, and his skirt is full. Duryodhana, Rāvaṇa, and Kichaka belong to this type. (3) *Chokannatadi* ("red beard"), power-drunk and vicious, is painted jet black from the nostrils upward. On both cheeks semicircular strips of white paper run from the upper lip to the eyes. He has black lips, white warts on nose and



Indian classical dance.
(Top left) *Bhārata-nātyam* traditional dance drama. (Top right) Male and female *kathākali* dancers. (Bottom left) *Kathak* school dancer in Mughal costume. (Bottom right) *Manipuri* style performance of *rās.*
(All but top right) Mohan Khokar, (top right) Foto Features

forehead, two long curved teeth, spiky silver claws, and a bloodred beard. (4) *Velupputadi* ("white beard") represents Hanuman, son of the wind god. The upper half of his face is black and the lower red, marked by a tracery of curling white lines. The lips are black, the nose is green, black squares frame the eyes, and two red spots decorate the forehead. A feathery gray beard, a large furry coat, and bell-shaped headgear give the illusion of a monkey. (5) *Karupputadi* ("black beard") is a hunter or forest dweller. His face is coal black with crisscross lines around the eyes. A white flower sits on his nose, and peacock feathers closely woven into a cylinder rise above his head. He carries a bow, quiver, and sword. (6) *Kari* ("black") is intended to be disgusting and gruesome. Witches and ogresses, who fall into this category, have black faces marked with queer patterns in white and huge, bulging breasts. (7) *Minnukku* ("softly shaded") represents sages, Brahmins, and women. The men wear white or orange dhotis. Women have their faces painted light yellow sprinkled with mica, and their heads are covered by saris.

Under a flower-decked canopy on a square, ground-level stage a tall brass worship lamp brimming with coconut oil burns brightly. The musicians and dancers bow before it before they start performing. Drummers standing in one corner pound the *cenda,* a barrel-shaped drum with a piercing, clattering sound suited for battle scenes, and continue throughout the performance, almost without respite. Two men hold a 12-foot by six-foot (four-metre by two-metre) embroidered hand curtain from behind which the principal characters make their entrances. They dance, grab the trembling curtain, and give vivid facial expressions with fearful glances and grunts. This "peering over the curtain," called *tiranokku,* is a close-up that offers an actor full scope to display his art. At a climactic moment the curtain is whisked away and the character enters in full splendour. The performance lasts all night, the singers singing the text that the dancers act out in an elaborate gesture language.

Well-known performers of *kathākali* include Guru Chandu Panikkar, Guru Kunju Kurup, Ramunni Nair, and Kalamandalam Krishna Nair. The dancers Guru Gopi Nath and Krishnan Kutty have both emphasized simplification of the use of towering headgear and thick-crusted, elaborate makeup, so that the art may be more commonly understood.

The kathak school. *Kathak*, born of the marriage of Hindu and Muslim cultures, flourished in North India under Mughal influence. *Kathak* dancers retain their 17th-century costumes but are steeped in Rādhā and Kṛṣṇa love lore. Kṛṣṇa, playing his flute in the Vṛndāvana woods on the bank of the Yamuna River, is surrounded by the *gopīs* ("milkmaids"). Their play is the eternal game of the god and his devotees, the hide-and-seek of man and woman. This spiritual relationship is deeply passionate, with erotic love-play. Slowly the dance degenerated and found shelter in bawdy houses, where nautch girls practiced the art to make themselves more tantalizing. In the beginning of the 20th century it was reclaimed and revived, however, mainly through the efforts of Kalkaprasad Maharaj, whose three sons Achchan, Lachchu, and Shambhu, perfected the art.

Because of its mixed *lāsya* and *tāṇḍava* temperament, *kathak* is popular with both females and males. In *bhārata-nāṭyam*, footwork is synchronized with hand gestures and eye movements, but *kathak* has no such rigid technique. It takes its movements from life, stylizes them, and adds complex rhythmic patterns. The mathematical precision in doubling and quadrupling the beat with quick transfers and shifts makes the onlookers dizzy.

Kathak performers

A female *kathak* dancer generally wears a brocade blouse, a long, wide, shimmering silk skirt, a transparent tissue scarf of gold threads, and a heavy cluster of ankle bells. A musician, generally the guru, sits beside the drummer on the floor and vocalizes the complicated syllables of the drum that the dancer beats out with her feet. *Kathak*'s basic dance posture and some of the steps can be traced to the *rāslīlā* of Braj Bhoomi. The musical refrain, called *lehra*, provides the base on which the drummer and the dancer execute a rich tapestry of rhythmic patterns. Beats are called *mātrā*s and the footwork *tatkar*. Important elements of the dance are *chakkar*s, *torah*s, and *tihai*s. *Chakkar* denotes whirling with great speed and stopping for a fraction of time after each whirl within the prescribed beat and maintaining the beauty of the form. *Torah* is a composition of rhythmic syllables. *Tihai* is the repetition of a phrase of rhythmic syllables used to adorn the concluding part of a *torah*. There are two styles of *kathak:* Jaipur *gharana* and Lucknow *gharana*. While the Lucknow *gharana* excels in *bhāva*, the Jaipur *gharana* specializes in brilliance of footwork.

In the 20th century the major performers of *kathak* include Shambhu Maharaj, who specialized in *bhavapradarśan* ("display of emotion"), and Sunder Prasad, who concentrated on the *tala* and *layakari* aspects of the dance. Birju Maharaj, Gopi Krishan, Sitara Devi, and Damayanti Joshi all have important reputations in India as well as abroad.

The manipuri school. *Manipuri* has survived in the sheltered valley of Manipur in the Assam Hills. It kept aloof not only from foreign influences but also from the main Indian trends. Its isolation was broken only in the 1920s, when Rabindranath Tagore visited the valley and invited a leading guru, Atomba Singh, to teach at his school in Santiniketan. The supple movements of *manipuri* dance were suitable for Tagore's lyrical dramas, and he employed them in his plays and introduced the dance as a part of the curriculum at his institution.

The *manipuri* dancer wears a large, stiff skirt glittering with round mirror pieces and a shimmering gauze veil. Her hair is done up in a high rolled crown with chains of white blossoms, and her luminous cheeks and forehead are decorated with dots of sandalwood paste.

Style of the *manipuri* dance

Known for its femininity, *manipuri* is marked by a slow, swooning rhythm. The dancer, with her hips thrust back and head tilted on one side, turns and sways and glides as if in a dream. The immobility of her face, like that of a mask, is in sharp contrast with the other three schools of dance, in which the face and eyes are a major source of expression.

The *manipuri* drummer, his naked torso in a white dhoti with a red border tucked up above his knees, dances while he plays on the drum. He slaps and thumps; the drum rumbles and howls and chuckles. Drunk with its rhythm, the drummer dances in wild, frenzied leaps. His energetic and electric movements are a masculine counterpart to the slow, undulating patterns woven by the female dancer.

Chief 20th-century exponents of *manipuri* include Atomba Singh, who preserved the tradition of *rās* dancing, and Amubi Singh.

The kuchipudi school. *Kuchipudi* dance-dramas owe their origin to the small village of Kuchipudi (Kuchelapuram) in Andhra Pradesh. Their form was originated in the 17th century by Sidhyendra Yogi, creator of the superb dance-drama *Bhama Kalapam*, which is the story of charming Satyabhāma, jealous wife of Lord Kṣṛṇa. Sidhyendra Yogi taught the art to Brahmin boys of Kuchipudi and gave a performance with them in 1675 for the Nawab of Golconda, who was so pleased that he granted Kuchipudi to the Brahmin Bhavathas for the preservation of this art. Even in the 20th century every Brahmin of Kuchipudi is expected to perform at least once in his life the role of Satyabhāma as an offering to Lord Kṛṣṇa.

The *kuchipudi* dance begins with worship rituals. A male dancer moves about sprinkling holy water, and then incense is burned. *Indra-dhvaja* (the flagstaff of the god Indra) is planted on the stage to guard the performance against outside interference. Women sing and dance with worship lamps, followed by the worship of Gaṇeśa, the elephant god, who is traditionally petitioned for success before all enterprises. The *bhagavatha* (stage manager-singer) sings invocations to the goddesses Sarasvatī (Learning), Lakṣmī (Wealth), and Paraśakti (Parent Energy), in between chanting drum syllables.

Two men hold up the traditional coloured curtain. A long gold-embroidered braid is hung on the curtain as a challenge to anyone among the spectators who dares to act and dance. If anyone takes up this braid, the hero playing the female character Satyabhāma will cut off "her" hair. The principal characters are introduced from behind the curtain after each one has done a brisk dance while the *bhagavatha* sings out the background and function of each. All roles are traditionally played by men (now by women also), and all the four elements of *abhinaya* are used—dance, song, costume, and psychological resources. Thus, *kuchipudi* differs from other classical dances in which the performers do not sing.

Among the major *kuchipudi* dancers of the 20th century are Guru Chinta Krishnamurthi, Vedantam Satyanarayana, and Yamini Krishnamurthi.

Odissi. *Odissi*, practiced in Orissa, claims to be over 2,000 years old and the true inheritor of the *Nāṭya-śāstra* tradition. It originated and was initially developed in the temples and later flourished in the courts as well. Many of the 108 basic dance units (*karaṇas*) mentioned in the *Nāṭya-śāstra* can be found only in *odissi*, and many of its dance poses are sculpted on the exterior of the temples of Bhubaneswar, Konārak, and Purī. Kelu Charan Mahapatra and Indrani Rehman are the principal 20th-century figures associated with *odissi*.

Bhagavatha mela, mohini attam, and *kuṟavañci*

Other classical dance forms. Among other classical or semiclassical dance forms are *bhagavatha mela, mohini attam,* and *kuṟavañci*. Performed at the annual Narasimha Jayanti festival in Melatur village in Tamil Nadu, the *bhagavatha mela* uses classical gesture language with densely textured Karnatak music. Its repertoire was enriched by the musician-poet Venkatarama Sastri (1759–1847), who composed important dance-dramas in the Telugu language. *Mohini attam* is based on the legend of the Hindu mythological seductress Mohinī, who tempted Śiva. It is patterned on *bhārata-nāṭyam* with elements of *kathākali*. It uses Malayalam songs with Karnatak music. *Kuṟavañci* is a dance-drama of lyrical beauty prevalent in Tamil Nadu. It is performed by four to eight women, with a gypsy fortune-teller as initiator of the story of a lady pining for her lover. Formally, it is a mixture of the folk and classical types of Indian dance.

Folk dance. Indian folk dances have an inexhaustible variety of forms and rhythms. They differ according to region, occupation, and caste. The half-naked Adivasis (aboriginal tribes) of central and eastern India (Murias, Bhīls, Gonds, Juangs, and Santāls) are the most uninhibited in their dancing. There is hardly a national fair or



Indian folk dance.
(Left) Chhau dance of Bihār showing boatman and wife. (Centre) *Kacchi ghori* dancers of Rājasthān. (Top right) *Ghoomar* dancers of Rājasthān. (Bottom) *Garabā* dancers of Gujarāt.
(Left, bottom) Mohan Khokar, (centre, top right) Foto Features

festival where these dances are not performed. The most impressive occasion occurs every January 26 on Republic Day, when dancers from all parts of India come to New Delhi to dance in the vast arena of the National Stadium and along a five-mile parade route.

Types of folk dances

It is difficult to categorize Indian folk dances, but generally they fall into four groups: social (concerned with such labours as tilling, sowing, fishing, and hunting); religious; ritualistic (to propitiate an angry goddess or demon with magical rites); masked (appears in all the above categories).

The *kolyacha* is among the better known examples of social folk dance. A fisherman's dance indigenous to the Konkan coast of western central India, the *kolyacha* is an enactment of the rowing of a boat. Women wave handkerchiefs to their male partners, who move with sliding steps. For wedding parties young Kolis dance in the streets carrying household utensils for the newlywed couple, who join the dance at its climax.

The national social folk dance of Rājasthān is the *ghoomar*, danced by women in long full skirts and colourful *chuneri*s (squares of cloth draping head and shoulders and tucked in front at the waist). Especially spectacular are the *kacchi ghori* dancers of this region. Equipped with shields and long swords, the upper part of their bodies each arrayed in the traditional attire of a bridegroom and the lower part concealed by a brilliant-coloured papier-mâché horse built up on a bamboo frame, they enact jousting contests at marriages and festivals. Bawaris, by tradition a criminal tribe, generally are expert in this form of folk dance.

In the Punjab, the most electrifying social folk dance is the male harvest dance, *bhangra*, which is also popular in the Punjab province of Pakistan. This dance is always punctuated by a song. At the end of every line the drum thunders. The last line is taken up by all the dancers in a chorus. In ecstasy they spring, bellow, shout, and gallop in a circle, madly wiggling their shoulders and hips. Any man of any age can join.

The Lambadi Gypsy women of Andhra Pradesh wear mirror-speckled headdresses and skirts and cover their arms with broad, white bone bracelets. They dance in slow, swaying movements, with men acting as singers and drummers. Their social dance is imbued with impassioned grace and lyricism and is less wild than that of Gypsies in other parts of the world.

The bison-horn dance of the Muria tribe in Madhya Pradesh is done by both men and women, who live on equal terms. The men wear a horned headdress with a tall tuft of feathers and a fringe of cowry shells dangling over their faces. A drum shaped like a log is slung around their necks. The women, their heads surmounted by broad, solid-brass chaplets and their breasts covered with heavy metal necklaces, carry sticks in their right hands like drum majorettes. Fifty to 100 dance at a time. The "bisons" attack and fight each other, spearing up leaves with their horns and chasing the female dancers in a dynamic interpretation of nature's mating season.

The Juang tribe in Orissa performs bird and animal dances with vivid miming and powerful muscular agility.

Religious folk dance

Some major examples of religious folk dances are the *dindi* and *kala* dances of Mahārāshtra, which are expressions of religious ecstasy. The dancers revolve in a circle, beating short sticks (*dindi*s) to keep time with the chorus leader and a drummer in the middle. As the rhythm accelerates, the dancers form into two rows, stamp their right feet, bow, and advance with their left feet, making geometric formations. The *kala* dance features a pot symbolizing fecundity. A group of dancers forms a double-tiered circle with other dancers on their shoulders. On top of this tier a man breaks the pot and splashes curds over the naked torsos of the dancers. After this ceremonial opening, the dancers twirl sticks and swords in a feverish battle dance.

Garabā, meaning a votive pot, is the best known religious dance of Gujarāt. It is danced by a group of 50 to 100 women every year for nine nights in honour of the goddess Ambā Mātā, known in other parts of India as Durgā or Kālī. The women move in a circle bending, turning, clapping their hands, and sometimes snapping their fingers. Songs in praise of the goddess accompany this dance.

Of the endless variety of ritualistic folk dances, many have magical significance and are connected with ancient cults. The *karakam* dance of Tamil Nadu state, mainly performed on the annual festival in front of the image of Māriyammai (goddess of pestilence), is to deter her from unleashing an epidemic. Tumbling and leaping, the dancer retains on his head without touching it a pot of uncooked rice surmounted by a tall bamboo frame. People ascribe this feat to the spirit of the deity, which, it is believed, enters his body. The Therayattam festival in Kerala is held to propitiate the gods and demons recognized by the pantheon of the Malayalis. The dancers, arrayed in awe-inspiring costumes and hideous masks, enact weird rituals before the village shrine. A devotee makes an offering of a cock. The dancer grabs it, chops off its head in one stroke, gives a blessing, and hands the bloody gift back to the devotee. This ceremony is punctuated by a prolonged and ponderous dance.

Masked folk dances

The greatest number of masked folk dances are found in Arunachal Pradesh (formerly North East Frontier Agency) union territory of India, where the influence of Tibetan dance may be seen. The yak dance is performed in the Ladākh section of Kashmir and in the southern fringes of the Himalayas near Assam. The dancer impersonating a yak dances with a man mounted on his back. In *sada topo tsen* men wear gorgeous silks, brocades, and long tunics with wide flapping sleeves. Skulls arranged as a diadem are a prominent feature of their grotesquely grinning wooden masks representing spirits of the other world. The dancers rely on powerful, rather slow, twirling movements with hops. The *chhau*, a unique form of masked dance, is preserved by the royal family of the former state of Saraikela in Bihār. The dancer impersonates a god, animal, bird, hunter, rainbow, night, or flower. He acts out a short theme and performs a series of vignettes at the annual Chaitra Parva festival in April. *Chhau* masks have predominantly human features slightly modified to suggest what they are portraying. With serene expressions painted in simple, flat colours, they differ radically from the elaborate facial makeup of *kathākali* or the exaggerated ghoulishness of the Nō and Kandyan masks. His face being expressionless, the *chhau* dancer's body communicates the total emotional and psychological tensions of a character. His feet have a gesture language; his toes are agile, functional, and expressive, like those of an animal. The dancer is mute; no song is sung. Only instrumental music accompanies him. In another form of *chhau*, practiced in the Mayūrbhanj district of Orissa, the actors do not wear masks, but through deliberately stiff and immobile faces they give the illusion of a mask. The style of their dance is vigorous and acrobatic.

Modern Indian dance. While in the West the theatrical elements of spoken words, music, and dance developed independently and evolved in the forms of drama, opera, and ballet, Indian theatrical tradition continued to combine the three in its dramas. Indian films still follow this rule (the heroine suddenly bursts into a song or dances for the hero), which offends Western sensibility, but in fact they are following their own classical and folk tradition. Recently, dance in the form of ballet with complex choreography in the Western sense has emerged as a distinct form.

Modern ballet

Modern Indian ballet started with Uday Shankar (1900–), who went to England to study the plastic arts and was chosen by the Russian ballerina Anna Pavlova to be her partner in the ballet *Radha and Krishna*. Young Shankar returned to India fired with enthusiasm. After studying the essentials of the four major styles of classical dance, he created new ballets with complex choreography and music, mixing the sounds from wooden clappers and metal cymbals with those of traditional instruments. He used classical and folk rhythms. Employing Western stage techniques, he presented his ballets with a skill and polish previously unknown to Indian audiences. These ballets included *Shiva-Parvati* and *Lanka Dahan* ("The Burning of Lanka"), in which he used wooden masks from Ceylon. In *Rhythm of Life* (1938) and in *Labour and Machinery* (1939), he employed contemporary political and social themes. He established a culture centre at Almora in 1939 and during its five years' existence created a whole generation of modern dancers.

Shanti Bardhan, a junior colleague of Uday Shankar, produced some of the most imaginative dance-dramas of the modern period. After founding the Little Ballet Troupe in Andheri, Bombay, in 1952 he produced *Ramayana*, in which the actors moved and danced like puppets. His posthumous production *Panchatantra* (*The Winning of Friends*) is based on an ancient fable of four friends (Mouse, Turtle, Deer, and Crow), in which he used masks and the mimed movements of animals and birds.

Mohan Khokar



"Ramayana," puppet-style modern dance drama originally produced and choreographed by Shanti Bardhan, *c.* 1950.

Narendra Sharma and Sachin Shankar, both pupils of Uday Shankar, have continued his tradition. Other important figures who have shaped modern Indian dance include Menaka, Ram Gopal, and Mrinalini Sarabhai, who has experimented with conveying modern themes through the *bhārata-nāṭyam* and *kathākali* styles.

Dance-training centres. Dance training in small academies and local *kala kendras* ("art centres") is available all over contemporary India. Most universities have introduced dance as a subject in their curricula. The gurus still impart specialized training to pupils who go to live with them in villages and learn the art over a number of years. But there are many state-run or public-financed training centres organized in the 20th century that attract students from all over the world. Among the most important of these are Kerala Kalamandalam (Kerala Institute of Arts), near Shoranūr; Kalakshetra at Adyar, Tamil Nadu; Kathak Kendra, a dance branch of the Bharatiya Kala Kendra in New Delhi; Triveni Kala Sangam (Centre of Music, Dance, and Painting), at New Delhi; Darpana Academy in Ahmadābād, Gujarāt; Visva-Bharati (founded by Rabindranath Tagore), at Santiniketan, West Bengal; and the Jawaharlal Nehru Manipuri Dance Academy, at Imphāl.

INDIAN THEATRE

Classical theatre. Classical Sanskrit theatre flourished during the first nine centuries of the Christian era. Aphorisms on acting appear in the writings of Pāṇini, the Sanskrit grammarian of the 5th century BC, and references to actors, dancers, mummers, theatrical companies, and academies are found in Kauṭilya's book on statesmanship, the *Artha-śāstra* (4th century BC). But classical structure, form, and style of acting and production with aesthetic rules were consolidated in Bharata Muni's treatise on dramaturgy, *Nāṭya-śāstra*. Bharata defines drama as a

Bharata Muni's rules

> mimicry of the actions and conduct of people, rich in various emotions, depicting different situations. This relates to actions of men good, bad and indifferent and gives courage, amusement, happiness, and advice to all of them.

Bharata classified drama into ten types. The two most important are *nāṭaka* ("heroic"), which deals with the exalted themes of gods and kings and draws from history or mythology (Kālidāsa's *Śakuntalā* and Bhavabhūti's *Uttararāmacarita* fall into this caterory), and *prakaraṇa* ("social"), in which the dramatist invents a plot dealing with ordinary human beings, such as a courtesan or a woman of low morals (Śūdraka's *Mṛcchakaṭika*, "The Little Clay Cart," belongs to this type). Plays range from one to ten acts. There are many types of one-act plays, including *bhaṇa* ("monologue"), in which a single character carries on a dialogue with an invisible one, and *prahasana* ("farce"), which is classified into two categories: superior and inferior, both dealing with courtesans and crooks. King Mahendravikramavarman's 7th-century-AD *Bhagavad-Ajjukiya* ("The Harlot and the Monk") and *Mattavilāsa* ("Drunken Revelry") are examples of *prahasana*.

Types of classical theatre

There are three structural types of classical theatre: oblong, square, and triangular, each further divided into large, medium, and small sizes. According to the *Nāṭya-śāstra*, the playhouse was "like a mountain cave" with two floors at different levels, small windows so that outside noise and wind would not interfere with the acoustics, and a backstage for actors to do makeup, costumes, and offstage noise effects. Bharata disapproved of a large playhouse and recommended the medium-size structure meant for court productions.

The ancient Hindus insisted on a small playhouse, because dramas were acted in a highly stylized gesture language with subtle movements of eyes and hands. Hindu theatre differed from its Greek counterpart in temperament and method of production. The three unities rigidly followed by the Greeks were totally unknown to Sanskrit dramatists. Less time was consumed by a Greek program of three tragedies and a farce than by a single Sanskrit drama, with its subsidiary plots and wide variety of characters and moods. The Greeks laid emphasis on plot and speech, the Hindus on the four types of acting and visual demonstration. People were audiences to the Greeks and spectators to the Hindus. The aesthetic rules also differed. Aristotle's theory of catharsis bears no resemblance to Bharata's theory of *rasa*. The Greek conception of tragedy is totally absent in Sanskrit dramas, as is the aesthetic principle that prohibits any death or defeat of the hero on stage.

Kinds of Hindu production

There were two types of Hindu productions: the *lokadharmī*, or realistic theatre, with natural presentation of human behaviour and properties catering to the popular taste, and the *nāṭyadharmī*, or stylized drama, which, using gesture language and symbols, was considered more artistic. In *Śakuntalā* the king enters riding an imaginary chariot, and Śakuntalā plucks flowers that are not there; in "The Little Clay Cart" the thief breaks through a nonexistent wall, and Maitreya passes through Vasantasena's seven courtyards by miming.

A classical play traditionally opened with the *nāndī*, a benediction of eight to 12 lines of verse in praise of the gods, after which the *sūtra-dhāra* (stage manager) entered with his wife and described the place and occasion of the action. The last sentence of his prologue served as a bridge leading to the action of the play. In *Śakuntalā* he refers to the bewitching song of his wife, which has made him forget his surroundings as the pursuit of a deer has made the king forget his state affairs. At this point the king enters, riding his hunting chariot, and the spectators are plunged into action of the play.

The *vidūṣaka* (clown) is a noble, good-hearted, blundering fool, the trusted friend of the hero. A bald-headed glutton, comic in speech and manners, he is the darling of the spectators. With the decline of Sanskrit drama the folk theatre in various regional languages inherited the conventions of the opening prayer song, the *sūtra-dhāra*, and the *vidūṣaka*.

The only surviving Sanskrit drama is *kudiyattam*, still performed by the Cakkayars of Kerala. Some principles of the *Nāṭya-śāstra* are evident in their presentations.

The earliest available classical dramas are thirteen plays edited in 1912 by Pandit Ganapati Sastri, who dug out their manuscripts in Trivandrum, the capital of Kerala state. These, ascribed to Bhāsa (1st century BC–1st century AD), include the one-act *Ūrubhaṅga* ("The Broken Thigh"), a tragedy that is a departure from Sanskrit convention, and the six-act *Svapnavāsavadatta* ("The Dream of Vāsavadattā").

The most acclaimed dramatist is Kālidāsa. Other important playwrights succeeding him include Harṣa, Mahendravikramavarman, Bhavabhūtī, and Viśakhādatta. An exception is King Śūdraka, whose work is perhaps the most theatrical in the entire Sanskrit range.

"The Little Clay Cart"

The title of "The Little Clay Cart" represents a departure from Sanskrit tradition, in which a *prakaraṇa* was generally named after its hero and heroine. *Mālavikāgnimitra*, for example, is the love story of Princess Mālavikā and King Agnimitra, *Vikramorvaśī* is the tale of King Purūravas and the heavenly nymph Urvaśī, and *Mālatī-Mādhava* is the love drama of Mālatī and Mādhava. Śūdraka, as if to mock tradition, chose an insignificant homely incident—the hero's son playing with a toy cart—and elevated this to the title.

"The Little Clay Cart" has a wide range of characters. The plot does not progress in a straight line but zigzags along a winding path. During its ten acts the hero does not appear in four of them, the heroine is absent from three, and the lustful villain disappears after the first act until the eighth. Each act is an almost independent play. The device used to link the acts is that of ending them with subtitles that sum up their particular themes or plots.

"The Little Clay Cart" has been successful in the West, whereas Indian audiences, still fed on poetic-flavoured characters and romances of an ethereal type, have favoured *Śakuntalā*. Western audiences find "The Little Clay Cart" more in their own tradition of realism and individualized characterization. Its "lisping villain," gamblers, and rogues have something in common with Shakespeare's comic characters and Molière's crooks. "The Little Clay Cart" is better theatre, whereas *Śakuntalā* is better poetry.

Folk theatre. After the decline of Sanskrit drama, folk theatre developed in various regional languages from the 14th through the 19th centuries. Some conventions and stock characters of classical drama (stage preliminaries, the opening prayer song, the *sūtra-dhāra*, and the *vidūṣaka*) were adopted into folk theatre, which lavishly employs music, dance, drumming, exaggerated makeup, masks, and a singing chorus. Thematically it deals with mythological heroes, medieval romances, and social and political events, and it is a rich store of customs, beliefs, legends, and rituals. It is a "total theatre," invading all the senses of the spectators.

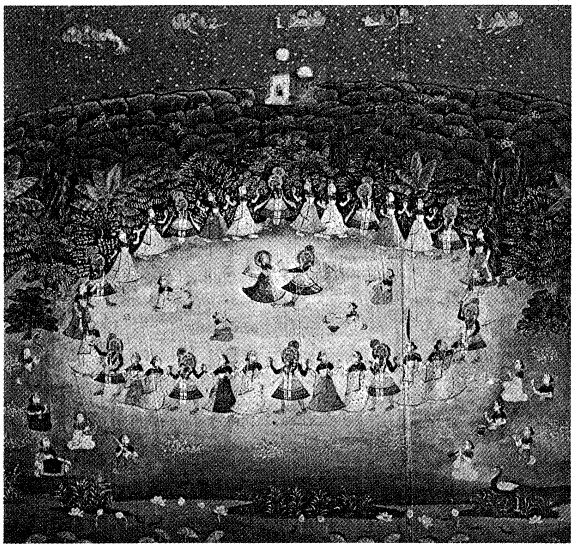
The most crystalized forms are the *jātrā* of Bengal, the *nautanki*, *rāmlīlā*, and *rāslīlā* of North India, the *bhavai* of Gujarāt, the *tamāshā* of Mahārāshtra, the *terukkūttu* of Tamil Nadu, and the *yakṣagāna* of Kanara.

Places and kinds of folk theatre

Folk theatre is performed in the open on a variety of arena stages; round, square, rectangular, multiple-set. The *bhavai*, enacted on a ground-level circle, and the *jātrā*, on a 16-foot (five-metre) square platform, have gangways that run through the surrounding audience and connect the stage to the dressing room. Actors enter and exit through these gangways, which serve a function similar to the *hanamichi* of the Japanese Kabuki theatre. In the *rāmlīlā*, the action sometimes occurs simultaneously at various levels on a multiple set. Actors in *nautanki* and *bhavai* sit on the stage in full view instead of exiting and sing or play an instrument as a part of the chorus. In the *rāmlīlā*, the actor playing Rāvaṇa removes his ten-headed mask when he is not acting and continues sitting on his throne, but for the spectators he is theatrically absent. Asides, soliloquies, and monologues abound. Scenes melt into one another, and the action continues in spite of change of locale.

In most folk forms the art of the actor is hereditary. He learns by watching his elders throughout childhood. He starts with drumming, then dancing, plays female roles, and then major roles.

All roles are played by men except that of the *tamāshā*



Rāslīlā folk drama of northern India, watercolour, late 19th century. In the collection of Mohan Khokar, New Delhi. 1.3 × 1.3 m.
Mohan Khokar

woman, who is always a dancer-singer-actress. Recently, women have started playing female roles in the *jātrā* but have failed to achieve the artistic stature of their professional male counterparts.

In the *rāmlīlā* and *rāslīlā*, the principal characters—Rāma and Kṛṣṇa—are always played by boys under 14, because tradition decreed they must be pure and innocent. They are considered representatives of the gods and are worshipped on these occasions. In the *rāmlīlā* the *vyas* ("director"), present on the stage throughout the performance, prompts and directs the characters loudly enough for the audience to hear. This is not regarded as disturbing because it is an accepted part of the tradition. Adult roles such as Rāvaṇa and Hanuman are sometimes played by the same individual throughout his life.

Non-religious folk dance

Of the nonreligious forms, the *jātrā* and the *tamāshā* are most important. The *jātrā*, also popular in Orissa and eastern Bihār, originated in Bengal in the 15th century as a result of the *bhakti* movement, in which devotees of Kṛṣṇa went singing and dancing in processions and in their frenzied singing sometimes went into acting trances. This singing with dramatic elements gradually came to be known as *jātrā*, which means "to go in a procession." In the 19th century the *jātrā* became secularized when the repertoire swelled with love stories and social and political themes. Until the beginning of the 20th century, the dialogue was primarily sung. The length has been cut from all night to four hours. The *jātrā* performance consists of action-packed dialogue with only about six songs. The singing chorus is represented by a single character, the *vivek* ("conscience"), who can appear at any moment in the play. He comments on the action, philosophizes, warns of impending dangers, and plays the double of everybody. Through his songs he externalizes the inner feelings of the characters and reveals the inner meaning of their outer actions.

The *tamāshā* (a Persian word meaning "fun," "play," or "spectacle") originated at the beginning of the 18th century in Mahārāshtra as an entertainment for the camping Mughal armies. This theatrical form was created by singing girls and dancers imported from North India and the local acrobats and tumblers of the lower-caste Dombari and Kolhati communities with their traditional manner of singing. It flourished in the courts of Marāthā rulers of the 18th and 19th centuries and attained its artistic apogee during the reign of Bājī Rāo II (1796–1818). Its uninhibited *lavani*-style singing and powerful drumming and dancing give it an erotic flavor. The most famous *tamāshā* poet and performer was Ram Joshi (1762–1812) of Sholāpur, an upper class Brahmin who married the courtesan Bayabai. Another famous singer-poet was Patthe Bapu Rao (1868–1941), a Brahmin who married a beautiful low-caste dancer, Pawala. They were the biggest *tamāshā* stars during the first quarter of the 20th century. The *tamāshā* actress, commonly called the *nautchi* (meaning "nautch girl," or "prostitute") is the life and soul of the performance. Because of their bawdy elements, women never see *tamāshā* plays, nor do respectable men.

In the 20th century, *jātrā* and *tamāshā* both have become highly organized and are commercially run. Troupes are in heavy demand and work for nine months. Over 700 *tamāshā* troupes with 2,000 dancer-actresses tour the rural areas, providing a living for about 40,000 people. The *jātrā* is the most successful commercially. Its star actors draw more than any other professional actor in the theatrical centre of Calcutta.

Puppet theatre of North India

Popular in North India are the *putliwalas* ("puppeteers") of Rājasthān, who operate marionettes made of wood and bright-coloured cloth. The puppet plays deal with kings, lovers, bandits, and princesses of the Mughal period. Generally, the puppeteer and his nephew or son operate the strings from behind, while the puppeteer's wife sits on her haunches in front of the miniature stage playing the drums and commenting on the action. The puppeteer chirps, whimpers, and squeals in animal–bird voices and creates battles and tragic moments, expresses pathos, anger, and laughter. In Andhra Pradesh the puppets, called *tholu bommalata* ("the dance of leather dolls"), are fashioned of translucent, coloured leather. These are projected on a small screen, like colour photographic transparencies. Animals, birds, gods, and demons dominate the screen. The puppeteer manipulates them from behind with two sticks. Strong lamps are arranged so that the size, position, and angle of the puppets change with the distance of the light. They are similar to the *wayang kulit* puppets of Indonesia but are much smaller and quicker moving.

In the absence of a powerful Indian city theatre (with the exception of a few in Calcutta, Bombay, and Tamil Nadu), folk theatre has kept the rural audiences entertained for centuries and has played an important part in the growth of modern theatres in different languages. The 19th-century dramatist Bharatendu Harishchandra, who was responsible for the birth of Hindi drama, used folk conventions—the opening prayer song, tableaux, comic interludes, duets, stylized speech—and combined these with Western theatrical forms in vogue at that time. Parsi companies adapted the popular folk techniques for their extravaganzas and were a major influence until the 1930s. Rabindranath Tagore, rejecting the heavy sets and realistic decor of the commercial companies, created a lyrical theatre of the imagination. Much influenced by the *baul* singers and folk actors of Bengal, he introduced the Singing Bairagi and the Wandering Poet (similar to the *vivek* of the *jātrā*) in his dramas. In the late 20th century, folk theatre has been viewed as a form that can add colour and vitality to contemporary theatre.

Modern theatre. Modern Indian theatre first developed in Bengal at the beginning of the 18th century as a result of Western influence. The other regional theatres more or less followed Bengal's pattern, and within the next 100 years they took the same meandering path, though they never achieved the same robust growth.

British influence on Bengali theatre

The British conquered Bengal in 1757 and influenced local arts by their educational and political systems. Their clubs performed Shakespeare, Molière, and Restoration comedies, introducing Western dramatic structure and the proscenium stage to the Indian intelligentsia. With the help of Golak Nath Dass, a local linguist, Gerasim Lebedev, a Russian bandmaster in a British military unit, produced the first Bengali play, *Chhadmabes* ("The Disguise"), in 1795 on a Western-style stage with Bengali players of both sexes. Subsequently, Bengali playwrights began synthesizing Western styles with their own folk and Sanskrit heritage. With growing national consciousness, theatre became a platform for social reform and propaganda against British rule. Among the most important playwrights were Michael Madhu Sudan (1824–73), Dina Bandhu Mitra (1843–87), Girish Chandra Ghosh (1844–1912), and D.L. Roy (1863–1913).

The success of Dina Bandhu Mitra's *Nildarpan* ("Mirror of the Indigo"), dealing with the tyranny of the British indigo planters over the rural Bengali farm labourers, paved the way for professional theatre. The actor-director-writer Girish Chandra Ghosh founded in 1872 the National Theatre, the first Bengali professional company, and took *Nildarpan* on tour, giving performances in the North Indian cities of Delhi and Lucknow. The instigatory speeches and lurid scenes of British brutality resulted in the banning of this production. To overcome censorship difficulties, playwrights turned to historical and mythological themes with veiled symbolism that was clearly understood by Indian audiences. The heroes and villains of these plays came to represent the Indian freedom fighter against the British oppressor. Girish's historical tragedies *Mir Qasim* (1906), *Chhatrapati* (1907), and *Sirajuddaulah* (1909) bring out the tragic grandeur of heroes who fail because of some inner weakness or betrayal of their colleagues. D.L. Roy emphasized the same aspect of nationalism in his historical dramas *Mebarapatan* (The Fall of Mebar), *Shahjahan* (1910), and *Chandragupta* (1911).

Girish introduced professional efficiency and showmanship. His style of acting was flamboyant, with fiery grace. Actors such as Amar Datta and Dani Babu carried his style into the early 1920s. The acting and production methods of the Star, the Minerva, and the Manmohan Theatres (all professional) were modelled on Girish's pioneer work.

The beginning of realism

The first elements of realism were introduced in the 1920s by Sisir Kumar Bhaduri, Naresh Mitra, Ahindra Chowdhuri, and Durga Das Banerji, together with the actresses Probha Devi and Kanka Vati. In his Srirangam Theatre (closed in 1954), Sisir performed two most memorable roles: the aging Mughal emperor Aurangzeb and the shrewd Hindu philosopher-politician Cānakya. Sisir's style has been refined by actor-director Sombhu Mitra and his actress wife Tripti, who worked in the Left-wing People's Theatre movement in the 1940s. With other actors they founded the Bahurupee group in 1949 and produced many Tagore plays including *Rakta Karabi* ("Red Oleanders") and *Bisarjan* ("Sacrifice"), so far unattempted by any professional company.

Rabindranath Tagore (1861–1941), steeped in Hindu classics and indigenous folk forms but responsive to European techniques of production, evolved a dramatic form quite different from those of his contemporaries. He directed and acted in his plays along with his cousins, nephews, and students. These productions were staged mostly at his school, Santiniketan, in Bengal as a nonprofessional and experimental theatre. The Calcutta elite and foreign visitors were attracted to these performances.

A painter, musician, actor, and poet, Tagore combined these talents in his productions. He used music and dance as essential elements in his latter years and created the novel opera-dance form in which a chorus sat on the stage and sang while the players acted out their roles in dance and stylized movements. Sometimes Tagore himself sat on a stool acting as the *sūtra-dhāra* and chanted to the accompaniment of music and drum as the dancing players became visual moving pictures.

In northern and western India, theatre developed in the latter half of the 19th century. The Bombay Parsi companies, using Hindi and Urdu, toured all over India. Their spectacular showmanship, based on a dramatic structure of five acts with songs, dances, comic scenes, and declamatory acting, was copied by regional theatres. The Maharashtrian theatre, founded in 1843 by Visnudas Bhave, a singer-composer-wood-carver in the court of the Raja of Sāngli, was developed by powerful dramatists such as Khadilkar and Gadkari, who emphasized Marāthā nationalism. The acting style in Maharashtrian theatre remained melodramatic, passionately arousing audiences to laughter or tears.

In the south, the popularity of dance-dramas has not allowed theatrical realism to flourish. Tamil commercial companies with their song and dance extravaganzas have dominated Andhra Pradesh, Kerala, and Mysore. The most outstanding Tamil company since the independence of India in 1947 has been the T.K.S. Brothers of Madras, famous for trick scenes and gorgeous settings. Also famous is the actor-producer-proprietor Rajamanickam, who specializes in mythological plays with an all-male cast, using horses, chariots, processions, replicas of temples, and even elephants.

Urdu and Hindi drama

Urdu and Hindi drama began with the production of *Indrasabha* by Nawab Wajid Ali Shah in 1855 and was developed by the Parsi theatrical companies until the 1930s.

Parsi theatre was an amalgam of European techniques and local classical forms, folk dramas, farces, and pageants. Mythical titans thundered on the stage. Devils soared in the air, daggers flew, thrones moved, and heroes jumped from high palace walls. Vampire pits, the painted back cloth of a generalized scene, and mechanical devices to operate flying figures were direct copies of the 19th-century Lyceum melodramas and Drury Lane spectacles in London.

The star film actor Prithvi Raj Kapoor founded Prithvi Theatres in Bombay in 1944 and brought robust realism to Hindi drama, then closed down in 1960 with a sense of completion after many tours throughout India. Prithvi's sons, nephews, and old associates worked in his large company, which became a training centre for many actors who later joined the films. Among these was the outstanding stage actress Zohra Sehgal, a former dance partner of Uday Shankar in the 1930s who had tremendous emotional depth and range, rare in actresses on the Hindi stage. Out of Prithvi's eight productions, in which he always played the lead, the best was *Pathan* (1946), which ran for 558 nights. It deals with the friendship between a tribal Muslim khān and a Hindu dewan and is set in the rugged frontier from which Prithvi came. This tragedy of two archtypes in which the khān sacrifices his son to save the life of his friend's son had intensity of action, smoldering passion, and unity of mood and achieved the highest quality of realism on the Hindi stage to this day.

Among the actors who molded regional-language theatres are Shri Narayan Rao Rajhans (popularly known as the Bala Gandharva of the Mahārāshtra stage), Jayashankar Bhojak Sundari of Gujarāt, and Sthanam Narasimhrao of Andhra. All three specialized in female roles and were star attractions during the first quarter of the 20th century.

In the last half of the 20th century, two outstanding actor-directors are Ebrahim Alkazi, director of the National School of Drama in New Delhi, and Utpal Dutt, who founded the Calcutta Little Theatre Group in 1947, which originally performed plays in English and in 1954 changed to productions in Bengali. Dutt is an actor fully committed to the revolutionary ideology of the Chinese Communist leader Mao Tse-tung. He acts on open-air stages in rural areas of Bengal, where he exerts a strong artistic and political influence.

Since Lebedev in 1795 there has been a continuous stream of Western-trained actors and producers who have been revitalizing regional-language theatrical groups. Nawab Wajid Ali Shah had visiting French opera composers in his mid-19th-century court. Tagore did his first opera, *Valmiki Pratibha* ("The Genius of Vālmīkī"), in 1881, after returning from England, where he became familiar with Western harmonies. Prithvi Raj Kapoor, E. Alkazi, and Utpal Dutt all had their earlier training in English productions. Norah Richards, an Irish-born actress who came to the Punjab in 1912, produced in 1914 the first Punjabi play, *Dulhan* ("The Bride"), written by her pupil I.C. Nanda. For 50 years she promoted rural drama and inspired actors and producers, including Prithvi Raj Kapoor.

India's genius still lies in its dance-dramas, which have a unique form based on centuries of unbroken tradition. There are very few professional theatre companies in the whole of India, but thousands of amateur productions are staged every year by organized groups. Out of this intense experimental activity, the Indians hope a contemporary national theatre will emerge, influenced by Western techniques but distinctly Indian in flavour.

Centres for training

Many centres for theatrical training have been established in the 20th century. Among the most important are the National School of Drama and the Asian Theatre Institute in New Delhi, Sangeet Natak Akademi (National Academy of Music, Dance, and Drama) in New Delhi, and the National Institute for the Performing Arts in Bombay. Bharatiya Natya Sangh, the union of all Indian theatre groups, was founded in 1949 and is centered in New Delhi. Affiliated with UNESCO's branch of the International Theatre Institute, it organizes drama festivals and seminars, as well as serving as a centre for information.

SRI LANKA (CEYLON)

The ritualistic dances of Sri Lanka have attained world fame for their weird mystical beauty. Literary drama has not flourished, because the monks of predominantly Buddhist Sri Lanka shunned theatre. Dramatic activity found expression in exorcism ceremonies and masked dramas that employed mime, song, dance, acrobatics, and bits of prose dialogue. Heavily influenced by India, Sri Lanka's Kandyan dance and *kōlam* plays have South Indian origins. But over the centuries these have been transformed and now have a distinctly Sri Lankan character.

It is difficult to divide Sri Lankan performing arts into dance and drama, because a *kōlam* play uses dance and song, and the devil dance has bits of improvised prose dialogue.

Dance. *Devil dance.* The Buddhists of Sri Lanka still believe in supernatural beings and the healing power of magical rites. Their devil dancing is the expression, however, of pre-Buddhist beliefs.

The devil dance is performed to cure a person gripped by disease, insanity, or bad luck believed to be caused by some malignant spirit and to propitiate demons and deities to bring good fortune. The dancers belong to a lower-caste community and are professional. During their performance the patient lies to one side. Several palm-leaf shrines are constructed outside his house, each dedicated to a particular demon to lure him into the arena. The role of the Vesamuni, king of all demons, is played by the chief exorcist.

Three types of supernatural beings have to be appeased: demons, deities, and others that are half demon, half deity. The most terrible is Riri Yakka (Demon of Blood), who inhabits cremation grounds and graveyards and rides a pig. His belly is smeared with blood, and he has a monkey's face and four clawed hands that hold a parrot, a sword, a rooster, and a human head.

Costumes of the devil dancers

The dancers, all men, each wear around their heads a red cloth fringed with long ribbons of palm leaves hanging down like female hair, a strip of cloth around their chests, 22 yards (20 metres) of thin white cloth wound so skillfully around their hips that it never comes loose during an entire night of violent activity, and clusters of bells fitted around their calves to make a deafening jingle. Their appearance is half female, half male.

The dance is punctuated by little pieces of mimes and magical actions, with naked-chested drummers pounding to the accompaniment of a chorus of singers. The climax is reached when the dancers, holding flaming torches in both hands, whirl and spin, forming circles of fire around themselves. The flames lick their bodies, but they remain unsinged. The dancers leap and dive through the air in seeming defiance of gravity. In this surcharged atmosphere they pause to put on masks representing various demons. These have terrifying expressions. They romp and stomp in circles, describing their identity and purpose of visiting. The particular demon associated with the malady enters the patient's body. The chief exorcist questions, threatens, tortures, beseeches, and offers bribes to appease the demon until it finally leaves and its victim is healed.

The *sanni yakku* dance, exorcising the disease demon, has a series of humorous impersonations. One is of the demon as a beautiful woman, then as a pregnant woman, and finally as a mother. The exorcists ask questions about her pregnancy, and she lists all the respectable men of the village.

Out of many devil-dance ceremonies, most picturesque and important are the *kohomba kankariya* (or "ritual of the god Kohomba"), performed to ensure prosperity and to get rid of pestilence, and the *bali*, danced to propitiate the heavenly beings.

Kandyan dance. The *kandyan* dance, shorn of occult ceremonies, is highly sophisticated and refined. It flourished under the Kandyan kings from the 16th through the 19th centuries, and today it is considered the national dance of Sri Lanka. It has four distinct varieties: The *pantheru, naiyadi, udekki,* and *ves* (the most artistic and renowned). Its energetic movements and postures are reminiscent of India's *kathākali.* Besides the above four styles, there are 18 *vannamas* (dance enactments) including the *gajaga vannama,* depicting the elephant; the *hanuma vannama,* the monkey; and the *mayura vannama,* the peacock. These beautiful animal movements and abstract impersonations have been distilled and perfected over several hundred years.

The *vannamas*

Hindu mythological themes were originally the subjects of *kandyan* dances, the most popular being Rāma's crossing over to Laṅkā with the help of his monkey general and his reunion with Sītā. Gradually, stories of kings and legendary heroes and mimes of birds and forest beasts were introduced. The Kandyan kings elevated the dance to such beauty and skill that the Buddhists began admitting it into their temple courtyards as a tribute to the glory of their religion. It became a part of the annual August Perahera festival, in which a procession of gilded elephants, palanquins, saffron-robed monks, drummers, and chanters move majestically to the Temple of the Tooth in Kandy, where Buddha's tooth is enshrined. The *kandyan* dancers are a glittering attraction as they perform enroute to the temple.

Ewing Krainin—Stockpile



Kandyan dancers and drummers from Sri Lanka.

A *kandyan* dancer wears a pagoda-like conical silver headpiece with glistening forehead fringe and huge earpieces, many-stranded bead necklaces of silver and ivory across his naked torso, beaten-silver epaulets on his biceps, and hollow silver anklets filled with silver beads to make them rattle. He spins with sudden leaps and reaches violent climaxes of geometric patterns. The sudden right and left turns of his head make the onlookers dizzy. When telling a story, he sings descriptive passages and enacts them with spurts of dancing.

Masked drama. Out of the four folk-drama forms—*kōlam, sokari, nadagam,* and *pasu*—the most highly developed and significant is the *kōlam,* in which actors wear brightly painted and intricately carved wooden masks. The word *kōlam* is of Tamil origin and means "costume," "impersonation," or "guise." The performance consists of the masked representation of many isolated characters,

The *kōlam*

such as kings, demons, deities, hunters, animals and birds, the washerman, the police constable, a pregnant woman —a British Museum manuscript concerning the *kōlam* lists 53 such characters. The most terrifying masks are of the demons, with twisted faces, protruding tusks, and cavelike nostrils for snorting fury. The *nāga* demon has a long, flaming, red tongue and dozens of cobras writhing around his head. Some old masks have only one large bulging eye, with a cobra hissing out from one nostril. The design of these masks lies between the strongly stylized tribal masks of Africa and the highly polished, sophisticated masks of the Japanese Nō Theatre. Five basic colours are used—red, blue, yellow, green, and black, the last two for lower rank characters. Exaggerated comicality, distortions, bulges, nightmarish whimsy, bright colours, and artful carving of the masks have been significant factors in keeping this form of drama alive.

The *kōlam* is performed once a year for seven to ten nights, starting at night after dinner and lasting through the early hours of the morning. The performance is generally held in the open courtyard of a house, to the accompaniment of two drummers, an instrumentalist, and a singing chorus with leader. After songs in praise of Lord Buddha and others (including the patron of the show), the *sabhapati* (master of ceremonies) describes the origin of *kōlam*—how an Indian king's pregnant wife expressed a desire to see a masked dance-drama and how a troupe was invited from a distant court. The *sabhapati* then introduces the masked characters as they enter and describes their various vocations and backgrounds.

Out of many, two plays are especially famous: the *Sandakinduru Katava* and the *Gothayimbala Katava*. The former deals with the legendary idyllic love between a half-human, half-bird couple singing and dancing in a forest. The King of Banaras comes hunting and, attracted by the beautiful Kinduri, kills her husband and makes advances to her. Rejected, he is ready to kill her when Lord Buddha appears and brings her husband back to life. In the *Gothayimbala Katava* the beautiful wife of the warrior Gothayimbala bathing in a pond attracts the attention of a demon, who falls in love with her. The enraged husband comes and chops off the demon's head, which, because of its magical power, reunites itself with the body every time it is cut off. Finally, the forest deity comes and rescues the warrior.

The recorded history of *kōlam* is not very old. There is only one known early eye witness account of *kōlam*, that of John Callaway, who in 1829 published 185 verses of a play with a description of the performance and some sketches of the masks and a brief introduction concerning the masquerade. According to Callaway, the dancers did not sing. The chanters described the characters in the third person and sometimes exclaimed to draw the attention of the audience to a particular action. The earliest *kōlam* text is preserved in the Colombo National Museum on palm leaves; another is in the British Museum inscribed on paper. The oldest printed text, edited in 1895 by A.G. Perera, is in the Colombo National Museum Library.

Masks are made of the light woods kaduru (Strychnos nux vomica) and ruk-attana (*Alstonia scholaris*) and after 50 years start decaying; consequently, the earlier masks are no longer in existence.

There has been an important revival of interest in drama in Sri Lanka since independence. E.R. Sarachchandara, a scholar of traditional Ceylonese theatre, has been responsible for a major breakthrough in revitalizing and adapting for the modern stage traditional dramatic forms such as the *kōlam*. Several new playwrights have become prominent in the mid-20th century. Foremost among them is Henry Jayasena. A producer-writer-actor, Jayasena has written and staged plays in Sinhalese and translations of foreign plays. But modern theatre is still very weak.

DANCE AND THEATRE IN KASHMIR

The Vale of Kashmir, predominantly populated by Muslims, has remained aloof from the main cultural currents of India. The ancient caves and temples of Kashmir, however, reveal a strong link with Indian culture at the beginning of the Christian Era. At one time the classical dances of the south are believed to have been practiced. When Islām was introduced, in the 14th century, dancing and theatrical arts were suppressed, being contrary to a strict interpretation of the Qur'ān. These arts survived only in folk forms and were performed principally at marriage ceremonies. The popular *hafiza* dance performed by Kashmiri women at weddings and festivals to the accompaniment of *sufiana kalam* (devotional music of the Muslim mystics known as Ṣūfīs) was banned in the 1920s by the ruling maharaja, who felt this dance was becoming too sensual. It was replaced by the *bacha nagma*, performed by young boys dressed like women. A popular entertainment at parties and festivals, it is also customarily included in modern stage plays.

In contrast to its natural scenic richness, Kashmir is theatrically a pauper. Theatrical productions are generally amateurish, since there is no regular performing company or any tradition of civic theatre.

Bhand jashna

There is only the *bhand jashna* ("festival of clowns"), a 300- to 400-year-old genre of Kashmiri folk theatre. Performed in village squares, it satirizes social situations through dance, music and clowning.

The Kashmiri-language theatre was founded in 1947, when a new national consciousness, the aftermath of the independence of the Indian subcontinent from Britain, inspired playwrights and folk actors to dramatize topical events and create a "visual newspaper" for the people. Left-wing propaganda plays such as *Zamin Sanz* ("Who Owns the Land?") and *Jangbaaz* ("The Warmonger"), though mediocre, had topical interest. Notable among those who tried their hand at writing for the stage is the poet Nadim, author of two operas, *Bambur-yambarzal* (*The Bumblebee*) and *Himal Nagraj* (*The Beautiful Woman and the Snake Prince*).

Since the 1960s, the Jammu and Kashmir Academy of Art, Culture, and Languages has been struggling to promote theatre in the Kashmiri and Dogri languages, but with little success. Its emphasis is on literary dramas and folk-dance festivals of regional appeal.

PAKISTAN

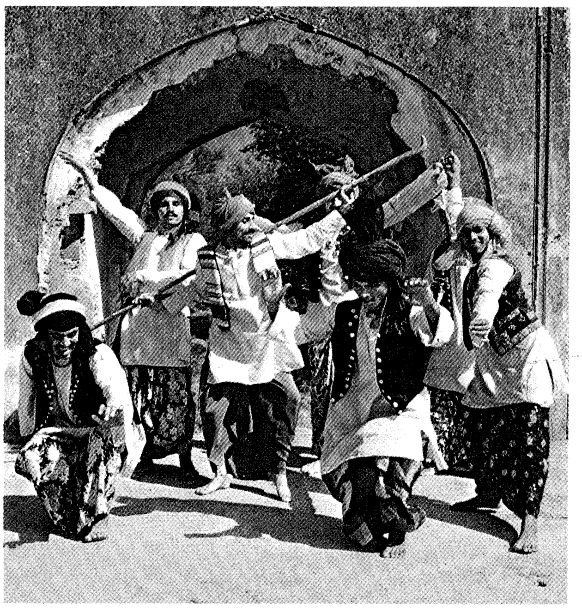
Muslim culture has frowned upon the performing arts, with the result that there is no Arabic or Persian classical theatre. The only possible sources of drama were the Persian passion plays dealing with the martyrdom of Ḥuysayn (grandson of Muḥammad) in the desert of Karbalā' in 680 AD, which have inspired some Urdu playwrights. Pakistan, a Muslim country, therefore either had to find a theatrical heritage in Urdu and Bengali theatre, which had been flourishing in India long before the partition, or look to the West. It did both. The Urdu-language theatre of Pakistan had started in the Lucknow court of Nawab Wajid Ali Shah in 1855 and was nurtured by both Muslim and Hindu artists. In Pakistan the *kathak* style is preferred because of its strong Muslim flavour and Mughal court associations. Cut off from Hinduism and its lore, Pakistani performers use these Indian classical dance styles to interpret the aspirations of a young nation, while their folk dances express the character of Pakistan's rural culture.

Folk Dance. Pakistan's dances are virile and explosive. *Bhangra* and *khatak* are the most popular. *Khatak* is a dance of the tribal Pathans, known for their hospitality and feuds in the rugged hills of the northwest. It originated in zealous preparations for raids and celebrations of victories. In the 20th century, any joyous event is the occasion for this community dance. The Pathans, dressed in baggy *salwars*, embroidered waistcoats, and skull-capped turbans, perform it holding a rifle in both hands. In a frenzy they spin and somersault, float and whirl, with sudden bursts of swordplay to the accompaniment of drums and pipes. Because of its popularity, *khatak* is presented to visiting dignitaries and for this purpose has been refined into choreographed productions.

Khatak

Female dances

Important dances by women are the *sammi, kikli, giddha,* and *luddi*. Except for the *sammi*, which has a slow rhythm accompanied by a sad song because of its associa-



Bhangra, folk dance of the Punjab region of Pakistan and India.
Lustig—FPG

tion with the tragic love legend of Princess Sammi and Prince Dhola, all the other forms are charged with energy and fast rhythms. The *kikli* is performed by teen-age girls in groups of two. The partners cross their arms, interlock their fingers, and, touching the toes of their feet, stretch backward and whirl. The *giddha* is danced in a circle, the participants keeping the rhythm by clapping their hands. Two women impulsively leave the circle, jump into the centre, and perform a hilarious mimetic dance enacting a *boli* (two-line song) and again join the circle to dance in a ring and allow another couple to take the centre. In the *luddi*, women click their fingers and clap their hands, moving in a circle by jumps and half-turns and accelerating their rhythm by stamping their feet.

Performing arts in the Punjab. The genius of Punjabis finds expression in love stories, lusty dancing, and humour. The *mirasi*s (professional wits), *naqalia*s (mummers), and *domani*s (female singer-actresses) are professional performers belonging to the lower classes. They exploit all the tricks of exaggeration, absurdity, malapropism, comic gags, and lewd references. In the performance of a *naqal* (comic sketch), two people constitute a troupe. The leader holds a leather folder and slaps his foolish partner, who leads his master to a hilarious situation through absurd replies. Expert in mime and clowning, these character types are distantly related to the Western court fool and the commedia dell'arte.

Theatre in Pakistan. Urdu theatre grew out of a spectacular production of *Indrasabha* ("The Heavenly Court of Indra"), an operatic drama written by the poet Agha Hasan Amanat and produced in 1855 in the palace courtyard of the last Nawab of Oudh, Wajid Ali Shah. The story deals with the love of a fairy and Prince Gulfam. The fairy takes her lover to heaven where the angry and jealous Indra hurls him down to earth. Finally, the fairy, through her songs and dances, wins the heart of Indra, and the two lovers are united. Wajid Ali Shah, an expert *kathak* dancer and author of many valuable treatises on stage techniques, composed some of the melodies and dances for his production and used folk conventions, gorgeous costumes, elaborate settings, and gold-inlaid masks. *Indrasabha* was a fantastic success; it was translated into almost all the regional languages, with many local variants. Its characters—Sabaz Pari (Green Fairy), Kala Deo (Black Devil), Lal Deo (Red Devil)—became a part of the theatrical vocabulary of the subcontinent.

Parsi theatre. During the second half of the 19th century, Urdu was the main spoken and written language of the northern half of the subcontinent and understood in almost all the principal cities. The Parsis (originally Zoroastrians from Iran who settled on the coast of Bombay), comprising a wealthy community with sharp business acumen, were the pioneers in establishing a commercial theatre, that lasted from 1873 to 1935 and influenced all the other regional theatres. Though located mainly in Bombay and Calcutta, the Parsi companies toured the subcontinent with huge staffs, sets, and an army of players.

The best known playwright of this period is Agha Hashr (1876–1935), a poet-dramatist of flamboyant imagination and superb craftsmanship. Among his famous plays are *Sita Banbas*, based on an incident from the *Rāmāyaṇa; Bilwa Mangal*, a social play on the life of a poet, whose blind passion for a prostitute results in remorse; and *Aankh ka Nasha* ("The Witchery of the Eyes"), about the treachery of a prostitute's love, with realistic dialogue of a brothel. Many of Hashr's plays were adapted from Shakespeare: *Sufayd Khūn* ("White Blood") was modelled on *King Lear*, and *Khūn-e Nāḥaq* ("The Innocent Murder") on *Hamlet*. His last play, *Rustam-o-Sohrab*, the tragic story of two legendary Persian heroes, Rustam and his son Sohrab, is a drama of passion and fatal irony.

Parsi productions

Productions by Parsi theatrical companies were large-budgeted affairs. Plays opened with the actors in full makeup and costume, their hands folded and eyes closed, singing a prayer song in praise of some deity, and generally ended in a tableau. Sometimes at curtain call the director rearranged the tableau in a split second and offered a variant. Actors were required to know singing, dancing, music, acrobatics, and fencing and to possess strong voices and good physical bearing. In improvised auditoriums with bad acoustics and packed with more than 2,000 people, actors' voices reached the farthest spectator. Plays began at 10 o'clock and lasted until dawn, moving from comedy to tragedy, from pathos to farce, from songs to the rattle of swords, all interspersed with moral lessons and rhyming epigrams. The droll humour and realism of the comic interludes remain unsurpassed in contemporary Urdu drama. Important playwrights of this period were Narain Prasad Betab, Mian Zarif, and Munshi Mohammed Dil of Lucknow. All took inspiration from Hindu mythology and Persian legends, transforming these tales into powerful dramas.

Imtiaz Ali Taj (1900–70) was a bridge between Agha Hashr and contemporary Pakistani playwrights. His *Anarkali* (1922), the tragic love story of a harem girl, Anarkali, and Crown Prince Salim (son of Akbar the Great), unfolds the love–hate relationship of a domineering emperor and his rebellious son. Brilliant in treatment and character analysis, this play has been staged hundreds of times by amateur groups and has entered the list of Urdu classics.

In the absence of a professional company, Urdu theatre has found it difficult to strike roots. After 1947 many Muslim actors and writers were absorbed by the Indian film industry in Bombay, and they found it difficult to adjust their great talent to amateur theatrical clubs. All the same, plays have been staged in Karāchi, Lahore, and Rāwalpindi. The best productions have been those dealing with topical themes—refugee problems, new adjustments, the corrupt bureaucracy, the Kashmir issue, and other sociopolitical issues. Agha Babar in Rāwalpindi produced *Burra Sahib* (1961: "The Big Boss"), an adaptation of Gogol's *Government Inspector*, setting it in Pakistan. *Tere Kuce se Jub Hum Nikle* ("Thrown Out of Your Lane"), by Naseer Shamshi, describes the pathetic condition of an aristocratic family in Delhi that is forced to leave home because of communal riots. In *Lal Qile se Lalukhet Tak* ("From the Red Fort to Lalukhet"), by Khwajah Moinuddin, the comedy arises out of the pitiable condition of the refugees who leave their well-settled existence in Delhi dreaming of prosperity, take a tedious journey, and arrive homeless in Karāchi to find shelter in thatched hovels. Ali Ahmed, an avant-garde actor-director in Karāchi, presents his plays with polished stagecraft and esoteric appeal.

Lahore remains the centre of amateur theatre based on the tradition of the late directors A.S. Bokhari and G.D.

Early Indian architecture and sculpture



Stūpa III at Sāñcī, Madhya Pradesh, sandstone, 1st century BC.



Detail of a sandstone relief sculpture from the *torana* of Stūpa I at Sāñcī, Madhya Pradesh, 1st century BC.

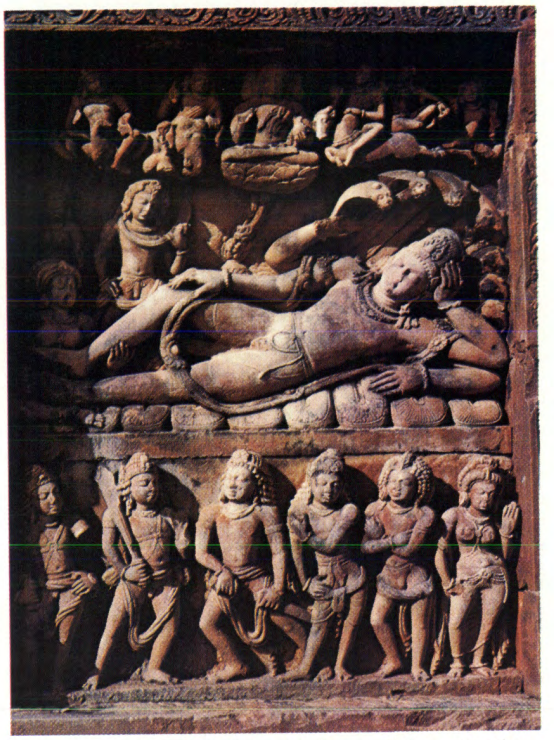


Fragment of a soapstone disc from Kausāmbī, Uttar Pradesh, 3rd century BC. In the Municipal Museum, Allāhābād, Uttar Pradesh. Height 5.7 cm.



Abduction scene, terra-cotta relief from Kausāmbī, Uttar Pradesh, *c.* 2nd century BC. In the Municipal Museum, Allāhābād, Uttar Pradesh.

Plates 1–8: P. Chandra



Visnu lying on the serpent Sesa; sandstone relief panel on the Visnu temple at Deogarh, Orissa, 5th century AD.



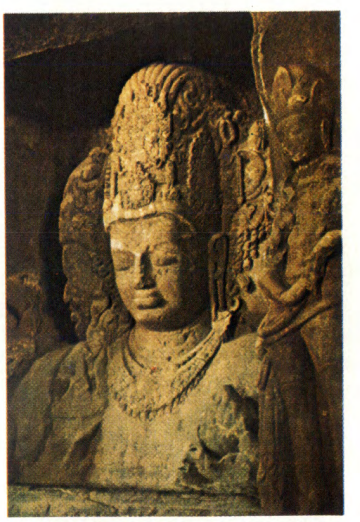


Lotus scroll painted on the ceiling of Cave 2 at Ajanta, Mahārāshtra, 5th century AD.

Indian art of the Golden Age: 5th to 6th centuries AD

(Below) The Pārvatī temple at Nāchna, Rājasthān, sandstone, 5th century AD. (Left) Amorous couples and lotus scroll, detail of the doorframe of the Pārvatī temple.





Siva-Mahesamūrti, the main image in the cave Temple at Elephanta, Mahārāshtra, 6th century AD.

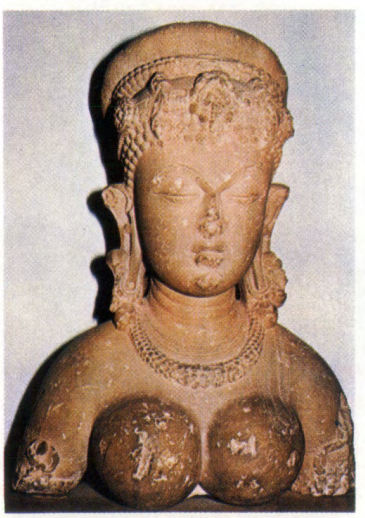
Medieval sculpture of North and South India



The hermitage by the Ganges, detail of a granite relief depicting the penance of Arjuna, from Mahābalipuram, Tamil Nadu, early 7th century AD.



The god Siva, detail of a doorframe from Singhpur, Madhya Pradesh, sandstone, 10th century AD.



Bust of a goddess from the fort at Gwalior, Madhya Pradesh. c. 9th century AD. In the Gwalior Museum, Madhya Pradesh. Height 54 cm.

Seated Avalokitesvara, gilt bronze sculpture from Nālandā, Bihār, 8th century AD. In the Nālandā Museum, Bihār. Height 28 cm.



Siva and his consort, bronze sculpture from Tiruvenkadu, Tamil Nadu, early 11th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu. Height of male 106 cm., height of female 94 cm.



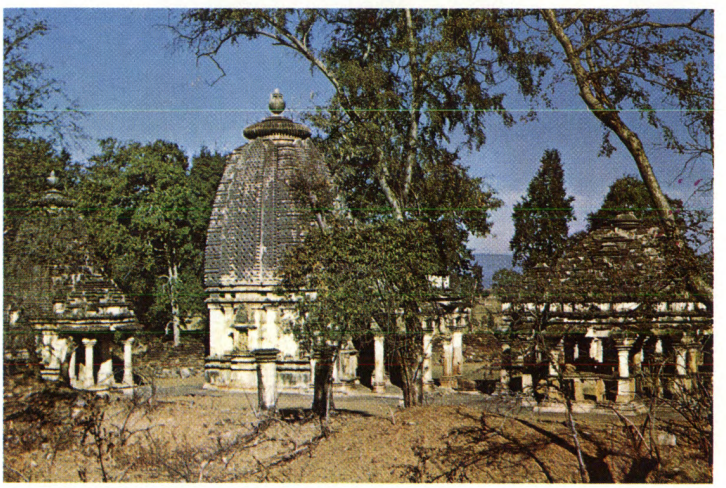
Medieval temple architecture: styles of North and South India



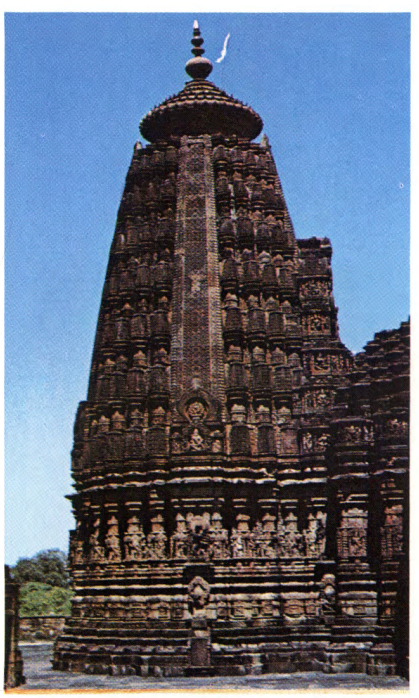
Temple with *kutina* superstructure, the Tālapurīsvara at Panamalai, Tamil Nadu, early 8th century AD.



Temples, tank, and *gopura* of the Siva temple at Cidambaram, Tamil Nadu, 12th–13th century AD.



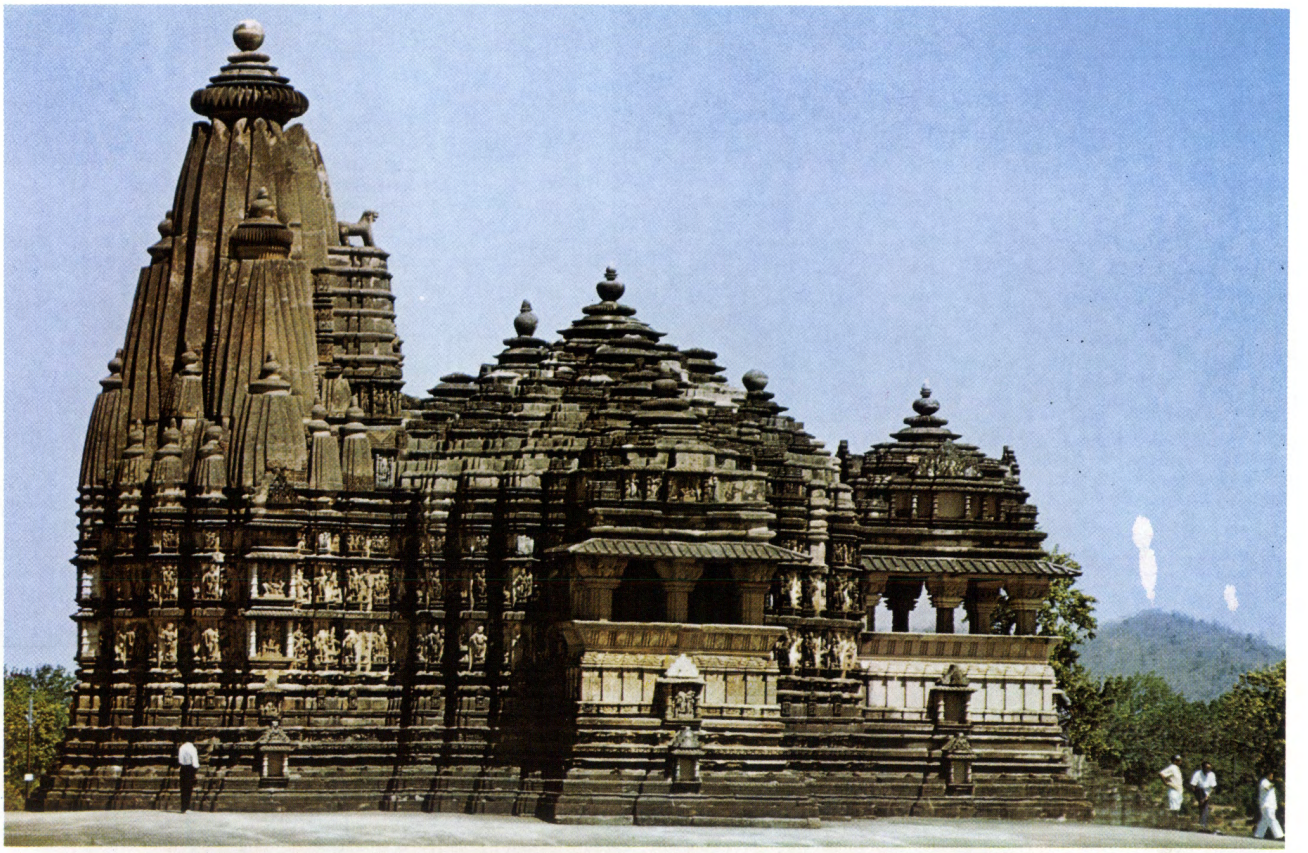
Group of temples at Badoli, Rājasthān, 10th century AD.

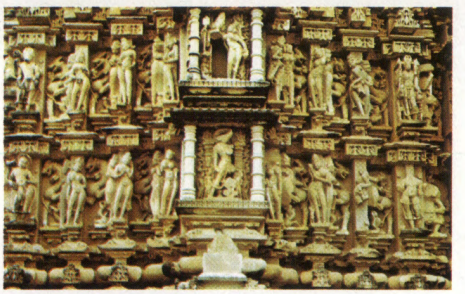


Temple with *bhumija* superstructure, the Udayesvara at Udaipur, Rājasthān, *c.* AD 1059–82.



Temple with *latina* superstructure at Umrī, Madhya Pradesh, 9th century AD.





(Above) The Citragupta temple at Khajurāho, Madhya Pradesh, sandstone, 11th century AD. (Left) Detail of the temple wall.





(Left) The Brhadīsvara temple at Thanjāvūr, Tamil Nadu, early 11th century AD. (Right) Detail of the temple wall.



The Kesava temple at Somnāthpur, Gujarāt, *c.* AD 1268.

Indian miniature painting: the indigenous tradition



The monk Kālaka addressing the Sāhi king, detail from a folio from a *Kālakācāryakathā* manuscript, Western Indian style, late 13th century AD. In the Prince of Wales Museum of Western India, Bombay. Dimensions of entire folio 7.6 × 7.6 cm.



Ladies in conversation, detail from a folio from a *Mahābhārata* manuscript, AD 1516. In the collection of the Asiatic Society of Bombay. Dimensions of entire folio 10.2 × 10.2 cm.



A prince and his lady, Rajasthani style, Mālwa, mid-17th century AD. In a private collection.



Krsna and Rādhā, miniature from a series illustrating the *Gitagovinda*, Rajasthani style, Mewār, mid-17th century AD. In a private collection.



A hill chief smoking, Pahari style, Basohlī, late 17th century AD. In the National Museum of India, New Delhi.

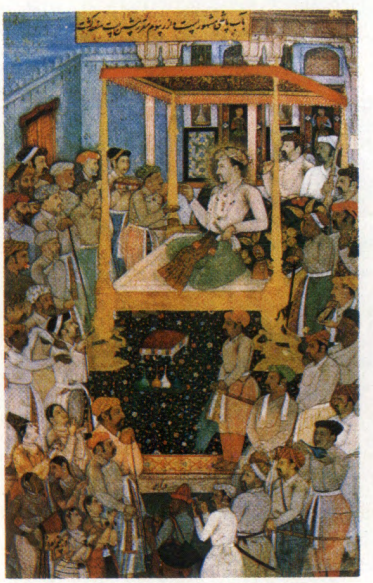
The Mughal style and its influences



A queen on a hunting expedition, Pahari style, Kāngra, *c.* AD 1760. In the National Museum of India, New Delhi.



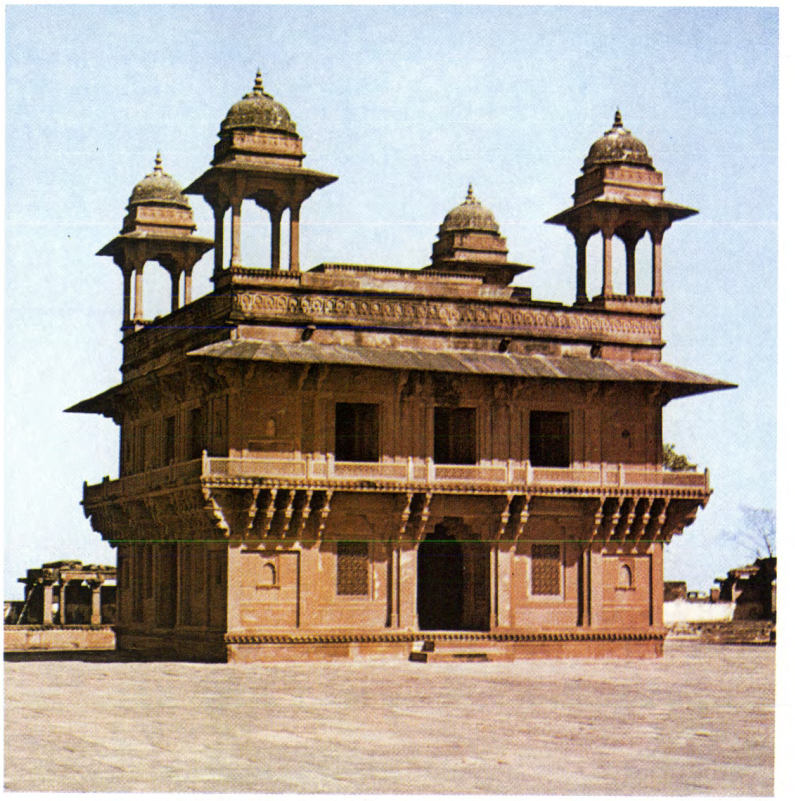
Court scene by Basāvan, folio from an illustrated manuscript of the *Anwār-e Suhayli*, Mughal style, AD 1596–97. In the Bharat Kala Bhavan, Vārānasi, Uttar Pradesh. 24.8 × 13.9 cm.



The feast of Nauroz at Jahāngīr's court, Mughal style, *c.* AD 1615. In the collection of the Reza Library, Rāmpur, Uttar Pradesh.



The musical mode Magha-Malara, Rajasthani style, Būndi, late 17th century AD. In a private collection.

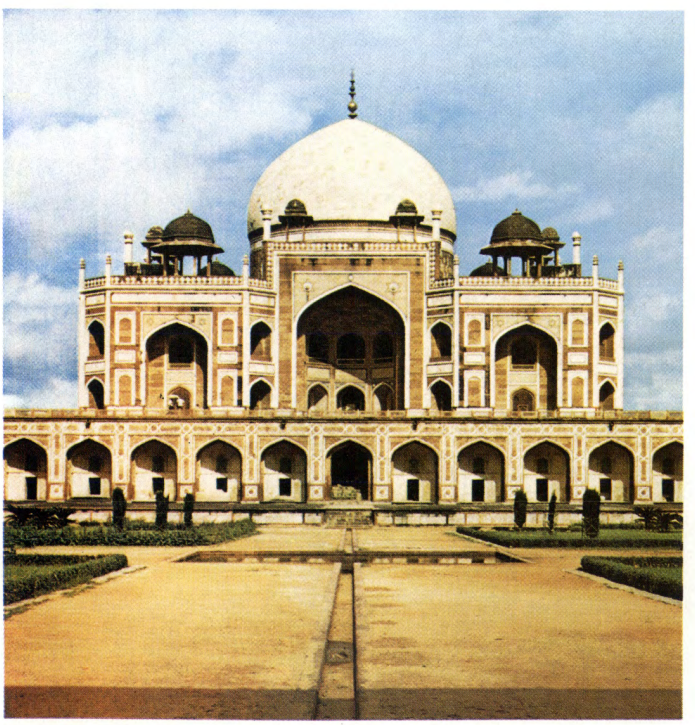


The so-called Dīvān-e Khāss at Fatehpur Sīkri, red sandstone, *c.* AD 1585.

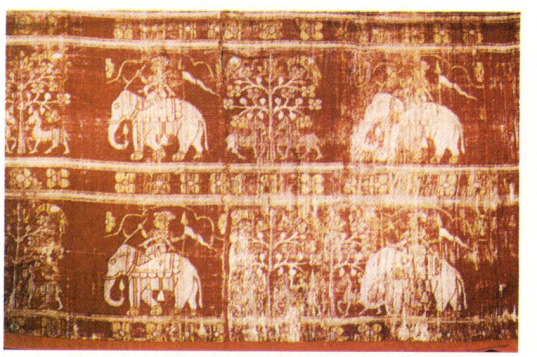
Copper bowl with hunting scene, Mughal, *c.* AD 1583. In the Prince of Wales Museum of Western India, Bombay.



Mughal architecture and decorative arts



The tomb of Humāyūn at Delhi, red sandstone and marble, *c.* AD 1564.



Detail of a silk tapestry, Mughal, early 17th century AD. In the Prince of Wales Museum of Western India, Bombay.



Jade dish inlaid with semiprecious stones and gold, Mughal, mid-17th century AD. In the Prince of Wales Museum of Western India, Bombay

Sondhi, both former principals of the Government College in Lahore. In 1942 G.D. Sondhi built the Open-Air Theatre, situated on a small artificial hillock in the Lawrence Gardens and perhaps the best in all of South Asia. It has remained the centre of dramatic contests and festivals and is a favourite of visiting dancers and actors.

The actor-playwright Rafi Peer, with his knowledge of Western theatre as a result of his training in Berlin in the 1930s, has helped to develop Pakistani theatre. Professional in approach, he has produced radio and stage plays and has been a critical colleague of A.S. Bokhari and Imtiaz in the revival of amateur theatre.

Plays are being written for radio and television that are readily adaptable for the stage, and vice versa. Saadat Hussan Mantoo (1912–55), the greatest writer of short stories and author of over 100 radio plays and features, is still the model for contemporary writers for plot construction, bitter realism, and whimsical dialogue. His collection of plays (1942–45), including *Mantoo ke Drama* ("Mantoo's Plays"), *Ao* ("Come"), and *Tin Aurten* ("Three Women"), have flashes of the then-unborn Theatre of the Absurd.

More dramas are written in Urdu today than are staged. The turnover is large because of their generally amateurish character and short runs. There is no major professional centre for the training of actors, nor a school for stagecraft and production. Notwithstanding, the young directors and playwrights have been enthusiastic about establishing a permanent Urdu stage.

BANGLADESH

East Bengal has continued the folk *jātrā* and used this form for themes concerning current political problems and historical events. A successful example of the latter is *Nawab Sirajudawla*, which deals with the fall of the last Muslim ruler of Bengal in 1757 through betrayal by his ambitious brother-in-law Mīr Jaʿfar, who joined the British. This *jātrā* is popular with both rural and urban audiences. Tales of Muslim kings and lovers from Persian legends also have been rendered into *jātrās*.

Contemporary theatre inherits the tradition of the prepartition Bengali stage. The poet-playwright Nazrul Islam followed the tradition of Tagore's verse plays and dance operas. Inspired by left-wing ideology, he wrote for the People's Theatre in East Bengal, championing the cause of the poor farmer. He has dealt with psychological problems and inner tensions in his *Shilpi* ("The Artist"), in which the artist is torn between love for his wife and for his art. Especially popular are historical themes of political significance, inspiring Muslims who for centuries were subjugated by the Hindus of East Bengal. Ebrahim Khan wrote *Kamal Pasha* (1926), a play about the Turkish liberator, a symbol of hope and reawakening, and *Anwar Pasha*, about the downfall of Anwar (Enver), who could not cope with the new historical forces.

East Bengal has a solid acting tradition and a rich repertoire of Bengali plays. Its amateur stage has professional actors, and it retains the impassioned lyricism and power of the mainstream of Bengali tradition. (B.Ga.)

IV. Visual arts

VISUAL ARTS OF THE INDIAN SUBCONTINENT

Indian art is the term commonly used to designate the art of the Indian subcontinent, which includes the present political divisions of India, Kashmir, Pakistan, and Bangladesh. Although a relationship between political history and the history of Indian art before the advent of Islām is at best problematical, a brief review will provide a broad context. The earliest urban culture of the subcontinent is represented by the Indus Valley civilization (*c.* 2500–1800 BC), which possessed several flourishing cities not only in the Indus Valley but also in Gujarāt and Rājasthān. The circumstances in which this culture came to an end are obscure. Although there is no clear proof of historical continuity, scholars have noticed several striking similarities between this early culture and features of later Indian civilization. The period immediately following the urban Indus Valley civilization is marked by a variety of essentially rural cultures. A second urbanization began to occur only around the 6th century BC, when flourishing cities started to reappear, particularly in the Gangetic Basin. The Buddha lived and preached in this period, which culminated in the great Maurya Empire, whose relatively few works are the earliest surviving remnants of monumental art. The Maurya dynast Aśoka (died 238 BC) is considered the greatest of Buddhist kings; and the majority of the monuments of the next 500 years appear to be dedicated to the Buddhist faith, though iconographical and other details suggest that the art also drew heavily on popular religion.

The Maurya Empire spread over almost all of what is modern India and Pakistan. Territories as extensive were never possessed by any other dynasty. With its fall, the empire broke up into a number of states ruled by many dynasties, some of which acquired considerable power and fame for varying periods of time. Among these, the Śuṅgas (*c.* 2nd–1st century BC) in the north and the longer-lived Sātavāhanas in the Deccan and the south are particularly noteworthy. Though these kings were Hindu by religion, Buddhist monuments form the great majority of surviving works.

Toward the end of the 1st century BC, northern India was subjected to a series of invasions by Scythian tribes, resulting finally in the establishment of the vast Kuṣāṇa empire, of which Mathurā was an important centre. The new rulers seemed to have followed Indian faiths, the great emperor Kaniṣka (*c.* AD 78) being a devout Buddhist. The schools of Gandhāra and Mathurā flourished during their rule, and, though much of the work is dedicated to the Buddhist religion, the foundations of later Hindu iconography were also laid in this period. While the Kuṣāṇa dynasty was sovereign in the north, the Sātavāhanas continued to rule in the south. The bulk of the work at Amarāvatī was produced during their hegemony.

The Golden Age of India

Around the mid-4th century, the Gupta dynasty, of indigenous origin, rapidly expanded its power, uprooting the last remnants of foreign rule and succeeding in bringing almost all of northern India under its sway. In the Deccan there arose at the same time the equally powerful Vākāṭakas, with whom the Guptas appear to have had friendly relations. The period extending from the 4th through the 5th centuries is marked by the most flourishing artistic activities. In addition to the Buddhist monuments, there are the first strong indications of specifically Hindu patronage. Works of remarkable beauty and elegance were produced in this period, which is commonly called the Golden Age of India.

The disintegration of these two empires toward the close of the 5th and the 6th centuries ushered in what has been called the medieval period (*c.* 8th–12th centuries), marked by the appearance of a large number of states and dynasties, often at war with each other. Their rise to power and their decline was part of a constantly recurring process, for none of them was able to hold onto a position of even relative paramountcy for any extended period of time. In the north, the great dynasties were the Gurjara-Pratīhāras, whose empire at its greatest equalled that of the Guptas; the Pālas, who ruled chiefly over northeastern India; and various other dynasties, such as the Kalacuris, the Candelas, and the Paramāras of north central India, the Cāhamānas of Rājasthān, the Cālukyas of Gujarāt. In the Deccan, also, several dynasties rose and fell, the most powerful of which were the Cālukyas of Bādāmi, the Rāṣṭrakūṭas, and the Cālukyas of Kalyāṇī. They were often at war not only with their powerful neighbours to the north but also with the great Pallava and Cōḷa kingdoms of southern India. Most of the dynasties of medieval India were Hindu, though some Jaina and a very few Buddhist kings are also known. The various faiths, however, existed in comparative harmony; and Buddhist and Jaina monuments continued to be built, though most of the surviving works are Hindu.

Although the effects of constant struggle were not as devastating as one might expect, largely as a result of the institutionalization of war and its confinement to appropriate castes, the Hindu kingdoms fell easy prey to the Islāmic invasions, which began as early as the 8th century AD but gathered strength only in the 11th century. By the

end of the 12th century, almost all of northern India had been conquered. Islāmic advances in the south were checked for a while by the Vijayanagara dynasty, but with its collapse almost all of India fell under various degrees of Islāmic hegemony. Large Hindu kingdoms enjoying differing degrees of independence continued to exist chiefly in Rājasthān and portions of southern India, but overall political supremacy was vested with the Islāmic states. The Muslim powers were also divided into many kingdoms, despite attempts made by the sultanate of Delhi, and later by the Mughals, to achieve paramountcy over large portions of India. These attempts were successful only for short periods of time. Although the initial impact of Islām on Indian art was generally destructive, Islāmic influences entering India were gradually transformed in the new environment and eventually resulted in the flowering of an extremely rich and important aspect of the Indian genius.

Ascendency of the European powers

The ascendency of the European powers in the 18th century, culminating in the establishment of the British Empire, laid the foundation of modern India's contacts with the West. As a whole, the European advent was marked by a relative insensitivity to native art traditions, but rising nationalism attempted a conscious revival of Indian art toward the end of the 19th century. In modern times, the absorption of European influence is a more natural, freer process that affects artistic development in a vital and profound way.

General characteristics. *The unity of Indian art.* Indian art is spread over a subcontinent and has a long, very productive history; but it nevertheless shows a remarkable unity and consistency. Works produced in the several geographical and cultural regions possess decidedly individual characteristics but at the same time have sufficient elements in common to justify their being considered manifestations of a general style. The existence of this style is evidence of the essential cultural unity of the subcontinent and to the uninterrupted contact between the various geographical units, at least from the historical period onward. Developments in one area have been quickly reflected in the others. The regional idioms have contributed to the richness of Indian art, and the mutual influences exercised by them have been responsible for the multi-faceted development of that art throughout the course of its long life.

The style of Indian art is largely determined not by a dynasty but by conditions of time and space. It has, essentially, a geographical rather than a dynastic basis, which is to say that the evolution of regional schools appears to have been largely independent of any particular dynasty that happened to rule over a specific region. The style does not change because of the conquest of one area by another dynasty; rather the influences exercised by one area on another are usually through the agency of factors other than conquest. Instances in which dynastic patronage changed the nature of a style are very few and confined mostly to the Islāmic period. The political history of India is itself quite vague, and the areas in possession of a dynasty at various points in its history are even less susceptible to precise definition. For all these reasons, the classification of Indian art adopted here is not based on dynasties, for such a division has little meaning. Nevertheless, names of certain dynasties are used, for these have passed into common usage. When this is done, however, the name must be understood as little more than a convenient way of labelling a particular period.

The materials of Indian art. Indian art employs various materials, such as wood, brick, clay, stone, and metal. Most wooden monuments of the early period have perished but have been imitated in stone. Clay and brick were also abundantly used; but, particularly in later times, the favoured material seems to have been stone, in the dressing (facing and smoothing) and carving of which the Indian artist attained great excellence. The material may have influenced the form somewhat, but essentially Indian art tends to impose the form on the material. Thus, materials are generally regarded as interchangeable: wooden and clay forms are imitated in stone, and stone is imitated in bronze, and in turn stone sculpture assumes qualities appropriate to metal. It is as though the nature of the material presented a challenge that had to be met and overcome. At the same time, Indian art stresses the plasticity of forms; sculpture is generally characterized by emphatic mass and volume; architecture is often sculpture on a colossal scale; and the elements of painting, particularly of the early period, are modelled by line and colour.

Emphasis on the plasticity of forms

Indian and foreign art. Thanks to its geographical situation, the Indian subcontinent has been constantly fed by artistic traditions emanating from West and Central Asia. The Indian artist has shown a remarkable capacity for accepting these foreign influences naturally and assimilating and transforming them to accord with the nature of his own style. The process occurred frequently: in the Maurya period; in the two centuries after Christ, when the Kuṣāṇa dynasty attained imperial supremacy in the north; and at a much later period, in the 16th century, when the Mughals patronized a new school of architecture and painting.

Just as India received influences, so it transmitted its own art abroad, particularly to Ceylon and the countries of Southeast Asia. Developments of great importance were thereby precipitated in Ceylon, Burma, Thailand, Indonesia, and Indochina, where the reinterpretation of Indian influences resulted in the creation of works of great originality.

Indian art and religion. Indian art is religious inasmuch as it is largely dedicated to the service of one of several great religions. It may be didactic or edificatory as is the relief sculpture of the two centuries before and after Christ; or, by representing the divinity in symbolic form (whether architectural or figural), its purpose may be to induce contemplation and thereby put the worshipper in communication with the divine. Not all Indian art, however, is purely religious, and some of it is only nominally so. There were periods when humanistic currents flowed strongly under the guise of edificatory or contemplative imagery, the art inspired by and delighting in the life of this world.

Although Indian art is religious, there is no such thing as a sectarian Hindu or Buddhist art, for style is a function of time and place and not of religion. Thus it is not strictly correct to speak of Hindu or Buddhist art, but, rather, of Indian art that happens to render Hindu or Buddhist themes. For example, an image of Viṣṇu and an image of Buddha of the same period are stylistically the same, religion having little to do with the mode of artistic expression. Nor should this be surprising in view of the fact that the artists belonged to nondenominational guilds, ready to lend their services to any patron, whether Hindu, Buddhist, or Jaina.

Religion, symbolism, and abstraction

The religious nature of Indian art accounts to some extent for its essentially symbolic and abstract nature. It scrupulously avoids illusionistic effects, evoked by imitation of the physical and ephemeral world of the senses; instead, objects are made in imitation of ideal, divine prototypes, whose source is the inner world of the mind. This attitude may account for the relative absence of portraiture and for the fact that, even when it is attempted, the emphasis is on the ideal person behind the human lineaments rather than on the physical likeness.

The artist and patron. Works of art in India were produced by artists at the behest of a patron, who might commission an object to worship for spiritual or material ends, in fulfillment of a vow, for the discharge of virtues enjoined by scripture, or even for personal glory. Once the artist received his commission, he fashioned the work of art according to his skill, gained by apprenticeship, and the written canons of his art, which possessed a holy character. There were prescribed rules for proportionate measurement, iconography, and the like, often with a symbolic significance. This is not to say that the individual artist was invariably aware of the symbolic meaning of the prescribed standards, based as these were on complex metaphysical and theological considerations; but the symbolism nevertheless formed part of the fabric of his work, ready to add an extra dimension of meaning to the initiated and knowledgeable spectator.

In these conditions it is not surprising that the artist as a person is for the most part anonymous, very few names of artists having survived. It was the skill with which the work of art was made to conform to established ideals, rather than the artist who possessed the skill, that held the place of first importance.

The appreciation of Indian art. According to Indian aesthetic theory, a work of art possesses distinct "flavours" (*rasa*), the "tasting" of which constitutes the aesthetic experience. Because the work of art operates at various levels, granting to the spectator what he is capable of receiving by virtue of his intellectual and emotional preparation, the appreciation of the beauty of form and line is considered an appropriate activity of the educated and cultured man. The supreme aesthetic experience, however, is believed to be much deeper and cognate to the experience of the Godhead. From this point of view, the work of art is in a sense irrelevant and unnecessary for a person at a high level of spiritual progress; and for the devout layman its excellence is measured by its efficacy in promoting spiritual development.

Architecture. The favoured material of early Indian architecture appears to have been wood, but little has survived the rigours of the climate. Wooden forms, however, affected work in other mediums and were sometimes quite literally copied, as, for example, in early cave temples of western India. The principles of wooden construction also played an important part in determining the shape of Indian architecture and its various elements and components.

Effect of wooden forms on other mediums

Baked or sun-dried brick has a history as ancient as that of wood; among the earliest remains are buildings excavated at sites of the Indus Valley civilization. The use of brick is once again evident from about the 6th century BC, and its popularity was undiminished in subsequent centuries. Many brick monuments have been discovered, particularly in areas in which good clay was easily available, such as the Gangetic Basin. Although more durable than wood, few brick buildings from before the 5th century AD have survived in a good state of preservation.

Traditions of stone architecture appear to be more recent than wood or brick, the earliest examples of the use of dressed stone for building purposes not predating the 6th century BC. The Indian architect, however, soon gained great proficiency in its use, and, by the 7th century AD, the use of stone for monumental buildings of considerable size had become quite popular. The preference for stone can also be seen in Islāmic monuments of India, which contrast markedly with the brick and tile structures popular in neighbouring West Asia.

Most surviving examples of Indian architecture before the Islāmic period are of a religious nature, consisting mainly of Buddhist shrines, or *stūpa*s, and temples. Monastic residences give some idea of civil architecture, but, surprisingly, very few examples of palaces and secular dwellings have been found.

Indus Valley civilization (c. 2500–1800 BC). From excavated remains, it is clear that the Indus Valley civilization possessed a flourishing urban architecture. The major cities associated with the civilization, notably Mohenjo-daro, Harappā, and Kalibangan, were laid out on a grid pattern and had provisions for an advanced drainage system. The residential buildings, which were serviceable enough, were mainly brick and consisted of an open patio flanked by rooms. For monumental architecture, the evidence is slight, the most important being a "sacred" tank (thought to be for ritual ablution) and associated structures. Corbel vaulting (arches supported by brackets projecting from the wall) was known, and, to a limited extent, timber was used together with brick; whatever architectural ornamentation existed must have been of brick or plaster.

The Maurya period (c. 321–185 BC). The state of Indian architecture in the period between the Indus Valley civilization and the rise of the Maurya Empire is largely unknown since most work was done in such perishable material as wood or brick. Excavations at Rājgīr, Kauśāmbī, and other sites, however, testify to the existence of fortified cities with *stūpa*s, monasteries, and temples of the type found at the later Maurya sites of Nagarī and Vidiśā; and there is evidence of the use of dressed stone in a palace excavated at Kauśāmbī. Considering the power of the Maurya Empire and the extensive territory it controlled, the architectural remains are remarkably few. The most important are *stūpa*s (later enlarged) such as a famous example of Sāñcī (Sanchi); the ruins of a hall excavated at the site of Kumrāhar in Patna (ancient Pāṭaliputra), the capital city; and a series of rock-cut caves in the Barābar and Nāgārjunī hills near Gayā, which are interesting because they preserve in the more permanent rock some types of wooden buildings popular at that time.

The *stūpa*

The *stūpa*, the most typical monument of the Buddhist faith, consists essentially of a domical mound in which sacred relics are enshrined. Its origins are traced to mounds, or tumuli, raised over the buried remains of the dead that were found in India even before the rise of Buddhism. *Stūpa*s appear to have had a regular architectural form in the Maurya period: the mound was sometimes provided with a parasol surrounded by a miniature railing on the top, raised on a terrace, and the whole surrounded by a large railing consisting of posts, crossbars, and a coping (the capping on the top course), all secured by tenons and mortices in a technique appropriate to craftsmanship in wood. The essential feature of the *stūpa*, however, always remained the domical mound, the other elements being optional.

Along with *stūpa*s were erected roofless, or hypaethral, shrines enclosing a sacred object such as a tree or an altar. Temples of brick and timber with vaulted or domical roofs were also constructed, on plans that were generally elliptical, circular, quadrilateral, or apsidal (*i.e.*, having an apse, or semicircular plan, at the sanctum end). These structures have not survived, but some idea of their shape has been obtained from the excavated foundations and the few examples imitating wooden originals that were cut into the rock, notably the Sudāmā and the Lomas Ṛṣi caves in the Nāgārjunī and Barābar hills near Gayā. The latter has an interesting entrance showing an edged barrel-vault roof (an arch shaped like a half cylinder) in profile supported on raked pillars, the ogee arch (an arch with curving sides, concave above and convex toward the top) so formed filled with a trellis to let in light and air. The interiors of most caves are highly polished and consist of two chambers: a shrine, elliptical or circular in plan with a domed roof (Sudāmā cave); and an adjacent antechamber, roughly rectangular and provided with a barrel vault. Remains of structural buildings have been excavated at Bairāt and Vidiśā, where wood and brick shrines with timber domes and vaults once existed. A temple (No. 40) at Sāñcī was apsidal in plan and perhaps had a barrel-vault roof of timber.

A hall excavated at Kumrāhar in Patna had a high wooden platform of most excellent workmanship, on which stood eight rows of ten columns each, which once supported a second story. Only one stone pillar has been recovered, and it is circular in shape and made of sandstone that has been polished to a high lustre. The capitals that topped them must have been similar to others found in neighbouring Lohanipur and almost certainly consisted of one or two pairs of addorsed (set back to back) animals, recalling Persepolitan examples. Indeed, there is much about Maurya architecture and sculpture to suggest Iranian influence, however substantially transformed in the Indian environment.

Early Indian architecture (2nd century BC–3rd century AD). Except for *stūpa*s, architectural remains from the 2nd century BC (downfall of the Maurya dynasty) to the 4th century AD (rise of the Gupta dynasty) continue to be rare, indicating that most of the work was done in brick and timber. Once again, examples cut into the rock and closely imitating wooden forms give a fairly accurate idea of at least some types of buildings in this period.

The *stūpa*s become progressively larger and more elaborate. The railings continue to imitate wooden construction and are often profusely carved, as at Bhārhut, Sāñcī II, and Amarāvatī. These were also provided with elaborate gateways, consisting of posts supporting from one to three architraves, again imitating wooden forms and

covered with sculpture (Bhārahut, Sāñcī I, III). In the course of time an attempt was made to give height to the *stūpa*s by multiplying the terraces that supported the dome and by increasing the number of parasols on top. In Gandhāra and southeastern India, particularly, sculptured decoration was extended to the *stūpa* proper, so that terraces, drums, and domes—as well as railing—were decorated with figural and ornamental sculpture in bas-relief. *Stūpa*s in Gandhāra were not provided with railings but, instead, had rows of small temples arranged on a rectangular plan.

West Indian cave temples

Cave temples of western India, cut into the scarp of the Western Ghāts and stretching from Gujarāt to southern Mahārāshtra, constitute the most extensive architectural remains of the period. Two main types of buildings can be distinguished, the temple proper (*caitya*) and the monastery (*vihāra, saṅghārāma*). The former is generally an apsidal hall with a central nave flanked by aisles. The apse is covered by a half dome; and two rows of pillars, which demarcate the nave, support a barrel-vault roof that covers the rest of the building. In the apsidal end is placed the object to be worshipped, generally a *stūpa*, the hall being meant for the gathered congregation. In front of the hall is a porch, separated from it by a screen wall provided with a door of considerable size, together with an arched opening on top clearly derived from wooden buildings of the Lomas Ṛṣi type and permitting air and dim light to filter into the interior. Other influences of wooden construction are equally striking, particularly in the vaulting ribs that cover the entire ceiling and that are sometimes actually of wood, as at Bhājā, where the pillars are also raked in imitation of the exigencies of wooden construction. The pillars are generally octagonal with a pot-shaped base and a capital of addorsed animals placed on a bell-shaped, or campaniform, lotus in the Maurya tradition. The most significent example is at Kārlī, dating approximately to the closing years of the 1st century BC. The Bhājā *caitya* is certainly the earliest, and important examples are to be found at Beḍsā, Kondane, Pītalkhorā, Ajantā, and Nāsik. Toward the end of the period, a quadrilateral plan appears more and more frequently, as, for example, at Kuda and Sailarwāḍī.

In addition to the *caitya*, or temple proper, numerous monasteries (*vihāra*s) are also cut into the rock. These are generally provided with a pillared porch and a screen wall pierced with doorways leading into the interior, which consists of a "courtyard" or congregation hall in the three walls of which are the monks' cells. The surviving rock-cut examples are all of one story, though the facade of the great monastery at Pitalkhorā simulates a building of several stories.

Monasteries carved into the rock are also known from Orissa (Udayagiri-Khandagiri), in eastern India. These are much humbler than their counterparts in western India, and consist of a row of cells that open out into a porch, the hall being absent. At Uparkot in Junāgadh, Gujarāt, is a remarkable series of rock-cut structures dating from the 3rd–4th century AD, which appear to be secular in character and in all probability served as royal pleasure houses.

The large number of representations of buildings found on relief sculpture from sites such as Bhārhut, Sāñcī, Mathurā, and Amarāvatī are a rich source of information about early Indian architecture. They depict walled and moated cities with massive gates, elaborate multi-storied residences, pavilions with a variety of domes, together with the simple, thatched-roofed huts that remained the basis of most Indian architectural forms. A striking feature of this early Indian architecture is the consistent and profuse use of arched windows and doors, which are extremely important elements of the architectural decor.

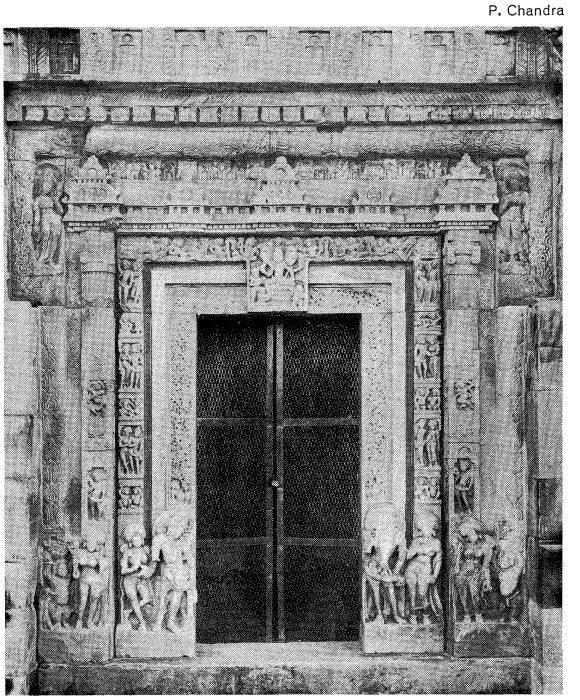
The Gupta period (4th–6th centuries AD). Dating toward the close of the 4th and the beginning of the 5th century AD is a series of temples that marks the opening phase of an architecture that is no longer content with merely imitating wooden building but initiates a new movement, ultimately leading to the great and elaborate temples of the 8th century onward.

Two main temple types of the Gupta period

Two main temple types have been distinguished in the Gupta period. The first consists of a square, dark sanctum with a small, pillared porch in front, both covered with flat roofs. This type of temple answers the simplest needs of worship, a chamber to house the deity and a roof to shelter the devotee. Temple No. 17 at Sāñcī is a classic example of this flat-roofed type. The plain walls are of ashlar masonry (made up of squared stone blocks), composed of sizable blocks, which are spanned by large slabs that constitute the ceiling. The pillars of the porch have a campaniform lotus capital, one of the last times this form appears in Indian architecture. Another temple of this type is the Kaṅkālī Devī shrine at Tigowā, which has more elaborate pillars, provided with the overflowing vase, or the vase-and-foliage (*ghaṭa-pallava*), capital that became the basic north Indian order.

It is the second type of temple that points the way to future developments. It also has a square sanctum, or cella, but instead of a flat roof there is a pyramidal superstructure (*śikhara*). Among the most interesting examples are a brick temple at Bhītargaon and the Viṣṇu temple at Deogarh, built entirely of stone. The pyramidal superstructure of each consists essentially of piled-up cornice moldings of diminishing size, which are decorated primarily with *candraśālā* (ogee arch) ornament derived from the arched windows and doors so frequently found in the centuries immediately before and after Christ. The sanctums of both temples are square in plan, with three sides provided with central offsets (vertical buttress-like projections) that extend from the base of the walls right up to the top of the *śikhara* (spire); the section of the central offset that extends across the wall is conceived in the form of a niche, in which is placed an image. The Deogarh temple is also noteworthy for the large terrace with four corner shrines (now ruined) on which it is placed, prefiguring the quincunx, or *pañcāyatana*, grouping (one structure in each corner and one in the middle) popular in the later period. The doorway sur-

P. Chandra



Carved sandstone doorframe of the Viṣṇu temple at Deogarh, Uttar Pradesh, India, 5th century AD.

round, too, is very elaborate, carved with several bands carrying floral and figural motifs. At the base of the surround are rows of worshippers, and in the crossette (projection at the corner) on top are images of graceful river goddesses.

The Pārvatī Devī temple at Nāchnā-Kutharā, also of this period, is interesting for the covered circumambulatory provided around the sanctum and the large hall in

front. When first discovered, the temple had an entire chamber above the sanctum (which subsequently collapsed). Though provided with a door, there seems to have been no access to it; thus, for all practical purposes it constituted a false story and, aside from a symbolic meaning, served no other purpose than to emphasize the importance of the sanctum. The principle of gaining height not by the superimposition of ornamental cornice moldings with *candraśālā* decoration but by the multiplication of stories, each imitating the story below, also distinguished the later architectural style of southern India.

The great Mahābodhi temple at Bodh Gayā, commemorating the spot where the Buddha attained enlightenment, though burdened with later restorations, is essentially a temple of this period. It has a particularly majestic *śikhara*, decorated with ornamental niches and *candraśālā*s, rising over a square sanctum to a great height.

Along with temples, *stūpa*s continued to be built. These also aspired to height, which was achieved by multiplication and heightening of the supporting terraces and elongating the drum and dome. A good example of this new form is the Dhamekh *stūpa* at Sārnāth. Along more conventional lines, but quite elaborate, are the brick *stūpa*s in Sind, notably a fine example at Mīrpur Khās.

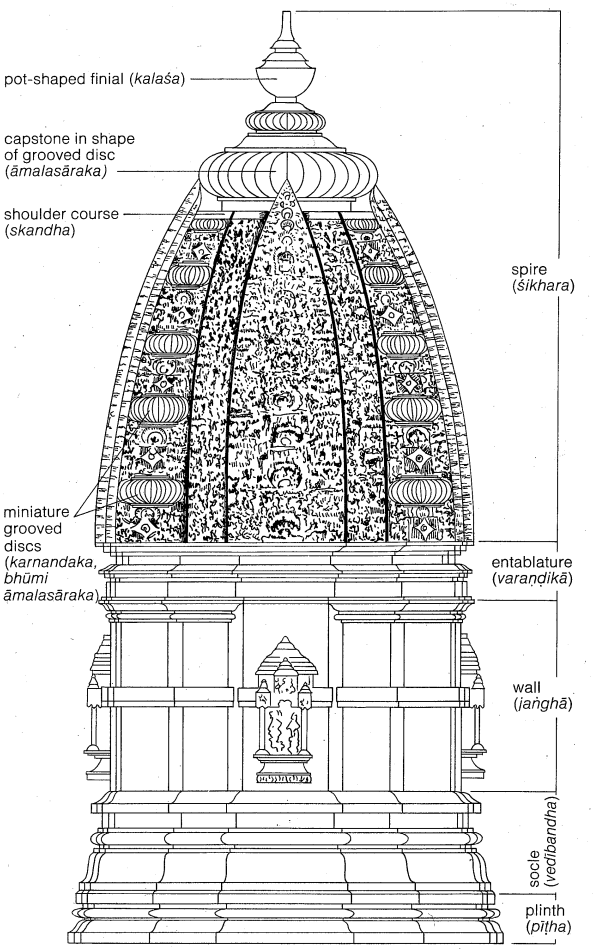
The rock-cut temple and monastery tradition also continued in this period, notably in western India, where the excavations—especially at Ajantā—acquire extreme richness and magnificence. The monasteries are characterized by the introduction of images into some of the cells, so that they partake of the nature of temples instead of being simple residences. Temples with an apsidal plan and barrel-vault roofs, however, soon went out of fashion, and are found very rarely in the subsequent period. The early 5th-century cave temples at Udayagiri, Madhya Pradesh, are similar to the simple flat-roofed temples with a hall and are not descended from ancient traditions as preserved in western India.

Medieval temple architecture. Architectural styles initiated during the 5th and 6th centuries found their fullest expression in the medieval period (particularly from the 9th to the 11th centuries), when great stone temples were built. Two main types can be broadly distinguished, one found generally in northern India, the other in southern India. To these can be added a third type, sharing features of both and found in Karṇāṭa and the Deccan. These three types have been identified by some scholars with the *nāgara, drāviḍa*, and *vesara* classes referred to in some Sanskrit texts, though the actual meaning of these terms is far from clear. In spite of the havoc wrought by the destructive Islāmic invasions, particularly in the Indo-Gangetic Plains, an extremely large number of monuments have survived in almost every other part of India, particularly in the south, and these continue to be discovered and recorded to the present day.

Period of the great stone temples

Medieval temple architecture: North Indian style. North Indian temples generally consist of a sanctum enshrining the main image, usually square in plan and shaped like a hollow cube, and one or more halls (called *maṇḍapas*), aligned along a horizontal axis. The sanctum may or may not have an ambulatory, but it is invariably dark, the only opening being the entrance door. The doorway surrounds are richly decorated with bands of figural, floral, and geometrical ornament and with river-goddess groups at the base. A vestibule (*antarāla*) connects the sanctum to the halls, which are of two broad types: the *gūḍhamaṇḍapa*s, which are enclosed by walls, light and air let in through windows or doors; and open halls, which are provided with balustrades rather than walls and are consequently lighter and airier. The sanctum almost invariably, and the *maṇḍapa*s generally, have *śikhara*s; those on the sanctum, appropriately, are the most dominant in any grouping. Internally, the sanctum has a flat ceiling; the *śikhara* is solid theoretically, though hollow chambers to which there is no access must be left within its body to lessen the weight. The ceilings of the halls, supported by carved pillars, are coffered (decorated with sunken panels) and of extremely rich design.

The sanctum is often set on a raised base, or a plinth (*pīṭha*), above which is a foundation block, or socle



Elevation of a North Indian temple with the *latina* type of superstructure.

(*vedībandha*), decorated with a distinct series of moldings; above the *vedībandha* rise the walls proper (*jaṅghā*), which are capped by a cornice or a series of cornice moldings (*varaṇḍikā*), above which rises the *śikhara*. One, three, and sometimes more projections extend all the way from the base of the temple up the walls to the top of the *śikhara*. The central offset (*bhadra*) is the largest and generally carries an image in a niche; the other projections (*rathas*), too, are often decorated with statuary.

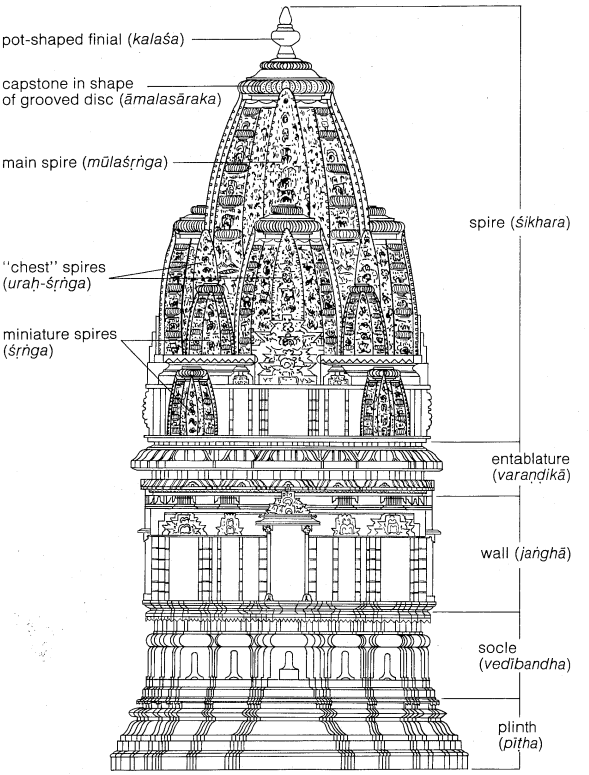
The entire temple complex, including sanctum, halls, and attendant shrines, may be raised on a terrace (*jagatī*), which is sometimes of considerable height and size. The attendant shrines—generally four—are placed at the corners of the terrace, forming a *pañcāyatana*, or quincunx, arrangement that is fairly widespread. The temple complex may be surrounded by a wall with an arched doorway (*toraṇa*).

Śikhara types

The *śikhara* is the most distinctive part of the North Indian temple and provides the basis for the most useful and instructive classification. The two basic types are called *latina* and *phāmsanā*. Curvilinear in outline, the *latina* is composed of a series of superimposed horizontal roof slabs and has offsets called *latā*s. The edges of the *śikhara* are interrupted at intervals with grooved discs, each one demarcating a "story." The surface of the entire *śikhara* is covered with a creeper-like tracery, or interlaced work, composed of diminutive ornamental *candraśālā*s.

The *śikhara* is truncated at the top and capped by a shoulder course (*skandha*), above which is a circular necking (*grīvā*), carrying a large grooved disc called the *āmalasāraka*. On it rests a pot and a crowning finial (*kalaśa*).

Unlike the *latina*, the *phāmsanā śikhara* is rectilinear rather than curvilinear in outline, and it is lower in height. It is composed of horizontal slabs, like the *latina*, but is capped by a bell-shaped member called the *ghaṇṭā*. The



Elevation of a North Indian temple with the *śekharī* type of superstructure.

surface of this type of *śikhara* may have projections, like the *latina śikhara*, and be adorned with a variety of architectural ornament.

From the 10th century onward, the *śekharī* type of spire, an elaboration of the *latina* type, became increasingly popular. In its developed form it consisted of a central *latina* spire (*mūlaśṛṅga*) with one or more rows of half spires added on the sides (*uraḥ-śṛṅga*) and the base strung with miniature spires (*śṛṅga*s). The corners, too, are sometimes filled with quarter spires, the whole mass of carved masonry recalling a mountain with a cluster of subsidiary peaks.

The *latina* and *śekharī* spires are generally found on the sanctum, while the *phāmsanā* and its variants are usually confined to the *maṇḍapas*, or halls. The sanctum spires also have a large and prominent projection in front (*śukanāsā*), generally rising above the vestibule (*antarāla*). These projections are essentially large ogee arches of complex form, which often contain the image of the presiding deity.

The *bhūmija śikhara*

A particularly rich and pleasing variety of North Indian *śikhara*, popular in Mālwa, western India, and northern Deccan, is the *bhūmija* type. It has a central projection on each of the four faces, the quadrants so formed filled with miniature spires in vertical and horizontal rows right up to the top.

Although basically reflecting a homogeneous architectural style, temple architecture in northern India developed a number of distinct regional schools. A detailed elucidation of all has yet to be made, but among the most important are the styles of Orissa, central India, Rājasthān, and Gujarāt. The style of Kashmir is distinct from the rest of northern India in several respects, and hardly any examples of the great schools that flourished in modern Uttar Pradesh, Bihār, and Bengal are left standing. The North Indian style also extended for some time into the Karṇāṭa territory, situated in the southern Deccan, though the architecture of Tamil Nadu was relatively unaffected by it.

Medieval temple architecture: North Indian style of Orissa. The greatest centre of this school is the ancient city of Bhuvaneśvara, in which are concentrated almost 100 examples of the style, both great and small, ranging in date from the 7th to the 13th century. Among the earliest is the Paraśurāmeśvara temple (7th–8th century), with a heavy, stately *latina śikhara*, to which is attached a rectangular *gūḍhamaṇḍapa* with double sloping roofs. The walls are richly carved, but the interiors, as in almost all examples of the style, are left plain. The Mukteśvara temple (10th century), which has a hall with a *phāmsanā* roof, is the product of the most exquisite workmanship. The enclosing wall and the arched entrance, or *toraṇa*, are still present, giving a clear idea of a temple with all its parts fully preserved. The Brahmeśvara temple, which is dated on the basis of an inscription to the mid-10th century, is a *pañcāyatana*, with subsidiary shrines at all of the corners. The most magnificent building, however, is the great Liṅgarāja temple (11th century), an achievement of Orissan architecture in full flower. The *latina* spire soars to a considerable height (over 125 feet [40 metres]); the wall is divided into two horizontal rows, or registers, replete with statuary; and the attached hall is exquisitely and minutely carved. The most famous of all Orissan temples, however, is the colossal building at Konārak, dedicated to Sūrya, the sun god. The temple and its accompanying hall are conceived in the form of a great chariot drawn by horses. The *śikhara* over the sanctum has entirely collapsed; and all that survives are the ruins of the sanctum and the *gūḍhamaṇḍapa*, or enclosed hall, and also a separate dancing hall. Of these, the *gūḍhamaṇḍapa* is now the most conspicuous, its gigantic *phāmsanā śikhara* rising in three stages and adorned with colossal figures of musicians and dancers.

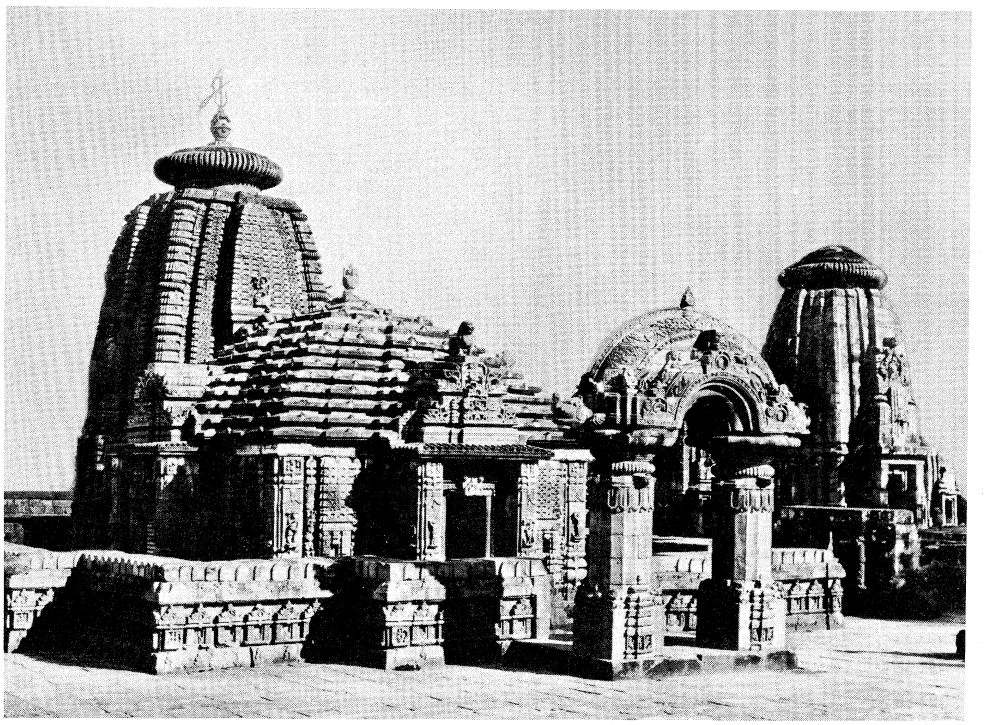
Because the Orissan style usually favours a *latina śikhara* over the sanctum, the *śekharī* spire of the Rājrānī temple (11th century) at Bhuvaneśvara (Bhubaneswar) is quite exceptional. Of particular interest as a late survival of early building traditions is the Vaitāl Deul (8th century), the sanctum of which is rectangular in plan, its *śikhara* imitating a pointed barrel vault. Besides Bhuvaneśvara, important groups of temples are to be found at Khiching and Mukhalingam.

Medieval temple architecture: North Indian style of central India. The area roughly covered by the modern state of Madhya Pradesh was the centre of several vigorous schools of architecture, of which at least four have been identified. The first flourished at Gwalior and adjacent areas (ancient Gopādri); the second in modern Bundelkhand, known in ancient times as Jejākabhukti; the third in the eastern and southeastern parts in the ancient country of Ḍāhala, of which Tripurī, near modern Jabalpur, was the capital; and the fourth in the west, in an area bordering Gujarāt and Rājasthān in the fertile land of Mālava (Mālwa).

The Gwalior temples

The earliest examples in the Gwalior area are a group of small shrines at Naresar, a few miles from Gwalior proper; dating to the 8th century, the shrines have *latina* spires and sparsely ornamented walls. In the 9th century a series of magnificent temples was built, including the Mālā-de at Gyāraspur, the Śiva temples at Mahuā and Indore, and a temple dedicated to an unidentified mother goddess at Barwa-Sāgar. The period appears to have been one of experimentation, a variety of plans and spires having been tried. The Mālā-de temple is an early example of the *śekharī* type in its formative stages; the Indore temple has a star-shaped plan; and the Barwa-Sāgar example has a twin *latina* spire over a rectangular sanctum. The masonry work is of the finest quality and the architectural ornament is crisply carved. (The figural sculptures are few.) The temple at Umrī, with a *latina* spire, is small and exquisitely finished; but the largest and perhaps the finest temple is the Telī-kā-Mandir on Gwalior Fort, rectangular in plan and capped by a pointed barrel vault, recalling once again the survival of ancient roof forms. The walls are decorated with niches (empty at present) topped by tall pediments (triangular gable ornament).

The style of this region became increasingly elaborate from the 10th century, during the supremacy of the Kacchapaghāṭa dynasty. The many examples from this period are distinguished by a low plinth and rich sculptural decoration on the walls. Outstanding among them are the Kākan-maḍh at Suhāniā (1015–35) and the Sās-Bahū temple (completed 1093) in Gwalior Fort. The several



Mukteśvara temple at Bhubaneswar, Orissa, India, late 10th century AD.
P. Chandra

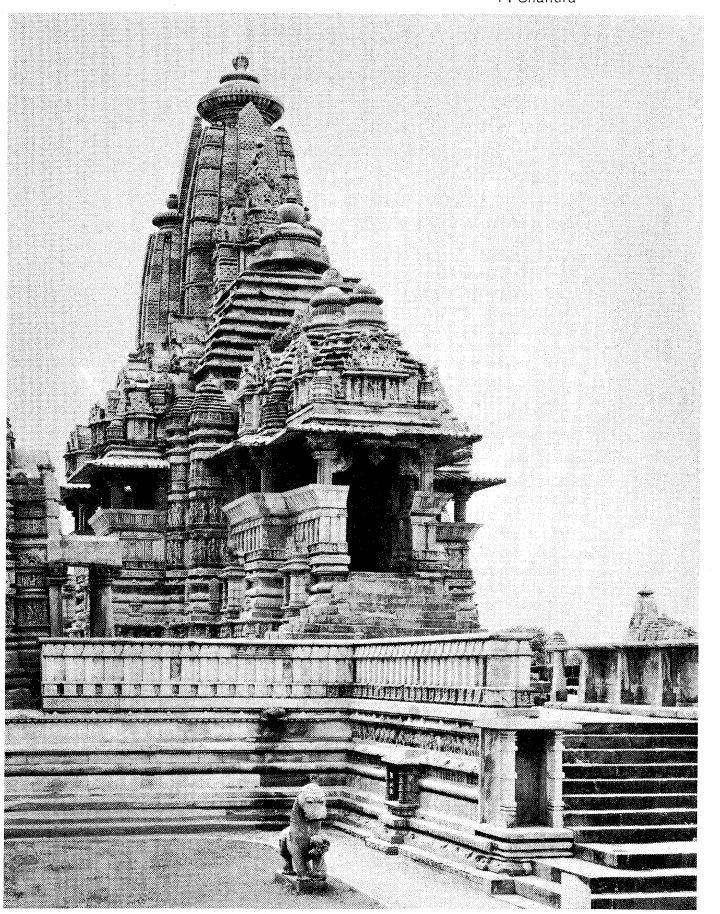
temples at Surwāyā and Kadwāhā, though smaller in size, are distinguished for their extremely rich and elegant workmanship.

The style is best represented by a large group of temples at Khajurāho, the capital of the Candella dynasty, though examples are also to be found in Mahoba and at several other sites in the Jhānsi district of Uttar Pradesh, notably Chāndpur and Dudhai. All of the distinctive characteristics of the fully developed style can be seen in the Lakṣmaṇa temple at Khajurāho (dated 941), which is a *pañcāyatana* placed on a tall terrace enclosed by walls. The sanctum has an ambulatory and, facing it, a series of halls, including the *gūḍhamaṇḍapa*, a porch, and a small intermediate hall. Both the ambulatory and the *gūḍhamaṇḍapa* are provided with lateral, balconied arms, or transepts, which let in light and air. Each hall has its own pyramidal *śikhara*, all skillfully correlated to ascend gradually to the main *śekharī* spire over the sanctum. Extraordinary richness of carving, both in the interior and on the exterior, where the walls carry as many as three rows of sculpture, and a skillful handling of the main spire to suggest ascent are distinguishing features of the style. The largest temple of the group, very similar to the Lakṣmaṇa, is the Kandāriyā Mahādeo; and among the most distinguished are the Viśvanātha and the Pārśvanātha temples. The Dūlādeo temple, which does not have an ambulatory, represents the closing phase of the group and probably belongs to the 12th century.

The Lakṣmaṇa temple

The earliest temples of the Ḍāhala area, dating from the 8th–9th century, are the simple shrines at Bāndhogarh, which consist of a sanctum with *latina* spire and porch. To the 10th century, when the local Kalacuri dynasty was rapidly gaining power, belong the remarkable Śiva temples at Chandrehe and Masaun, the former being circular in plan, with a *latina* spire covered with rich *candraśālā* tracery. The Golā Math at Maihar has the more conventional square sanctum, with a very elegant *latina śikhara*, the walls of which are adorned with two rows of figural sculpture. There must have existed at Gurgī a large number of temples, though all of them now are in total ruin. Judging from a colossal image of Śiva-Pārvatī and a huge entrance, which have somehow survived, the main temple must have been of very great size. Another important site is Amarkantak, where there are a large group of temples, the most important of which is the Karṇa. Although generally of the 11th century, they are quite simple, lacking the rich sculptural decoration so characteristic of the period. By contrast, the Virāṭeśvara temple at Sohāgpur, with an unusually tall and narrow *śekharī* spire, is covered with sculptural ornamentation as rich as that of Khajurāho.

P. Chandra



Lakṣmaṇa temple at Khajurāho, Madhya Pradesh, India, *c.* AD 941.

Temples of the Mālava area

The Mālava region, ruled largely by the Paramāra dynasty, appears to have been the first to develop the *bhūmija* type of *śikhara* (10th century). The finest and most representative group of these structures is at Un. Though, unfortunately, they are considerably damaged, judging from the remains, they must have been very elegant structures. The best preserved and easily the finest *bhūmija* temple is the Udayeśvara (1059–82), situated at Udaipur in Madhya Pradesh. The *śikhara*, based on a stellate plan, is divided into quadrants by four *latās*, or offsets, each one of which has five rows of aediculae. The large hall has three entrance porches, one to the front and two to the sides, and walls that are richly carved. The whole complex, including seven subsidiary shrines, is placed on a broad, tall platform. The Siddheśvara temple at Nemāwar (early 12th century) is even larger than the Udayeśvara, though the proportions are not as well balanced and the quality of the carving is inferior. Structures in the *bhūmija* manner continued to be made in Mālava up to the 15th century; the Malvai temple at Alīrājpur is a good example of the late phase.

From Mālava, the *bhūmija* style spread to the neighbouring regions. To the north in Rājasthān, the Mahānāleśvara temple at Menāl (*c.* 11th century), the Sun temple at Jhālrapātan (11th century), the Śiva temple at Rāmgarh (12th century), and the Uṇḍeśvara temple (12th century) at Bījoliān are important examples. To the west, in Gujarāt, are temples at Limkheda and Sarnāl of the 11th and 12th centuries. The style was particularly favoured in Mahārāshtra, to the south. Among surviving examples, the most impressive is the Ambarnāth temple near Bombay (11th century); Balsāṇe and Sinnar also have pleasing temples. The style continued up to the 16th century, many examples having been found in north Deccan and Berār. The *bhūmija* style also spread to the east of Mālava; the Bhāṇḍ Dewal at Arang (11th century), for example, is a Ḍāhala adaptation.

Medieval temple architecture: North Indian style of Rājasthān. A group of temples at Osiān, dating to about the 8th century, represents adequately the opening phases of medieval temple architecture in Rājasthān. They stand on high terraces and consist of a sanctum, a hall, and a porch. The sanctum is generally square and has a *latina* spire. The walls, with one central and two subsidiary projections, are decorated with sculpture, often placed in niches with tall pediments. The halls are generally of the open variety, provided with balustrades rather than walls, so that the interiors are well lit. The surrounds of the doorway sanctum are quite elaborate, with four or five bands of decoration and the usual river-goddess groups at the base. The pillars, with *ghaṭa-pallava* (vase-and-foliage) capitals, are also decorated, richness of sculpture and architectural elaboration being a characteristic of this group of monuments. The Mahāvīra temple, which is the largest, belongs to the 8th century, though renovated in later times, when the *toraṇa* (gateway) and the *śikhara* were added. Other important temples are Harihara Nos. 1, 2, and 3 and two temples dedicated to Viṣṇu. The ruined Harshat Mātā temple at Ābānerī, of a slightly later date (*c.* 800), was erected on three stepped terraces of great size and is remarkable for the exquisite quality of the carving. Some of the finest temples of the style date from the 10th century, the most important of which are the Ghaṭeśvara temple at Bāḍolī and the Ambikā Mātā temple at Jagat. The simple but beautiful Bāḍolī temple consists of a sanctum with a *latina* superstructure and an open hall with six pillars and two pilasters (columns that project a third of their width or less from the wall) supporting a *phāmsanā* spire. Only the central projections of the sanctum walls are decorated with niches containing sculpture. A large open hall was built in front of the temple at a later date. The Ambikā-Mātā temple at Jagat, of the mid-10th century, is exceptionally fine. It consists of a sanctum, a *gūḍhamaṇḍapa*, or enclosed hall, and a parapeted porch with projecting eaves. The walls of the sanctum and the hall are covered with fine sculpture, the superstructures being of the *śekharī* and the *phāmsanā* types.

Temples, too numerous to mention, dating from the 10th and—to an even greater extent—the 11th century onward, are found throughout Rājasthān. The styles of Rājasthān and neighbouring Gujarāt during these centuries grew closer and closer together until the differences between them were gradually obliterated. This coalescence resulted in the emergence of a composite style found throughout Gujarāt and Rājasthān. Temples situated in the two areas are discussed separately here, but this is for the sake of convenience and does not signify any real stylistic difference.

Growing resemblance between styles of Rājasthān and Gujarāt

The temples at Kirāḍu in Rājasthān, dating from the late 10th and 11th centuries, are early examples of the style shared by Rājasthān and Gujarāt. The Someśvara temple (*c.* 1020) is the most important and clearly shows the movement toward increasing elaboration and ornamentation. Each of the constituent parts became more complex; the moldings of the plinth, for example, are multiplied to include bands of elephants, horses, and soldiers. The walls are covered with sculpture, and the spire is of the rich *śekharī* type. Situated in Rājasthān, but again in the composite style, are the extraordinarily sumptuous temples known as the Vimala Vasahī (1031) and the Lūṇa Vasahī (1230) at Mt. Ābū. The Vimala Vasahī consists of a sanctum, a *gūḍhamaṇḍapa*, and a magnificent assembly hall added in mid-12th century. The plain, uncarved exterior walls of the rectangular enclosure of the temple have on the inside rows of cells containing images of divinities. The interiors are very richly carved, the coffered ceilings loaded with a wealth of detail. The Lūṇa Vasahī is even more elaborate, though the quality of the work had begun to decline perceptibly.

Traditional architecture continued even after the Islāmic invasions, particularly during the reign of Rāṇā Kumbhā of Mewār (*c.* 1430–69). During this period, the tall nine-storied Kīrttistambha and other temples at Chitor and also the great Chaumukha temple at Ranakpur (1438) were built.

Medieval temple architecture: North Indian style of Gujarāt. Gujarāt was the home of one of the richest regional styles of northern India. A temple at Gop (*c.* 600), with a tall terrace and a cylindrical sanctum with high walls capped by a *phāmsanā* roof, and other temples in Saurāshṭra show the formative phases of the style. Its distinctive features are clear in an interesting group of temples from Roḍā (*c.* 8th century). The sanctum is square in plan and has *latina* spires that are weighty and majestic. The walls are relatively plain, with niches, housing images, provided only on the central projection. The masonry work is exceptionally good, a characteristic of Gujarāt architecture throughout its history. The Rāṇakdevī temple at Wadhwān, of the early 10th century, is also characterized by plain walls and a *latina* spire, while the Śiva temple at Kerākoṭ has a *śekharī* spire and also a *gūḍhamaṇḍapa*. The great Sun Temple at Modhera, datable to the early years of the 11th century, represents a fully developed Gujarāt style of great magnificence. The temple consists of a sanctum (now in ruins), a *gūḍhamaṇḍapa*, an open hall of extraordinary richness, and an arched entrance in front of which was the great tank. The Navalakhā temple at Sejakpur continued this tradition. The Rudramāla at Siddhapur, the most magnificent temple of the 12th century, is now in a much ruined condition, with only the *toraṇa* (gateway) and some subsidiary structures remaining. Successively damaged and rebuilt, the Somanātha at Prabhāsa Patan was the most famous temple of Gujarāt, its best known structure dating from time of Kumārapāla (mid-12th century). It has been now dismantled, but a great temple built at the site in recent years testifies to the survival of ancient traditions in modern Gujarāt.

The temple at Gop

The hills of Satrunjaya and Girnār house veritable temple cities. Most of the shrines, which are of late date, are picturesque but otherwise of little significance. With the Islāmic conquest, the Gujarāt architect adapted his considerable skills to meet the needs of a patron of different religion and quickly produced a totally successful Indian version of Islāmic architecture.

Medieval temple architecture: North Indian style of Karṇāṭa. The North Indian style was largely confined to

India above the Vindhyas, though for a short period it also flourished in a region of southern India known as Karṇāṭa from ancient times and now largely part of Mysore state. Here, temples of the northern and the southern styles are found next to each other, notably at Aihole and Pattadkal. The earliest appears to be the Lāḍh Khān at Aihole, closely related to the 5th-century temple at Nāchnā Kutharā in northern India. The northern style was also cultivated at Pattadkal, where the most important examples are the Kāśīviśvanātha, the Galaganātha, and the Pāpanātha. Ālampur, now in Andhra Pradesh, has eight temples of the northern style with *latina* spires. These belong to the late 7th and early 8th centuries and are the finest and among the last examples of the northern style in southern India.

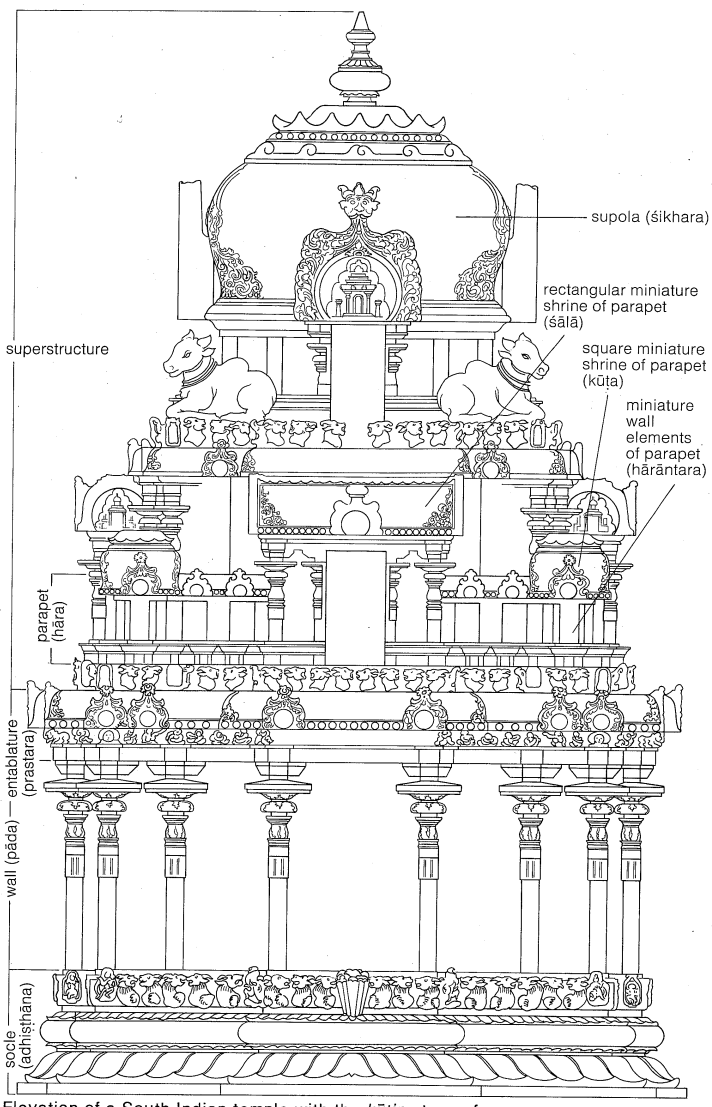
Distinct architectural style of Kashmir

Medieval temple architecture: North Indian style of Kashmir. The architectural style of the Kashmir region is quite distinct: unlike other regions, in which the sanctum usually has a *latina* or *śekharī* spire, the roof of the Kashmir sanctum is of the *phāmsanā* type, with eaves raised in two stages. The greatest example to survive is the ruined Sun Temple at Mārtaṇḍ (mid-8th century), which, though its *śikhara* is missing, gives a good idea of the characteristic features of the style. The temple is placed in a rectangular court enclosed by a series of columns. Access to the court is through an imposing entrance hall, the walls of which have doorways with gabled pediments and a trefoil (shaped like a trifoliate leaf) recess. The Avantisvāmī temple of the mid-9th century, now quite ruined, must have been similar, though much more richly ornamented. The style continued up to the 12th century; the Rilhaṇeśvara temple at Pāndrenṭhan is a comparatively well-preserved example of this period.

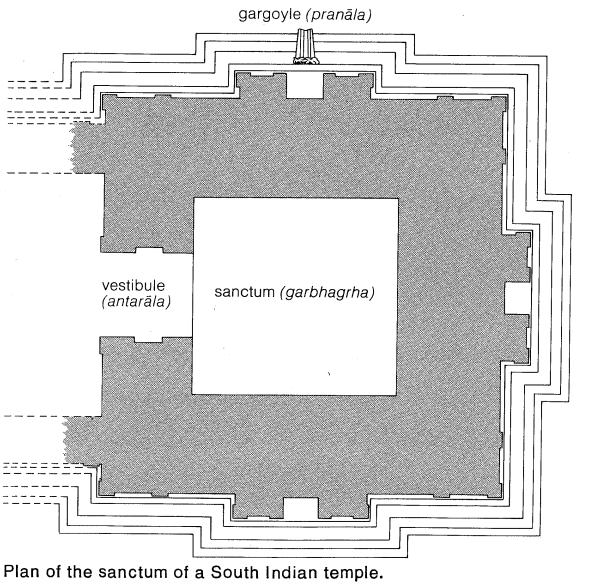
Medieval temple architecture: South Indian style. The home of the South Indian style, sometimes called the drāviḍa style, appears to be the modern state of Tamil Nadu; examples, however, are found all over southern India, particularly in the adjoining regions of Karṇāṭa and Andhradeśa, now largely covered by the modern states of Mysore and Andhra Pradesh, respectively. Both Andhradeśa and Karṇāṭa developed variants, particularly Karṇāṭa, which evolved a distinct manner, basically South Indian but with some features of North Indian origin. The Karṇāṭa style extended northward into Mahārāshtra, where the Kailāsa temple at Ellora is the most famous example.

A typical South Indian temple consists of a hall and a square sanctum that has a superstructure of the *kūṭina* type. Pyramidal in form, the *kūṭina* spire consists of stepped stories, each of which simulates the main story and is conceived as having its own "wall" enclosed by a parapet. The parapet itself is composed of miniature shrines strung together: square ones (called *kūṭa*s) at the corners and rectangular ones with barrel-vault roofs (called *śālā*s) in the centre, the space between them connected by miniature wall elements called *hārāntara*s. (Conspicuous in the early temples, these stepped stories of the superstructure with their parapets became more and more ornamental, so that in the course of time they evolved into more or less decorative bands around the pyramidal superstructure.) On top of the stepped structure is a necking that supports a solid dome, or cupola (instead of the North Indian grooved disc), which in turn is crowned by a pot and finial. The walls of the sanctum rise above a series of moldings, constituting the foundation block, or socle (*adhiṣṭhāna*), that differ from North Indian temples; and the surface of the walls does not have the prominent offsets seen in North Indian temples but is instead divided by pilasters. In the Karṇāṭa version, particularly from the late 10th century onward (sometimes called the *vesara* style), this arrangement of the superstructure is loaded with decoration, thus considerably obscuring the component elements. At the same time, these elements—particularly the central offset with its subsidiaries that carry *candraśālā* motifs—are so manipulated that they tend to form distinct vertical bands, in this respect closely recalling the *śikhara*s of northern India.

The design of the hall-temple roofed by a barrel vault, popular in the centuries before and after Christ, was adopted in southern India for the great entrance buildings, or *gopura*s, that give access to the sacred enclosures in which the temples stand. Relatively small and inconspicuous in the early examples, they had, by the mid-12th century, outstripped the main temple in size.



Elevation of a South Indian temple with the *kūṭina* type of superstructure.



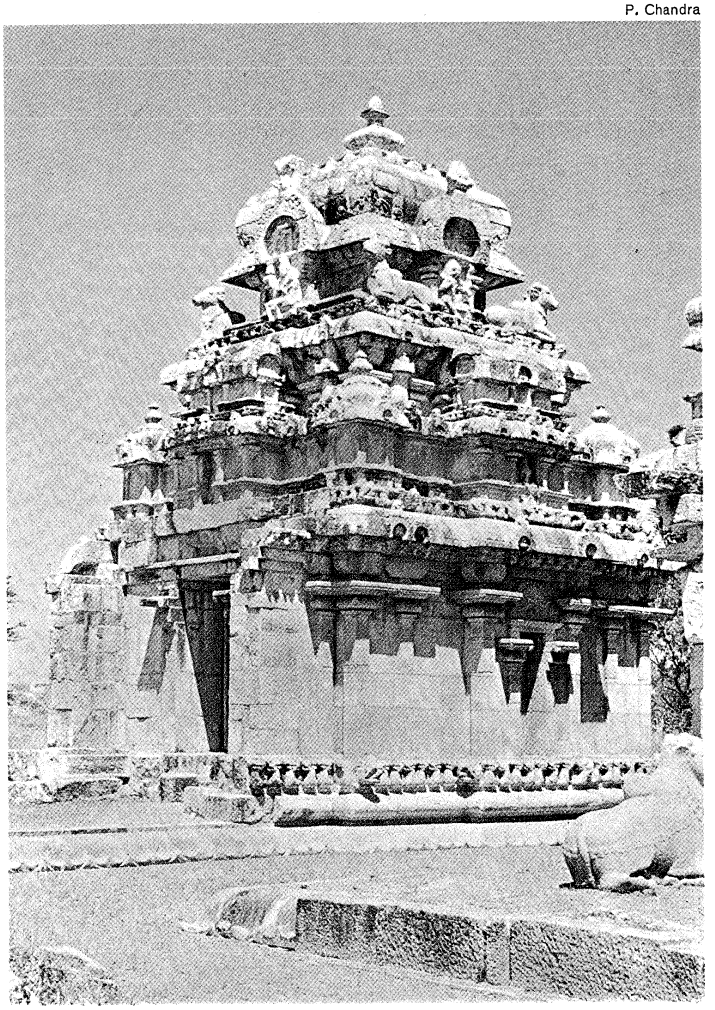
Plan of the sanctum of a South Indian temple.

Medieval temple architecture: South Indian style of Tamil Nadu (7th–18th century). The early phase, which, broadly speaking, coincided with the political supremacy of the Pallava dynasty (*c.* 650–893), is best represented by the important monuments at Mahābalipuram. Besides a fine group of small cave temples (early 7th century), among the earliest examples of their type in southern India, there are here several monolithic temples carved out of the rock, the largest of which is the massive three-storied Dharmarāja-ratha (*c.* 650). The finest temple at this site and of this period is an elegant complex of three shrines called the Shore Temple (*c.* 700), not cut out of rock but built of stone. The Tālapurīśvara temple at Panamalai is another excellent example. The capital city of Kānchipuram also possesses some fine temples—for example, the Kailāsanātha (dating a little later than the Shore Temple), with its stately superstructure and subsidiary shrines attached to the walls. The enclosure wall has a series of small shrines on all sides and a small *gopura*. Another splendid temple at Kānchipuram is the Vaikuṇṭha Perumāl (mid-8th century), which has an interesting arrangement of three sanctums, one above the other, encased within the body of the superstructure.

The Shore Temple

The 9th century marked a fresh movement in the South Indian style, revealed in several small, simple, but most elegant temples set up during the ascendancy of the Cōḻa and other contemporary dynasties. Most important of a large number of unpretentious and beautiful shrines that dot the Tamil countryside are the Vijayālaya Cōḻīśvara temple at Nārttāmalai (mid-9th century), with its circular sanctum, spherical cupola, and massive, plain walls; the twin shrines called Agastyīśvara and Cōḻīśvara, at Kīḻaiyūr (late 9th century); and the splendid group of two temples (originally three) known as the Mūvarkovil, at Koḍumbāḷūr (*c.* 875).

P. Chandra



Mūvarkovil at Kodumbālūr, Tamil Nadu, India, *c.* AD 875.

These simple beginnings led rapidly (in about a century) to the mightiest of all temples in the South Indian style, the Bṛhadīśvara, or Rājarājeśvara, temple, built at the Cōḻa capital of Thanjāvūr. A royal dedication of Rājarāja I, the temple was begun around 1003 and completed about seven years later. The main walls are raised in two stories, above which the superstructure rises to a height of 190 feet (60 metres). It has 16 stories, each of which consists of a wall with a parapet of shrines carved in relatively low relief. The great temple at Gaṅgaikoṇḍacōḻapuram, built (1030–40) by the Cōḻa king Rājendra I, is somewhat smaller than the Bṛhadīśvara; but the constituent elements of its superstructure, whose outline is concave, are carved in bolder relief, giving the whole a rather emphatic plasticity. The Airāvateśvara (1146–73) and Kampahareśvara (1178–1223) temples at Dārāsuram and Tribhuvanam follow the tradition of the 11th century but are smaller and considerably more ornate. They bring to a close a great phase of South Indian architecture extending from the 11th to the 13th century.

From the middle of the 12th century onward, the *gopuras*, or entrance buildings, to temple enclosures began to be greatly emphasized. They are extremely large and elaborately decorated with sculpture, quite dominating the architectural ensemble. Their construction is similar to that of the main temple except that they are rectangular in plan and capped by a barrel vault rather than a cupola, and only the base is of stone, the superstructure being made of brick and plaster. Among the finest examples are the Sundara Pāndya *gopura* (13th century) of the Jambukeśvara temple at Tiruchchirāppalli and the *gopuras* of a great Śiva temple at Chidambaram, built largely in the 12th–13th century. Even larger *gopuras*, if not of such fine quality, continued to be built up to the 17th century. Such great emphasis was placed on the construction of *gopuras* that enclosure walls, which were not really necessary, were especially built to justify their erection. In the course of time several walls and *gopuras* were successively built, each enclosing the other so that at the present day one often has to pass through a succession of walls with their *gopuras* before reaching the main shrine. A particularly interesting example is the Ranganātha temple at Srīrangam, which has seven enclosure walls and numerous *gopuras*, halls, and temples constructed in the course of several centuries. The *gopuras* of the Mīnākṣī temple at Madurai are also good representative examples of this period.

In addition to the *gopuras*, temples also continued to be built. Although they never achieved colossal size, they are often of very fine workmanship. The Subrahmaṇya temple of the 17th century, built in the compound of the Bṛhadīśvara temple at Thanjāvūr, indicates the vitality of architectural traditions even at this late date.

Medieval temple architecture: South Indian style of Karṇāṭa. The early phase, as in Tamil Nadu, opens with the rock-cut cave temples. Of the elaborate and richly sculptured group at Bādāmi, one cave temple is dated 578, and two cave temples at Aihole are early 8th century. Among structural temples built during the rule of the Cālukyas of Bādāmi are examples in the North Indian style; but, because the Karṇāṭa region was more receptive to southern influences, there are a large number of examples that are basically South Indian with only a few North Indian elements. The Durgā temple (*c.* 7th century) at Aihole is apsidal in plan, echoing early architectural traditions; the northern *latina śikhara* is in all probability a later addition. The Mālegitti Śivālaya temple at Bādāmi (early 8th century), consisting of a sanctum, a hall with a parapet of *śālās* and *kūṭas* (rectangular and square miniature shrines), and an open porch, is similar to examples in Tamil Nadu. The Virūpākṣa at Pattadkal (*c.* 733–746) is the most imposing and elaborate temple in the South Indian manner. It is placed within an enclosure, to which access is through a *gopura*; and the superstructure, consisting of four stories, has a projection in the front, a feature inspired by the prominent projections, or *śukanāsā*, of North Indian temples. Belonging to the 9th century is the triple shrine (the three sanctums sharing the same *maṇḍapa*, or hall) at Kambaḍahalli and the

Rock-cut cave temples of South India

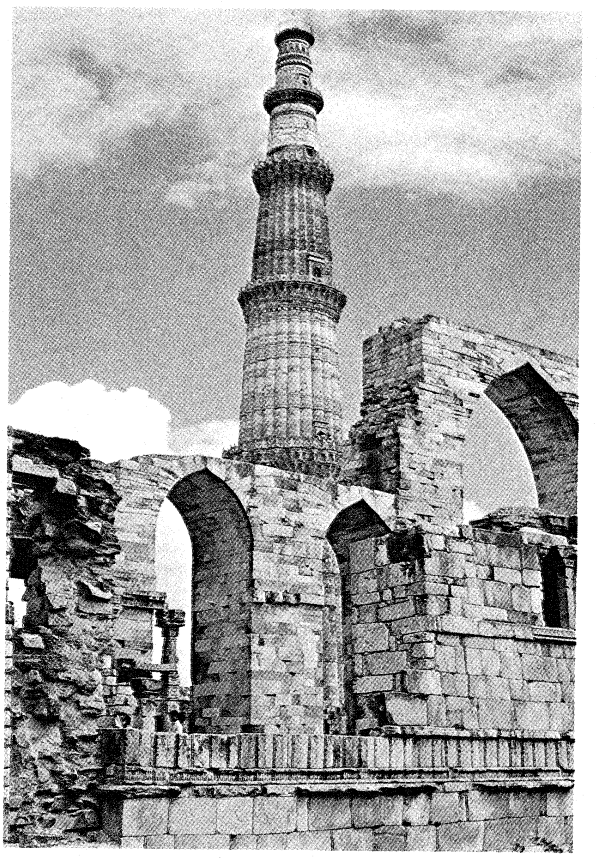
extremely refined and elaborately carved Bhoganandīśvara temple at Nandi. The Chāvuṇḍarāyabasti (*c.* 982–995) at Śravaṇa-Beḷgoḷa is also an impressive building, with an elegant superstructure of three stories.

With the 10th century, the Karṇāṭa idiom begins to show an increasing individuality that culminates in the distinctive style of the 12th century and later. The Kalleśvara temple at Kukkanūr (late 10th century) and a large Jaina temple at Lakkundi (*c.* 1050–1100) clearly demonstrate the transition. The superstructures, though basically of the South Indian type, have offsets and recesses that tend to emphasize a vertical, upward movement. The Lakkundi temple is also the first to be built of chloritic schist, which is the favoured material of the later period and which lends itself easily to elaborate sculptural ornamentation. With the Mahādevā temple at Ittagi (*c.* 1112) the transition is complete, the extremely rich and profuse decoration characteristic of this shrine being found in all work that follows. Dating from the reign of the Hoysaḷa dynasty (*c.* 1141) is a twin Hoysaḷeśvara temple at Halebīd, the capital city. The sanctums are stellate in form but lack their original superstructures. The pillars of the interior are lathe-turned in a variety of fanciful shapes; and the exterior is almost totally covered with sculpture, the walls with the usual complement of images; the base, or socle, with several bands of ornamental motifs and a narrative relief. Among other temples of this style, the most important are the Chenna Keśava temple at Belūr (1117), the Amṛteśvara temple at Amritpur (1196), and the Keśava temple at Somnāthpur (1268).

Medieval temple architecture: South Indian style of Mahārāshtra, Andhradeśa, and Kerala. The traditions of cave architecture are stronger in Mahārāshtra than in any other part of India; there, great shrines were cut out of rock right up to the 9th century AD and even later. Of those belonging to the early phase, the most remarkable is a temple at Elephanta (early 6th century); equally impressive are numerous temples at Ellora (6th–9th centuries). The Karṇāṭa version of the South Indian style extended northward into Mahārāshtra, where the Kailāsa temple at Ellora, erected in the reign of the Rāṣṭrakūṭā Kṛṣṇa I (8th century), is its most stupendous achievement. The entire temple is carved out of rock and is over 100 feet (30 metres) high. It is placed in a courtyard, the three sides of which are carved with cells filled with images; the front wall has an entrance *gopura*. The tall base, or plinth, is decorated with groups of large elephants and griffins, and the superstructure rises in four stories. Groups of important temples in the southern style are also found in the Andhra country, notably at Biccavolu, ranging in date from the 9th to the 11th centuries. The 13th-century temples at Palampet are the counterparts of the elaborate Karṇāṭa style of the same period, but without its overpowering elaboration. The temples of Kerala represent an adaptation of the South Indian style to the great main fall of this region and are provided with heavy sloping roofs of stone that imitate timber originals required for draining away the water.

Strong traditions of cave architecture in Mahārāshtra

Islāmic architecture in India: period of the Delhi and provincial sultanates. Although the province of Sind was captured by the Arabs as early as 712, the earliest examples of Islāmic architecture to survive in the subcontinent date from the closing years of the 12th century; they are located at Delhi, the main seat of Muslim power throughout the centuries. The Qūwat-ul-Islām mosque (completed 1196), consisting of cloisters around a courtyard with the sanctuary to the west, was built from the remains of demolished temples. In 1198 an arched facade (*maqṣūrah*) was built in front to give the building an Islāmic aspect, but its rich floral decoration and corbelled (supported by brackets projecting from the wall) arches are Indian in character. The Quṭb Mīnār, a tall (288 feet high), fluted tower provided with balconies, stood outside this mosque. The Aṛhāi-dīn-kā-jhompṛā mosque (*c.* 1119), built at Ajmer, was similar to the Delhi mosque, the *maqṣūrah* consisting of engrailed (sides ornamented with several arcs) corbel arches decorated with greater restraint than the Quṭb example. The earliest Islāmic tomb



Quṭb Mīnār and the Qūwat-ul-Islām mosque at Delhi, *c.* AD 1196.
P. Chandra

to survive is the Sultān Gharī, built in 1231, but the finest is the tomb of Iltutmish, who ruled from 1211 to 1236. The interior, covered with Arabic inscriptions, in its richness displays a strong Indian quality. The first use of the true arch in India is found in the ruined tomb of Balban (died 1287). From 1296 to 1316, ʿAlāʾud-Dīn Khaljī attempted to expand the Qūwat-ul-Islām mosque, already enlarged in 1230, to three times its size; but he was unable to complete the work. All that has survived is the Alai Darwāzah, a beautiful entrance.

In contrast to this early phase, the style of the 14th century at Delhi, ushered in by the Tughluq dynasty, is impoverished and austere. The buildings, with a few exceptions, are made of coarse rubble masonry and overlaid with plaster. The tomb of Ghiyās-ud-Dīn Tughluq (*c.* 1320–25), placed in a little fortress, has sloping walls faced with panels of stone and marble. Also to be ascribed to his reign is the magnificent tomb of Shāh Rukn-e ʿĀlam at Multān in Pakistan, built of brick and faced with exquisite tile work. The Koṭla Fīrūz Shāh (1354–70), with its mosques, palaces, and tombs, is now in ruins but represents the major building activity of Fīrūz Shāh, who took a great interest in architecture. Many mosques and tombs of this period and of the 15th century are found in Delhi and its environs; the most notable are the Begampur and Khiṛkī mosques and an octagonal tomb of Khān-e Jahān Tilangānī. In the early 16th century, Shēr Shāh Sūr refined upon this style, the Qalʿah-e Kuhnah Masjid and his tomb at Sasarām (*c.* 1540) being the finest of a series of distinguished works created during his reign.

14th-century style impoverished and austere

The provinces, which gradually became independent sultanates, did not lag behind in architectural activity. In West Bengal, at Pandua, is the immense Ādīna Masjid (1364–69), which utilized remains of Indian temples. In Jaunpur, Uttar Pradesh, are a group of elegant mosques, notably the Aṭalā Masjid (1377–1408) and the Jāmiʿ Masjid (*c.* 1458–79), characterized by *maqṣūrah*s that have the aspect of imposing gateways. The sultans of Mālwa built elegant structures at Māndu and at Chanderi in the middle of the 15th century. The sultanate of Gu-



Tomb and palace of Fīrūz Shāh (Hawz-e Khāss) at Delhi, *c.* AD 1380.
P. Chandra

jarāt is notable for its great contribution to Islāmic architecture in India. The style, which is basically indigenous, reinterprets foreign influences with great resourcefulness and confidence, producing works notable for their integrity and unity. The city of Ahmadābād (Ahmedabad) is full of elegant buildings; the Jāmiʿ Masjid (*c.* 1424), for example, is a masterly exposition of the style. Fine examples dating from the second half of the 15th century are the small but exquisite mosques of Muḥāfiz Khān (1492) and Rānī Sabra'i (1514) at Amadābād and the handsome Jāmiʿ Masjid at the city of Chāmpāner.

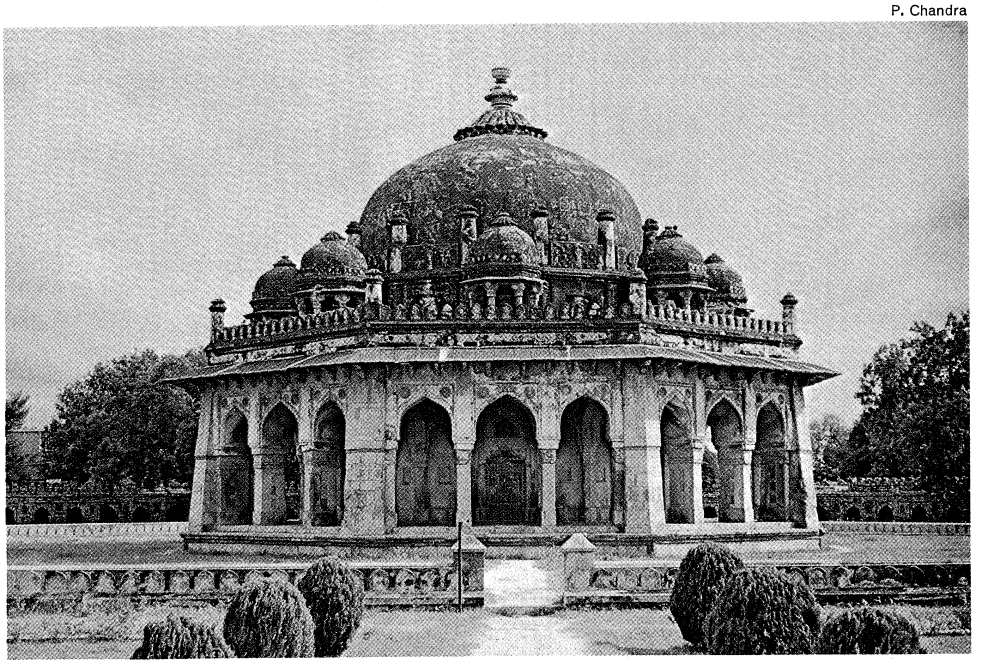
The Deccan was another great centre, but in contrast to Gujarāt it took little from the indigenous building traditions. Among the earliest works is the Jāmiʿ Masjid at Gulbarga (1367), with its extraordinary cloisters consisting of wide arches on low piers, producing a most solemn effect. The city of Bīdar possesses many remains, including a remarkable series of 12 tombs, the most elaborate of which is that of ʿAlā-ud-Dīn Aḥmad Bahmanī (died 1457), which has extremely fine decorations in coloured tile. Some of the finest examples of Islāmic architecture in the Deccan, however, are in Bijāpur. The most important buildings of this city are the great Jāmiʿ Masjid (begun in 1558) with its superb arched cloisters; the ornate Ibrāhīm Rawẓa; and the Gōl Gunbad (built by Muḥammad ʿĀdil Shāh), a tomb of exceptional size and grandeur, with one of the largest domes in existence.

The Hindu kingdoms that managed to retain varying degrees of independence during the period of Islāmic supremacy also produced important works. These naturally bore the imprint of what survived of traditional Indian architecture to a greater extent than did those monuments patronized by Muslims. Among the Hindu structures of this period are the extensive series of palaces, all in ruin, built by Rāṇā Kumbhā (*c.* 1430–69) at Chitor, and the superb Mān Mandir palace at Gwalior (1486–1516), a rich and magnificent work that considerably influenced the development of Mughal architecture at Fatehpur Sīkrī.

Islāmic architecture in India: Mughal style. The advent of the Mughal dynasty marks a striking revival of Islāmic architecture in northern India: Persian, Indian, and the various provincial styles were successfully fused to produce works of unusual refinement and quality. The tomb of Humāyūn, begun in 1564, inaugurates the new style. Built entirely of red sandstone and marble, it shows

Striking revival of Islāmic architecture

P. Chandra

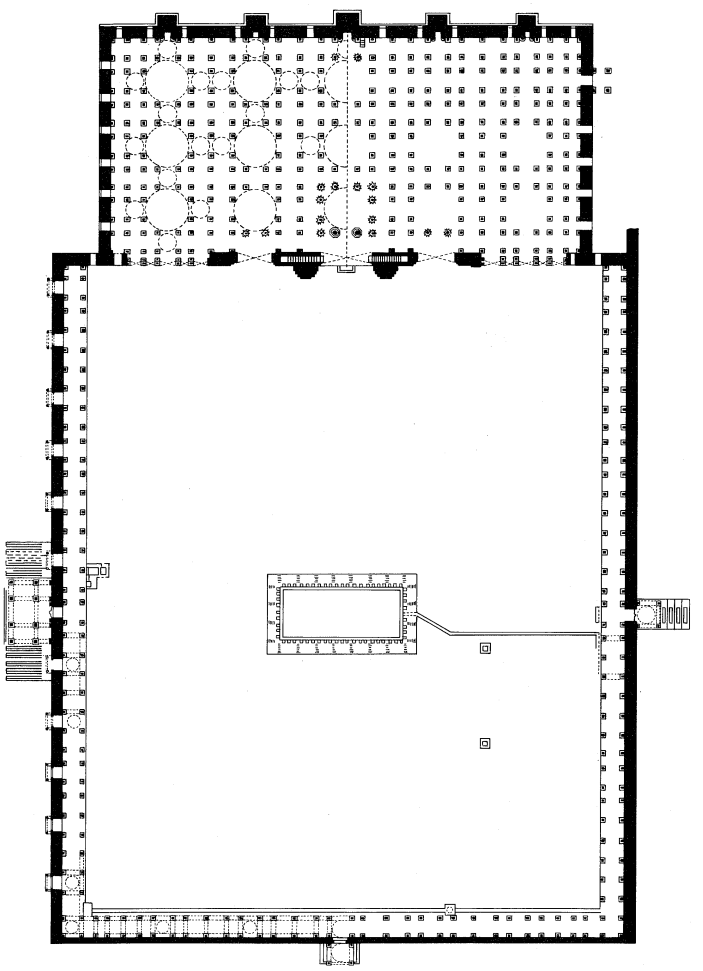


Tomb of 'Isā Khān at Delhi, AD 1547.

considerable Persian influence. The great fort at Āgra (1565–74) and the city of Fatehpur Sīkri (1569–74) represent the building activities of the emperor Akbar. The former has the massive so-called Delhi gate (1566) and lengthy and immense walls carefully designed and faced with dressed stone throughout. The most important achievements, however, are to be found at Fatehpur Sīkri; the Jāmiʿ Masjid (1571), with the colossal gateway known as the Baland Darwāzah, for example, is one of the finest mosques of the Mughal period. Other notable buildings include the palace of Jodhā Bāī, which has a strongly indigenous aspect; the exquisitely carved Turkish Sultānā's house; the Panch-Maḥal; the Dīvān-e ʿĀmm; and the so-called hall of private audience. Most of the buildings are of trabeate (lintel, or horizontal beam) construction, arches being used very sparingly. The tomb of the emperor, at Sikandarā, near Āgra, of vast proportions and unique design, consists of a series of stepped terraces on which rests a centotaph (a tomb erected in honour of a person whose body lies buried below). Begun about 1602, it was completed in 1613, during the reign of Akbar's son Jahāngīr. Architectural undertakings in this emperor's reign were not very ambitious, but there are fine buildings, chiefly at Lahore. The tomb of his father-in-law Iʿtimād-ud-Dawla, at Āgra, is small but of exquisite workmanship, built entirely of delicately inlaid marble. The reign of Shāh Jahān (1628–58) is as remarkable for its architectural achievements as was that of Akbar. He built the great Red Fort at Delhi (1639–48), with its dazzling hall of public audience, the flat roof of which rests on rows of columns and pointed, or cusped, arches, and the Jāmiʿ Masjid (1650–56), which is among the finest mosques in India. But it is the Tāj Mahal (*c.* 1632–49), built as a tomb for Queen Mumtāz Maḥal, which is the greatest masterpiece of his reign. All the resources of the empire were put into its construction. In addition to the mausoleum proper, the complex included a wide variety of accessory buildings of great beauty. The marble mausoleum rises up from a tall terrace (at the four corners of which are elegant towers, or minārs) and is crowned by a graceful dome. Other notable buildings of the reign of Shāh Jahān include the Motī Masjid (*c.* 1648–55) and the Jāmiʿ Masjid at Āgra (1548–55).



Buildings at Fatehpur Sīkri, Uttar Pradesh, India, *c.* AD 1571.
P. Chandra



Plan of the Jāmiʿ Masjid, Ahmadābād, Gujarāt.

Architectural monuments of the reign of Aurangzeb

Architectural monuments of the reign of Aurangzeb represent a distinct decline; the tomb of Rābīʿah Begam at Aurangābād, for example (1679), is a poor copy of the Tāj Mahal. The royal mosque at Lahore (1673–74) is of much better quality, retaining the grandeur and dignity of earlier work; and the Motī Masjid at Delhi (1659–60) possesses much of the early refinement and delicacy. The tomb of Ṣafdar Jang at Delhi (*c.* 1754) was among the last important works to be produced under the Mughal dynasty and had already lost the coherence and balance characteristic of mature Mughal architecture.

European traditions and the modern period. Buildings imitating contemporary styles of European architecture, often mixed with a strong provincial flavour, were known in India from at least the 16th century. Some work was of considerable merit, particularly the baroque architecture of the Portuguese colony of Goa, where splendid buildings were erected in the second half of the 16th century. Among the most famous of these to survive is the church of Bom Jesus, which was begun in 1594 and completed in 1605.

The 18th and 19th centuries witnessed the erection of several buildings deeply indebted to neoclassic styles; these buildings were imitated by Indian patrons, particularly in areas under European rule or influence. Subsequently, attempts were made by the British, with varying degrees of success, to engraft the neo-Gothic and also the neo-Saracenic styles onto Indian architectural tradition. At the same time, buildings in the great Indian metropolises come under increasing European influence; the resulting hybrid styles gradually found their way into cities in the interior. In recent years an attempt has been made to grapple with the problems of climate and function, particularly in connection with urban development. The influence of the Swiss architect Le Corbusier, who worked on the great Chandīgarh project, involving the construction of a new capital for Punjab, in the early 1950s, and that of other American and European masters has led to a modern architectural movement of great vitality, which is in the process of adapting itself to local needs and traditions.



Church of Bom Jesus at Velha Goa, India, AD 1594–1605.
P. Chandra

Sculpture. On the Indian subcontinent, sculpture seems to have been the favoured medium of artistic expression. Even architecture and the little painting that has survived from the early periods partake of the nature of sculpture. Particularly is this true of rock-cut architecture, which is often little more than sculpture on a colossal scale. Structural buildings are also profusely adorned with sculpture that is often inseparable from it. The close relationship between architecture and sculpture has to be taken into account when considering individual works that, even if complete in themselves, are also fragments belonging to a larger context. Indian sculpture, particularly from the 10th century onward, thus cannot be studied in isolation but must be considered as part of a larger entity to the total effect of which it contributes and from which it in turn gains meaning.

The subject matter of Indian sculpture is almost invariably religious. This does not mean that it cannot be understood as a work of art apart from its religious significance; but, at the same time, an understanding of its motivation and intent enriches one's appreciation. Much of what is represented is the recounting of legend and myth, particularly in the two centuries before Christ, when narrative relief was much in vogue. The work at this time, didactic and edificatory in intent, generally expresses itself in forms that are surprisingly earthy and sensuous. The anthropomorphic representation of the Buddha is avoided, and the subsidiary gods and goddesses are very much creatures of this earth. The Buddha image formulated around the 1st century AD is not what one would expect of the meditative, compassionate, Master of the Law; he is presented rather as an energetic, earthy being radiating strength and power.

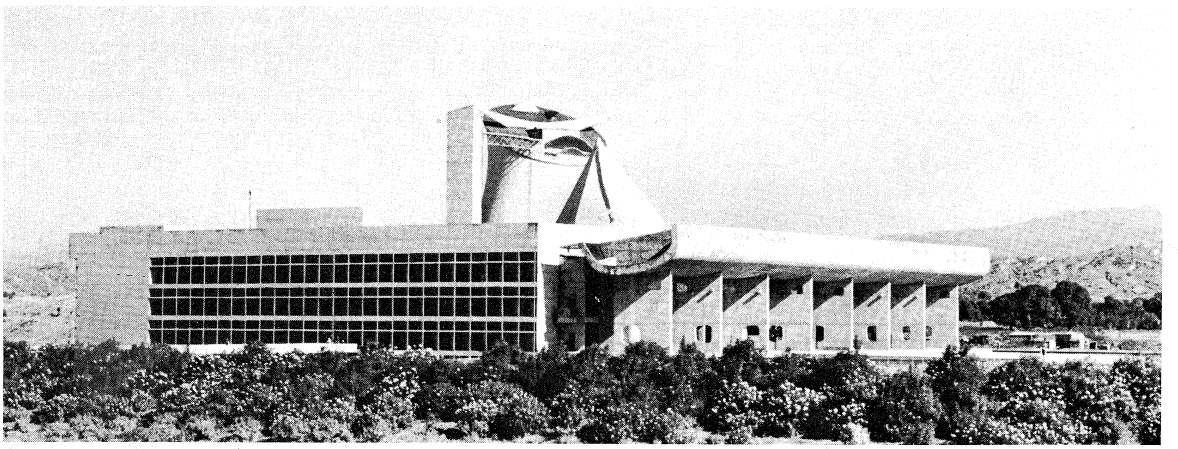
Buddhist influences

The foundations of traditional Hindu imagery were also laid about the same time that the Buddha image was first formulated: images with several arms, and sometimes heads, representing the Indian mind's attempt to define visually the infiniteness of divinity. In subsequent periods the image with many arms became a commonplace in Hindu, Buddhist, and Jaina iconography. Although the various pantheons expanded, they continued to share features of common derivation, expressing the belief that beyond the phenomenal multiplicity of forms lay the unity of the Godhead.

In addition to the major religions, there has always existed in India a substratum of folk beliefs and cults dedicated to the worship of powers that preside over the operation of the life processes of nature. These fertility cults, best expressed in the worship of the male and female divinities *yakṣas* and *yakṣis*, played an important part in the development of Indian art. Among the perennial motifs that spring from the cults, those expressing life and abundance—such as the lotus, the pot overflowing with vegetation, water, or the like, the tree, the amorous couple, and above all the *yakṣas* and *yakṣis* themselves—are most significant. The images of these divinities, in particular, are the source of a great deal of artistic imagery and played a leading part in the development of iconographic types such as the images of the Buddha, the goddess Śrī, and other divinities. The maternal as the ideal of female beauty, manifested in the emphasis on full breasts and wide hips, can be traced to the same beliefs. The very richness and exuberance of much Indian art is an expression of the view of life that equates beauty with abundance.

It is difficult to generalize about the style of a sculptural tradition that extended over a period of almost 5,000 years, but it is nevertheless clear that the distinguishing quality of Indian sculpture is its emphatic plasticity so obvious in Sāñcī I and Mathurā sculpture from the 1st–3rd century AD. Forms are seen as swelling from within in response to the power of an inner life, the sculptor's function being to make these more manifest. At the same time a vision of form that is carved from without rather than modelled from within is also present, as for example at Bhārhut. The history of much of Indian sculpture, marked by periods of high achievement bursting with creativity followed by periods in which the potentialities so postulated are gradually worked out, is essentially the interaction of these two dominant tendencies.

Indus Valley civilization (c. 2500–1800 BC). Sculpture found in excavated cities consists of small pieces, generally terra-cotta objects, soapstone, or steatite, seals carved for the most part with animals, and a few statuettes of stone and bronze. The terra-cotta figurines are summarily modelled and provided with elaborate jewelry, which was fashioned separately and applied to the surface of the piece. Most of the work is simple, but a small group of human heads with horns are very sensitively modelled. Animal figures are common, particularly bulls, which are



The Legislative Assembly chambers of the states Haryana and Punjab in Chandīgarh, India, designed by Le Corbusier, 1952.
Paolo Koch—Rapho Guillumette

often carved with a sure understanding of their bulky, massive form. This plastic quality is also found in the humped bulls engraved on steatite seals, where the modelling is more refined and sensitive. A humpless beast, generally called a "unicorn," is another favourite animal, but it is frequently quite stylized. In addition to bisons, elephants, rhinoceroses, and tigers, seals are carved with images of apparent religious significance, often strongly pictographic.

Terra-cotta sculpture and seals

The terra-cotta sculpture and the seals both show two clear and distinct stylistic trends, one plastic and sensuous, the other linear and abstract. These appear during the same period and are also seen in the small group of stone and bronze sculptures that date from this period (National Museum, New Delhi). Of extraordinarily full and refined modelling is a fragmentary torso from Harappā, barely four inches (ten centimetres) high but of imposing monumentality; the same feeling for massive form is present in a lesser known bronze buffalo. A jaunty bronze dancing girl with head tilted upward (about 4½ inches [11 centimetres] high), from Mohenjo-daro, and a headless figure of a male dancer from Harappā, shoulders twisted in a circular movement, clearly demonstrate, in the attenuated and wiry tension of their forms, the second component of Indus Valley art. Of great interest is a famous bearded figure from Mohenjo-daro wearing a robe decorated with a pattern composed of trefoil motifs. The tight, compressed shape of the body and the expansive modelling of the head demonstrate that the two aspects of form revealed in Indus Valley art were not compartmentalized but interacted with each other. This can also be seen in the interplay of modelled form and textured surface frequently found in works produced by this civilization.

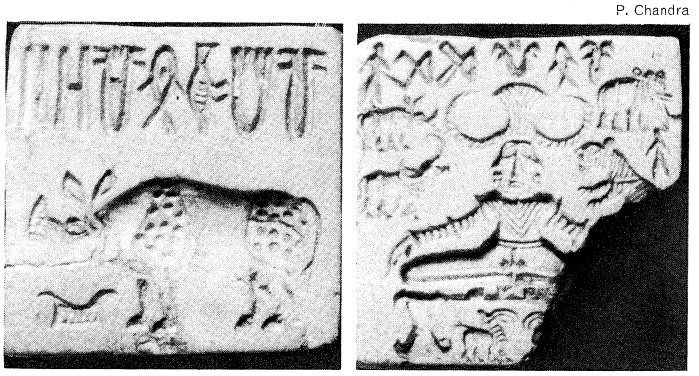
Maurya period (c. 3rd century BC). Little is known of Indian art in the period between the Indus Valley civilization and the reign of the Maurya emperor Aśoka. When sculpture again began to be found, it was remarkable for its maturity, seemingly fully formed at birth. The most famous examples are great circular stone pillars, products of Aśoka's imperial workshop, found over an area stretching from the neighbourhood of Delhi to Bihār. Made of fine-grained sandstone quarried at Chunār near Vārānasi (Benares), the monolithic shafts taper gently toward the top. They are without a base and, in the better preserved examples, are capped by campaniform lotus capitals supporting an animal emblem. The entire pillar was carefully burnished to a bright lustre commonly called the "Maurya polish." The most famous of these monuments is the lion capital at Sārnāth, consisting of the front half of four identical animals joined back to back. There is a naturalistic emphasis on build and musculature, and the modelling is hard, vigorous, and energetic, stressing physical strength and power. Very similar, if not at the same level of achievement, is the quadruple lion capital at Sāñcī. Single lions are found at Vaiśālī (Bakhra), Rāmpurvā, and Lauriya Nandangarh. The Vaiśālī pillar is heavy and squat, and the animal lacks the verve of the other animals—features, according to some, designating it as an early work, executed before the Maurya style attained its maturity. By contrast, the Rāmpurvā lion, finished with painstaking and concise artistry, represents the style at its best. His smooth, muscled contours, wiry sinews, rippling, flamboyant mane, and alert stance reveal the work of a superior artist. An example at Lauriya Nandangarh is interesting because the pillar and the lion are both complete and in their original place, giving a clear idea of the column as it appeared to its contemporaries.

The Rāmpurvā lion

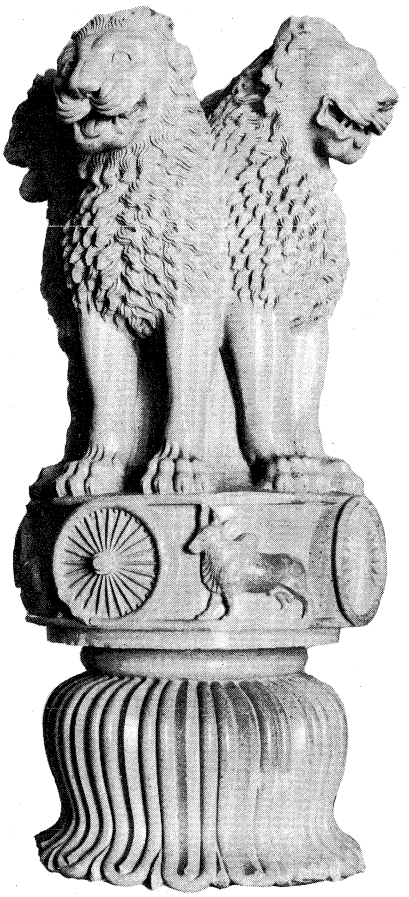
The lion was the animal most often represented, but figures of elephants and bulls are also known. At Dhauli in Orissa, the fore part of an elephant is carved out of rock on a terrace above a boulder that carries several of Aśoka's edicts. The modelling here is soft and gentle, and the plump, fleshy qualities of the young animal's body, seen as emerging from the rock, are suffused with warmth and natural vitality. Since the contrast with the rather formal, heraldic lions could not be more complete, the sculpture clearly testifies to the simultaneous existence of a style different from that of the lion capitals. The style might very well represent the indigenous tradition of plastic form that appears consistently in later art and also in some of the animal capitals made in the imperial atelier, notably the damaged elephant that once crowned the pillar at Saṅkīsa and, above all, the splendid bull from Rāmpurvā. In this great work of art, the two opposing concepts of form merge in a work of harmonious power. The pronounced naturalism comes from the same source as do the lions, but the tense line and hard modelling yield to a form that wells from within and at the same time is given stability and strength by a vision imposed from without.

The sudden appearance of Maurya art with seemingly no tradition behind it has led to speculation that it was the creation of foreign artists, either Achaemenian or Hellenistic. Persian influence, particularly in the lotus

P. Chandra



Steatite seals of the Indus Valley civilization (c. 2500–1800 BC). In the National Museum of India, New Delhi.



Lion capital from Sārnāth, Uttar Pradesh, India, Chunār sandstone, mid-3rd century BC. Height 2.13 m.
P. Chandra

capitals and the figures of lions can hardly be denied, but what is remarkable is the drastic reinterpretation of alien forms by Indian artists. This is a process that is repeatedly seen in the history of Indian art.

Besides the animal sculpture, some human figures, more or less life size, can also be assigned to the Maurya period, though scholarly opinion is by no means unanimous on the point. Among the most important are three images discovered at Patna (ancient Pāṭaliputra, the Maurya capital), two of which are representations of *yakṣas*, the popular male divinities associated with cults of fertility, and the third, found at Dīdarganj (a section of Patna), a representation of a *yakṣī*, or female divinity. Stylistically the images are very similar. The standing *yakṣas* (Indian Museum, Calcutta) are powerful creatures; the ponderous weight of their bodies, together with a certain refined appreciation of the soft flesh, is admirably rendered. The Dīdarganj *yakṣī* (Patna Museum), a masterpiece, displays the Indian ideal of female beauty, the heavy hips and full breasts strongly emphasizing the maternal aspect. In a nude torso discovered at Lopanipur, the sophisticated and sensitive treatment of the surfaces and the gentle blending planes that avoid all harsh accents produce a work of much refinement.

Small stone discs

Small stone discs (also called ring stones because several of them are perforated in the centre), found from Taxila to Patna, are clearly connected with the cult of a nude mother goddess. They represent Maurya sculpture on a smaller and more intimate scale but characterized by the same refined and exquisite workmanship. They are executed in bas-relief, which became the favourite form of sculpture in the subsequent period.

The terra-cotta art of the Maurya period is best represented by a substantial group of figurines, modelled for the most part, the clay sculptor performing work in his medium at the same level as the artist working in stone. Patna has yielded a large number of these, but examples are found throughout the Gangetic Plain. The clothing and jewelry are heavy and elaborate, the modelling, particularly of the head, is sensitive, and the expression is often one of great charm and refinement. There are also more archaic examples, with flat bodies, enormous hips, and modelled heads and breasts.

Indian sculpture in the 2nd and 1st centuries BC. The Maurya Empire collapsed in the early years of the 2nd century BC, and with it passed the art with which it was intimately related. The sculpture that is found throughout India from the middle of the 2nd century BC is startlingly different, but the process by which this change took place in a relatively short period of time is not fully understood. Several schools, sharing common features but nevertheless possessing distinct individual characteristics, are known to have existed. The history of the schools of northern India is somewhat obscure, largely due to the great destruction wrought in the Gangetic heartland; but there appears to have flourished there and in adjacent areas a school of great importance represented by the remains discovered at Bhārhut, Sāñcī, Mathurā, and Bodh Gayā. Western India had its own school, as revealed in the sculptures decorating the cave temples, notably those of Bhājā, Pītalkhorā, and Kārlī. In the southeast, the important school of Andhradeśa flourished in the Krishna River Valley at Amāravatī, Jaggayyapeta, and associated sites; and in eastern India, what is now the modern state of Orissa, made its contribution in the rock-cut sculptures at Udayagiri-Khandagiri. The distinctive schools, though spread over a subcontinent, were not isolated from each other. The contacts fostered by a flourishing trade and by the constant movement of pilgrims were always very close, and it was never long before developments in one part of India were echoed in another.

Judging from extant remains, artists of the earlier period (*c.* 3rd century BC) preferred figures carved in the round, relief sculpture being quantitatively quite insignificant. By contrast, it was sculpture in low relief that was favoured in the first two centuries before Christ; the earlier tradition was not quite forgotten, but figures carved in the round are relatively few. Although there is no stylistic difference, relief sculpture is here considered first according to the various regional schools, and sculpture in the round is treated separately.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of northern and central India. Among the most important, and perhaps the earliest, remains in northern India are reliefs from the great *stūpa* at Bhārhut, dating approximately to the middle of the 2nd century BC. The work, suggesting a style imitating wooden



The monkey chief and the king of Vārānasi, red sandstone relief from Bhārhut, Madhya Pradesh, India, mid-2nd century BC. In the Indian Museum, Calcutta.
P. Chandra

Reliefs from the *stūpa* at Bhārhut

sculpture, is characterized by essentially cubical forms, flat planes that meet at sharp angles, and very elaborate and precisely detailed ornamentation of surfaces. Most of the sculpture was confined to the railing of the *stūpa.* Some of the supporting posts bear large image of *yakṣas* and *yakṣīs* of popular religion, now clearly pressed into the service of Buddhism, while most of the others are decorated with medallions in the centre and crescent-shaped motifs, or lunates, at the top and bottom, all filled with lotus motifs. Some medallions contain amorous couples, the overflowing pot, the goddess Śrī standing on lotuses while being ceremonially bathed by elephants and other symbols of abundance; still others contain the earliest illustrations of events in the Buddha's life and of narratives of his former incarnations as related in the *Jātaka* tales (a collection of tales about the Buddha). Although compositions are crowded, great economy of expression is evident because the artist confines himself to the representation of essentials. Figures are often carved in horizontal rows, sometimes asymmetrically, adapting themselves awkwardly to the circular space of the medallion. Continuous narrative, in which events succeeding in time are shown in the same space, is often resorted to—the first occurrence of what was to become a favourite narrative technique. There is no attempt at establishing any interrelationship, psychological or compositional, between the various figures, each of which is strictly confined within its own space. The faces are masklike, without trace of emotion, lending a solemn and hieratic quality to their expression. Trapped between the background and a frontal plane beyond which they are not allowed to project, the figures are in a sense strictly two-dimensional, more so than in any other style of Indian sculpture. Often, however—particularly in the treatment of animals—the artist is more relaxed, giving glimpses of intimate observation and a natural rendering that anticipates the direction of future development. Like the posts, the top part, or coping, of the stone rail is also carved on both faces; on one of them is a continuous creeper bearing lotus flowers, leaves, and buds; on the other, again the winding stem of a creeper, but bearing other good things of life—such as clothes, jewelry, and fruits—and also scenes illustrating the *Jātaka* stories.

Bhārhut is an extremely important monument inasmuch as it seems to mark a new beginning after the refined and naturalistic art of the Maurya Empire. The sophistication, in spite of the archaic, hieratic manner, would indicate that a considerable body of sculptural tradition, particularly in wood, preceded it; but of this no traces have survived. Be that as it may, Bhārhut states for the first time, and at some length, themes and motifs that would henceforth remain a part of Indian sculpture.

Stray finds of sculpture at Mathurā and other sites in modern Uttar Pradesh indicate that the Bhārhut style was spread over a large part of northern India, particularly the region roughly between that city and Vārānasi and Bodh Gayā in the east. A closely related style is also found at Sāñcī in eastern Mālava, where a representative example is the sculpture of the railing of Stūpa II. Although the themes and motifs found at Bhārhut occur here, narrative representations are all but absent. The style is almost identical; the stiff and rigid contours are a little softer, but both the scale and richness of Bhārhut are missing.

Principal glory of Sāñcī

It is the sculpture of the four gateways (*toraṇas*) of the Great Stūpa (Stūpa I) at Sāñcī, however, that is the principal glory of that site, carrying the promise of the Bhārhut style to its fulfillment. The *toraṇas,* four in number, were attached to the plain railing around the middle of the 1st century BC. They consist of square posts with capitals supporting a triple architrave, or molded band, with voluted (turned in the shape of a spiral, scroll-shaped ornament) ends and a top crowned with Buddhist symbols. Bracket figures, in the form of *yakṣīs*, serve as additional supports. All parts of these gates, strongly reminiscent of wooden construction, are covered from top to bottom with the most exquisite sculpture. Subjects and motifs found at Bhārhut are also found here, the same profusely flowering lotus stem and associated motifs, the same compositions with figures basically arranged in horizontal rows, the same love for clear detail; but to all of these are added a truly voluminous sense of form, a smoother and more energetic movement, and a keen appreciation for the forms of nature, all of which endow the sculpture with a naïve and sensuous beauty unparalleled in Indian art.

Departures from the Bhārhut style are particularly striking in the narrative reliefs. Their greater depth, taken together with their crowded composition, results in the background, visible at Bhārhut, being submerged in shadow. The figures, in all their richness and abundance, flow out from the dark ground, secured in place by the frame of the panels. The Bhārhut angular silhouette and the rigid, severe outline of the body yields at Sāñcī to a gently swelling plasticity, animated by a soft, breathing quality that molds the contours without strain or tension. There is a pronounced concern with the organization of composition, and the narration is often leisurely and discursive; the artist does not just tell the basic story but also lingers over the details, amplifying them to give a vivid picture of everyday life. The emotional monotone of Bhārhut survives in some Sāñcī sculptures, but in others it is superseded by joyous faces and the emotional impact of vivid gesture and movement. Dejection is written large on the faces of the soldiers of Māra's army, who had tried to disturb the Buddha's meditation, as they stagger away from the scene of defeat, and the sensuousness of the amorous scenes is successfully evoked by the tender and intimate gestures of the couples. No longer transfixed in their own space, they turn to look at each other lovingly, responding to each other with a deeply felt understanding.

Long and elaborate bas-reliefs carved on the architraves of the *toraṇas* are the summit of the Sāñcī sculptor's art. Among the finest are representations of the wars for the relics, the defeat of Māra, the *Viśvāntara Jātaka,* and the *Ṣaḍḍanta Jātaka.* The compositions are rich and crowded with figures, and are arranged with great skill. Particularly striking is the masterly handling of animals, notably the elephant, whose fleshy body and graceful movement are captured unerringly. Deer, water buffaloes, bulls, monkeys—all of the beasts and birds of the forests—are rendered with a sense of intimacy indicating the artist's sense of the fellowship of man and animal in the world of nature. The lush Indian landscape is often carved with ornamental trees, waterfalls, pools, mountains, and rivers. The Sāñcī sculptor also shows a marked preference for architectural settings, filling his compositions with numerous buildings that often provide the spatial context for the action. Entire cities, with surrounding walls, elaborate gate houses, and palatial mansions, are depicted. Depth is achieved by rendering side views, and multiple perspective continues to be the rule.

Images of *yakṣīs*

The several large images of *yakṣīs* serving as brackets supporting the lowermost architraves of the *toraṇas* are unique achievements. Like the same goddesses at Bhārhut, they are shown in association with a tree to which they cling, but the style is remarkably different. The modelling shows a concern for the charms of the body, stressing the tactile nature of its flesh. The heavy jewelry and clothing that conceal the body are drastically reduced, revealing its nudity. The soft, melting sensuousness of the female form is so greatly emphasized that the belly and the folds of flesh at the waist are almost flabby, redeemed only by the smooth, firm breasts and the tender arms and limbs.

By comparison, reliefs adorning the railing around the Mahābodhi temple at Bodh Gayā (of about the same date or a little earlier) are in a somewhat impoverished idiom, lacking the rich proliferation both of Bhārhut and Sāñcī. The posts have the usual medallions, lunates filled with lotuses, and reliefs depicting the familiar scenes of Buddhist myth and legend. The artistry of Bodh Gayā, however, is of a lower level of achievement than that at either Bhārhut or Sāñcī: the relief is deeper than that at Bhārhut but shallower than that at the Great Stūpa of Sāñcī; and crowded compositions are lacking, as are the clear and precise ornament and the rich floral motifs. The

Bodh Gayā sculptor, however, though abbreviating even further the iconography of Bhārhut, breaks up, as does the Sāñcī sculptor, the spatial isolation that so uncompromisingly separated each individual figure at that site.

The great school of Mathurā, also, seems to have come into existence about the 2nd century BC, though its period of greatest activity falls in the first two centuries after Christ. The city was repeatedly sacked in the course of the centuries, which may account for the paucity of materials, but enough has been discovered to reveal that the style, in its early stages, was very similar to that of Bhārhut, characterized by flat two-dimensional sculpture decorated with abundant and precise ornament. Several fragments discovered at the site show the gradual stages by which this style evolved, leading to the sculpture of the Great Stūpa at Sāñcī on the one hand and to Bodh Gayā on the other.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of Andhradeśa. Besides the schools of northern India, a very accomplished style also existed in southeast India; the most important sites are Jaggayyapeta and Amarāvatī, activity at the latter site extending well into the 2nd century AD. The early remains are strikingly similar to those at Bhārhut, the relief generally even shallower and the modelling comparatively flat. In contrast to those found in northern India, the proportions of the human body are elongated; but in its flat, cubical modelling, angular, halting contours, and precise, detailed ornamentation, the style is essentially similar to contemporary work elsewhere, right down to the same conventional clothing and jewelry. The nervous, fluid treatment of surfaces, so characteristic of subsequent Andhra sculpture, is already present here. The preferred material is marble rather than the sandstone invariably used in the north.

The style of the Andhradeśa school developed in a manner consistent with other regions of India, becoming more voluminous and shedding the early rigidity fairly rapidly. A group of sculptures at Amarāvatī are characterized by the same qualities that distinguish the work at the Great Stūpa of Sāñcī: full and lissome forms, modelling that emphasizes mass and weight, and sensuously rendered surfaces.

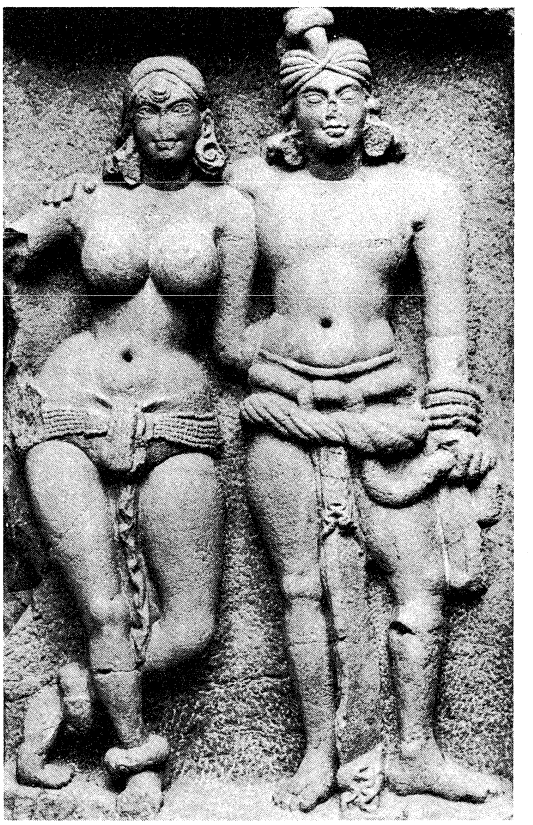
Rock-cut cave temples of the Western Ghāts

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of western India. The numerous rock-cut cave temples in the Western Ghāts are, comparatively speaking, much less profusely adorned with sculpture than remains from other parts of India. The earliest works are undoubtedly the bas-reliefs on a side wall of the porch of a small monastery at Bhājā. They are commonly interpreted as depicting the god Indra on his elephant and the sun god Sūrya on his chariot but are more probably illustrations of the adventures of the mythical universal emperor Māndhātā. What is immediately evident is that these sculptures are not imitations of wooden prototypes, like those at Bhārhut, but, rather, reflect a tradition of terra-cotta sculpture, abundant examples of which are found in northern India and Bengal, where this medium was very popular because of the easy availability of fine clay. The terra-cotta tradition is reflected in the amorphous, spreading forms of Bhājā and in the fine striations used in depicting ornaments and pleated cloth, techniques natural and appropriate to the fashioning of wet clay. The fact that there are some similarities to the Bhārhut style—the stilted postures of the figures and the flat contours of the body, for example—indicates that the beginnings of the western Indian school would also have to be placed about the middle of the 2nd century BC.

The next major group of sculptures in western India have been found at Pītalkhorā. The colossal plinth of a monastery decorated with a row of elephants, the large figures of the door guardians, and several fragments recovered during the course of excavations are among the more important remains. A great proportion of the work represents an advance over the style of Bhājā, though features derived from terra-cotta sculpture continue to be found: the figures are carved in greater depth and volume, but the texture of the drapery, the soft contours of the body, and the high relief of the jewelry, which sometimes gives the impression of having been fashioned separately and then applied, testify to the continuing strength of the terra-cotta tradition. Although the hard line and sharp cutting of some sculpture is reminiscent of the earlier, wood-carving tradition as seen at Bhārhut, the forms are more appropriate to the stone medium. Moreover, the expression is more explicit; and for the first time, both gently smiling and boldly laughing figures of *yakṣa*s appear, as well as the figure of a lover blissfully drunk on wine offered to him by his beloved. These features are also found in the later sculpture of the Great Stūpa at Sāñcī and, to a more pronounced extent, in the sculpture of the Mathurā school of the 1st centuries AD—for example, in the happily smiling *yakṣī*s from Bhutesar.

The cave temple at Kondane has, above the entrance hall, four beautiful panels depicting pairs of dancers. The forms retain the robust and full modelling of the more developed sculpture at Pītalkhorā, but to this is added an ease of movement and considerable rhythmic grace. Traces of the terra-cotta tradition are now totally absent; nor do they occur in the next phase, best represented by a group of sculptures found in the rock-cut temples and monasteries at Beḍsā and Nāsik and in the *caitya,* or temple proper, at Kārlī. Sculpture at all these sites shows many affinities to the Great Stūpa at Sāñcī and should be approximately contemporary or a little earlier. Easily the most outstanding achievements of this region and period, and for that matter one of the greatest achievements of the Indian sculptor, are the large panels, depicting amorous couples, located in the entrance porch of the

P. Chandra



Amorous couple, detail of the *caitya* at Kārlī, Mahārāshtra, India, 1st century BC.

Kārlī *caitya.* Here the promise of early work achieves its fulfillment, the full weighty forms imbued with a warm, joyous life and a free, assured movement. The resemblance to work at the Great Stūpa of Sāñcī is obvious, though these figures at Kārlī are on a much larger scale and possess a massiveness and monumentality that is a characteristic of the distinct western Indian idiom.

Indian sculpture in the 2nd and 1st centuries BC: relief sculpture of Orissa. Sculpture decorating the monasteries cut into the twin hills of Udayagiri and Khandagiri

Sculpture decorating rock-cut monasteries

in Orissa represents yet another early Indian local idiom. The work is not of one period but extends over the first two centuries before Christ; the stages of development roughly parallel the styles observed at Sāñcī Stūpa No. II, Bodh Gayā, and the Great Stūpa at Sāñcī, but they possess, like other regional schools, fairly distinct and individual features. The earliest sculptures are the few simple reliefs found in the Alakāpurī cave, humble works that recall the bas-reliefs of Sāñcī Stūpa II. The Mañcapurī, Tatowā Gumphā, and Anantā cave sculptures—particularly the image of Sūrya riding a chariot—are more advanced and resemble work at Bodh Gayā. The forms are heavy and solid and lack the accomplished movement of the later cave sculpture adorning the Rānī Gumphā monastery. These, like other sculptures here, are in a poor state of preservation, but they represent the finest achievements at the site. Most remarkable is a long frieze, stretching between the arched doorways of the top story, representing a series of incidents that have not yet been identified. The work parallels that of the Great Stūpa at Sāñcī, with the same supple modelling and crowded compositions. At the same time there is a nervous agitation, a fluid, agile movement together with a decided preference for tall, slender human figures. The reliefs on the guard rooms of Rānī Gumphā are also quite remarkable, depicting forested landscapes filled with rocks from which waterfalls flow into lakes that are the sporting grounds of wild elephants. The fine work of this cave strikes a romantic and lyrical note seldom found in Indian art.

Indian sculpture in the 2nd and 1st centuries BC: sculpture in the round and terra-cotta. The most important sculpture in the round are the life-size or colossal images of *yakṣa*s and *yakṣī*s, which reinterpret forms established by the two Patna *yakṣa*s and the Dīdarganj *yakṣī* of the Maurya period—very much as a few animal capitals, particularly the *makaras* (a crocodile-like creature) from Kauśāmbī and Vidiśā (Besnagar), echo the tradition of the superb Maurya animal capitals. It is the *yakṣa* figures, however, that deserve special attention, for they played a significant part in the iconographic developments of the 1st century AD and later and contributed substantially to the imagery of the anthropomorphic Buddha icon.

The most famous of the *yakṣa* images is a colossal figure recovered from the village of Pārkham, near Mathurā (Archaeological Museum). It is about 8⅔ feet (2.6 metres) in height, and, though the two hands are broken and the head is considerably damaged, it is an image of great strength. Its squat neck, its head set close to the body, which tends toward corpulence, its swelling belly restrained by a flat band, and a broad chest adorned with necklaces—all of these features contribute to an image turgid with earthy power. The back is flat and cursively finished, so that the figure has the appearance more of a bifacial relief than of an image carved in the round. Although the forms retain some of the cubical modelling of Bhārhut, the swelling limbs and torso have a massive weightiness that makes the image an appropriate representation of a divinity that presides over the productive processes of nature and endows plenty and abundance on his worshippers.

Mathurā *yakṣa* images

The Mathurā region seems to have been an important centre of *yakṣa* worship, for several images, most of them fragmentary, have been discovered there. Some images have also been found from the ancient city of Vidiśā (Vidisha Museum), one of which is even larger than the Pārkham example and is in a better state of preservation. The god holds a bag in one hand (the other was held below the chest), and the hair is tied in a large top knot over the forehead. The image is accompanied by a female consort (*yakṣī*), wide-hipped and full-breasted, who also emphasizes and personifies the powers of fertility.

The widespread nature of the cult is evidenced by the occurrence of *yakṣa* images throughout India. Fragments in the round (not to speak of the relief representations in a Buddhist context) of the 2nd to 1st centuries BC have been found from Madhyadeśa, Orissa, Rājasthān, Andhradeśa, and Mahārāshtra. At Pītalkhorā there is an exceptionally fine image of a *yakṣa* conceived as a pot-bellied dwarf carrying a shallow bowl on his head; the features, with a gently laughing mouth, are suffused with good humour. Similar *yakṣa*s, employed as atlantes (male figures used as supporting elements), are also found on the western gateway of the Great Stūpa at Sāñcī and at other sites, notably Sārnāth.

The latest in the series of cult images is the image of the Yakṣa Maṇibhadra, from Pawāyā (Gwalior Museum). The sculpture is at present headless, but the rest of the body is well preserved. The right hand holds a fly whisk that flares over the shoulder; the modelling of the legs and torso is sensitive, and the folds of the garment wrapped around the body are full and voluminous, recalling the style of sculpture at Sāñcī.

The terra-cotta sculpture of the period consists mainly of relief plaques made from molds found at numerous sites in northern India. These generally depict popular divinities; a richly dressed female figure loaded with profuse jewelry, obviously a mother goddess, is the favoured subject. Scenes from daily life also abound—as well as what appear to be illustrations of current myths and stories. Superb examples have been found from Mathurā, Ahichhatrā, Kauśāmbī, Tāmlūk, and Chandraketugarh. The workmanship is often of the most exquisite clarity and delicacy, the style paralleling that of contemporary stone sculpture.

Indian sculpture from the 1st to 4th centuries AD. This period is characterized by the dominance in northern India of the ancient school of Mathurā. Other schools, such as those that flourished at Sārnāth and Sāñcī in the first two centuries before Christ, for example, were markedly restricted in their artistic output. Much of their sculpture was imported from Mathurā, and the few images they produced locally were strongly influenced by Mathurā work. The narrative bas-relief tradition, consisting of elaborate compositions of edificatory character, was on the wane, and the emphasis was on carving individual figures, either in high relief or in the round. For the first time, images appear of the Buddha, *bodhisattva*s, and various other divinities including specifically Hindu images representing the gods Viṣṇu, Śiva, Varāha, and Devī slaying the buffalo demon; some of these figures begin to feature several arms, a characteristic of later iconography. There are also many images of *yakṣī*s, often in most alluring attitudes and gestures. Their enticing bodies are now presented as unified organic entities, lacking all traces of the stiff, puppet-like aspect that had not been entirely overcome even at the Great Stūpa of Sāñcī. During this period, also, a fresh incursion of foreign influence by way of western Asia was received, quickly assimilated, and transformed in the characteristic manner of Indian art.

The school of Gandhāra

The school of Gandhāra, with Taxila in Pakistan as its centre and stretching into eastern Afghanistan, flourished alongside the Kuṣāṇa school of Mathurā. It is of a startlingly different aspect, stressing a relatively naturalistic rendering of form, ultimately of Greco-Roman origin. The school evolved a distinct type of Buddha image and was also rich in relief sculptures depicting Buddhist myth and legend. Drawing largely on Indian traditions of composition, it nevertheless reinterpreted them in its own manner. The schools of Mathurā and Gandhāra were in close proximity and undoubtedly influenced each other, but essentially each adheres to its own concept of style.

The ancient Indian relief style found its fullest expression and development at neither Mathurā nor Gandhāra but in Andhradeśa, notably at the great sites of Amarāvatī and Nāgārjunikoṇḍa. Railing pillars and other parts of *stūpa*s decorated with *Jātaka* tales and scenes from the Buddha's life are found in great number and are of the most exquisite quality. Free-standing images of the Buddha, on the other hand, are relatively rare, being found only toward the close of the period.

Indian sculpture from the 1st to 4th centuries AD: Mathurā. One of the most important contributions of the school of Mathurā was the development of the cult image of the Buddha, who had been previously represented by aniconic (not made as a likeness) symbols. There is a certain amount of controversy about whether Mathurā

or Gandhāra originated the Buddha image, which appears to be insoluble in view of the circumstantial nature of the evidence. It is possible that the two schools independently developed their own separate types of images; but, at least as far as the Mathurā image is concerned, it is clear that it is a natural development from the tradition of large *yakṣa* sculptures found in this region. The development can easily be seen in a famous image (discovered at Sārnāth and now in the Sārnāth Museum) of Mathurā manufactured and dedicated by the monk Bala. Carved in the round, the image is shown in a pose of strict

P. Chandra



A Mathuriā image of the Buddha discovered at Sārnāth, Uttar Pradesh, India, red sandstone, *c.* AD 81. In the Sārnāth Museum, Uttar Pradesh, India.

frontality, the left hand held at the waist and the right arm, now damaged, originally raised to the shoulder—a posture immediately recalling that of the *yakṣa* images. The jewelry, however, is appropriately omitted, and the body is clothed in simple monastic garments. The modelling throughout is strong and sensuous, and the radiant energy of the body, its affirmative, outgoing movement, is more appropriate to the personality of a *yakṣa* than to that of the Buddha. This standing Buddha image, as seen in the Bala statue, is the standard Mathurā type, several examples of which are known. Along with this one, a similar, seated type developed, of which the best example is the splendid image known as the Kaṭra Buddha (Archaeological Museum). The modelling of the body is refined, the breasts characteristically heavy and prominent, and the flesh of the torso, with its subtle modulations, as convincingly rendered as the Bala image.

The new trends formulated early by the Mathurā school do not indicate a sharp break from the traditions of the earlier schools. This is clear in a series of magnificent *āyāgapaṭas*, or stone tablets originally set up outside *stūpas* to receive worship and offerings. They are usually square or rectangular and richly decorated with auspicious and religious symbols as well as angelic and mythical beings. The extremely decorative, lavish surface treatment gives the immediate impression of a great profusion of multiple forms, akin in feeling to the sculpture of the Great Stūpa of Sāñcī. The organization of these forms, however, has none of the easy freedom of Sāñcī. The figures, for example, are often cast in a regular, winding shape imitating the movement of the undulating lotus creeper. The same movement is seen in rows of animals depicted with haunches raised and chests touching the ground, features seen in earlier art but now much more emphatically stylized. The bodies of the animals also begin to be overpowered by vegetal forms, the tails, for example, terminating in foliate tips; in a later age, this tendency results in the almost total disintegration of animal shapes under the pressure of the floral.

Stone tablets

It is not to these bas-reliefs, however, that one turns for the most delightful creations of the Mathurā school (for they are in fact the last vestiges of a style rapidly passing out of favour) but to the large number of railing pillars usually carved with representation of *yakṣīs* engaged in playful and enticing activities such as plucking blossoms from trees or leaning on its branches, dancing, bathing under a waterfall, and adorning themselves. Among the most beautiful is a group recovered from Kaṅkālī Ṭīlā and now in the State Museum at Lucknow. The modelling of the figures is generally heavy, the soft, plump bodies suffused with a slow, languorous movement. What is important, however, is the emotion, which is no longer expressed in the face alone but in the whole attitude of the body. The pensive mood of a woman holding a lamp, for instance, is evoked not only by the serene features of the face but by the gentle sway of the relaxed body. Present throughout is a fresh movement of life, a marked striving for diverse and varied effects of posture, movement, expression, and even dress and ornament that brings about vital changes in the nature of Indian sculpture. A remarkable group of railing posts decorated with *yakṣī* images recovered from Bhūteśśar in Mathurā (Archaeological Museum) represent an even more refined achievement than the Kaṅkālī Ṭīlā figures. The heavy proportions, in spite of the full breasts and the wide hips, have been overcome; the happy faces express carefree joy, and the postures of the body are so alive with rhythm as to give the impression of a dancing figure.

Mathurā, during this period, was ruled by the Kuṣāṇa dynasty. A group of portrait sculptures of these rulers (Archaeological Museum), recovered from a village called Māt in the environs of Mathurā, gives an interesting glimpse of the foreign influences entering India at the time. One of them (unfortunately lacking the head) represents the emperor Kaniṣka wearing heavy boots, a tunic, and a coat, and leaning on a mace. The image is quite different not only in dress but also in style from other contemporary works, being essentially linear, with the forms entirely set into the surface. The surfaces have little ornamentation and are marked by extreme simplicity; they are also uncompromisingly stiff and rigid. It is possible that these images represent attempts by a Mathurā artist to imitate a style preferred by his imperial masters; but it was not long before the foreign elements were assimilated into the Mathurā style proper, for later images of Kuṣāṇa chiefs have the same expanding and voluminous form that characterizes other sculptures of this school. A large number of ornamental motifs that now appear in India for the first time undergo a similar process of transformation.

The extent of Mathurā influence on Indian art of this period can be gauged by the sculpture of the school found at several sites in different parts of northern India, notably Ahichhatrā, Kauśāmbī, Sārnāth, and Sāñcī. Most of these sites had been flourishing centres earlier, but only a very limited amount of sculpture was produced during the ascendancy of the Mathurā school; and whatever local sculpture was produced at this time was heavily



Headless portrait statue of the emperor Kaniṣka from Mathurā, Uttar Pradesh, India, *c.* late 1st century AD. In the Archaeological Museum, Uttar Pradesh, India. Height 1.85 m.
P. Chandra

influenced by the Mathurā style. At Sārnāth, for example, both the Bala Buddha imported from Mathurā and its local imitations have been found.

Ivory plaques of Begrām

Ivory plaques discovered at Begrām (Bagrām) in Afghanistan are closely related to the school of Mathurā. These are of great importance; for, though ivory must have been a favourite medium of sculpture, little has been preserved of the early work. Most of it is in very low engraved relief, with fluent, sweeping outlines. The figures are depicted in easy and elegant postures, and the workmanship often attains considerable virtuosity.

Indian sculpture from the 1st to 4th centuries AD: Gandhāra. Contemporary with the school of Mathurā, and extending almost into the 6th century, is the Gandhāra school, whose style is unlike anything else in Indian art. It flourished in a region known in ancient times as Gandhāra, with its capital at Taxila in the Punjab, and in adjacent areas including the Swāt Valley and eastern Afghanistan. The output of the school was very large; numerous images, mostly of Buddhas and *bodhisattva*s, and narrative reliefs illustrating scenes from the Buddha's life and legends have been found. The favoured material is gray slate or blue schist and, particularly during the later phases, stucco. Except for objects excavated at a few well-known sites (such as Taxila, Peshāwar, and the Swāt Valley, in Pakistan, and Jalālābād, Haḍḍā, and Bāmiān, in Afghanistan), most of the finds have been the result of casual discovery or clandestine treasure hunts and plunder, so that their correct provenance is not known. If to this are added the large variety of idioms that appear to have existed simultaneously and the total absence of securely dated images, the wide divergence of scholarly opinion with regard to the schools' evolution can be understood. In the present day, there is general agreement, however, that its most flourishing period probably coincided with Kuṣāṇa rule, particularly the reigns of the emperor Kaniṣka and his successors, and that the school did not long outlast the growth of the Gupta school in the 5th century.

The origins of the Gandhāra style are ultimately Greco-Roman, though, recently, emphasis has been placed on Roman art as the more immediate source. It has also been suggested that the school was created by foreign craftsmen imported into India and by their Indian pupils.

The Gandhāra school is also credited by some scholars with the invention of the anthropomorphic Buddha image. Whether this is correct or not, the Gandhāra image is quite different from that of Mathurā and illustrates the difference between the two schools. Instead of the powerful images directly descended from *yakṣa* prototypes, the Gandhāra version is an adaptation of an Apollo figure, with rather sweet and sentimental features. The definite volume and substance given to the pleated folds of the monastic robes make this image more naturalistic than anything found in Indian art. At the same time, the iconographical features are of Indian origin. Large numbers of *bodhisattva* images conceived in the image of royalty, some with strongly individualized facial features, have also been found.

Narrative relief sculpture

In contrast to Mathurā, narrative relief sculpture was very popular in Gandhāra art. Again, in composition and iconography these reliefs are largely dependent on the earlier Indian schools, but the style is quite distinct. Instead of continuous narrative, incidents separated in time are separately represented, though often arranged in sequence. Violent emotions are realistically rendered. The compositions range from simple horizontal placement of figures to rich and complex arrangements, which often attempt to render space illusionistically.

In the course of time, Indian influence was increasingly felt in the art of Gandhāra, and an abstract vision began to obscure the Greco-Roman naturalism of the earlier forms. In spite of the new influence (and the many grace-

P. Chandra



Bodhisattva of the Gandhāra school, schist sculpture, *c.* 2nd century AD. In the Municipal Museum, Allāhābād, Uttar Pradesh, India.

ful but cloying stucco sculptures that are representative of this late phase) the style shows no signs of vital change. This conservatism, together with the large artistic production, gives an overall impression of considerable monotony. Without any real roots in India and with marked foreign features, the avenues of natural development seem to have been closed to the school, which thus finally disappeared. Nevertheless it made vital contributions to the art of Central and eastern Asia, and several features, drastically transformed, were incorporated in Gupta art.

Indian sculpture from the 1st to 4th centuries AD: Andhradeśa. Besides the schools of Mathurā and Gandhāra, a most accomplished school of sculpture flourished in Andhradeśa during the three centuries after Christ, the most important centres being Amarāvatī and Nāgārjunakoṇḍa. The remains consist mainly of carved railings and rectangular slabs that decorated the great Buddhist *stūpas*, which have largely disappeared. The finds are thus fragmentary and belong to several phases of construction or to separate monuments spanning the 1st, 2nd, and 3rd centuries AD.

Unlike the school of Mathurā, which concentrates on the carving of single figures, the Amarāvatī school carried to the fullest limit of its development the ancient tradition of relief sculpture, which flourished in the two centuries before Christ at sites such as Bhārhut, Sāñcī, and Amarāvatī itself. The marble railing posts are decorated with central medallions and lunates at the top and bottom, all filled with lotus flowers of a very rich design. Often the medallions also contain reliefs illustrating scenes from the Buddha's life and from the *Jātaka* stories, and these are the principal glory of the site.

Two broad phases in the development of narrative relief can be distinguished. In the first, the artist builds on the achievements of early relief sculpture as seen on the Great Stūpa of Sāñcī. The forms are still comparatively heavy, the figures increasingly soft and fleshy, the movement freer but still pervaded by a sense of calm repose. This type of work, represented by relatively few examples, is followed by a phase in which the compositions achieve an extraordinary elaboration and complexity. Most striking is the restless, energetic movement, often nervous and flurried, that possesses the participants in any given scene. Complex relationships and patterns are established between the figures; and space is so articulated that the eye participates in the swirling inner movement of the composition that effectually dissolves the ground on which the figures are carved, while the figures themselves flow out in an endless movement from the ground. The setting is dramatic in the extreme. The loving workmanship, reminiscent of ivory carving, and the superb technical proficiency mark the Amarāvatī reliefs as the culminating point of the entire relief style.

The figures, of both men and women, are of unprecedented suppleness and plasticity, the forms rendered in every variety of torsion and flexion. A fluent, gliding line, often more appropriate to painting than to sculpture, encloses the figures, and pervading the whole is a subtle voluptuousness. The reliefs are often only nominally religious, a pretext for the sculptor's pleasure in representing the leisured and sophisticated life of the time.

Nāgārjunakoṇḍa sculpture marks the last phase of the relief style. The figures become stiffer and puppet-like, the patterns of movement frozen and mechanical but still possessing the energy and richness that always characterize this style.

Buddha images of Andhradeśa

The Buddha is represented in Andhradeśa by both symbolic and anthropomorphic forms. The iconographic formula developed shows him clad in a rather thick garment with stylized folds, and the postures are not as formal and hieratic as the Mathurā. This type of Buddha exercised considereable influence in the development of the Buddha image in Ceylon. In several other features as well, the Andhra style also contributed to the development of early sculpture in Southeast Asia.

Indian sculpture from the 1st to 4th centuries AD: terracotta. The quality of terra-cotta figurines of this period is generally inferior to work produced in the first two centuries BC. Many heads of crude workmanship, with protruding eyes, apparently representing foreigners, were found at sites such as Mathurā, Ahichhatrā, and Kauśāmbī. At the same time, there are some well-modelled heads that imitate the style of stone sculpture and are equally expressive.

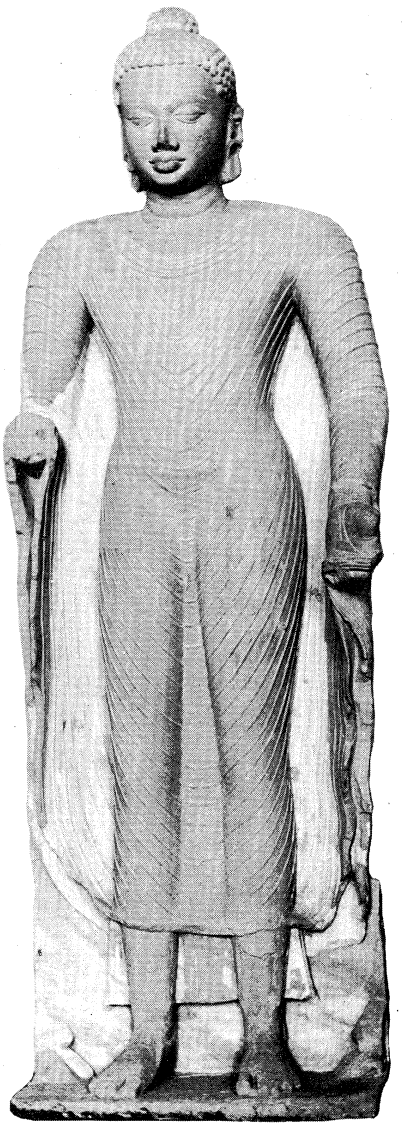
Gupta period (c. 4th–6th centuries AD). During the 4th and the 5th centuries, when much of northern India was ruled by the Gupta dynasty, Indian sculpture entered what has been called its classic phase. The promise of the earlier schools was now fully realized, and at the same time new forms and artistic ideals were formulated that served as the source for development in succeeding centuries. The more or less sensuous and earthy rendering of form was drastically transformed, so that artistic expression closely conformed to the religious vision. The forms are refined and treated with sure and unsurpassed elegance. The volumes, impelled by an inner life, still swell from within but are restrained and controlled, made to flow in smooth and abstract rhythms in an organic and unified concept in which the sensual and the spiritual are inextricably blended. The edificatory, didactic intent of early relief sculpture is abandoned, the works produced are pronouncedly meditative; and the repose and calm that settles on the images of the Buddha, the master of the inner contemplative life, is also seen on images of other divinities. Decorative ornament is in perfect harmony with the volumes it adorns, each emphasizing the other, so that in every respect this is a style of great composure and perfect balance.

Gupta period: Mathurā. The impetus for the new schools seems to have come from Mathurā, which is hardly surprising in view of the preponderant role played by the city in the preceding period. The transformation into the new idiom is best illustrated by a splendid image of the Buddha on which is inscribed "anno 64/A.D. 384" (Indian Museum, Calcutta). Memories of the rather massive and ponderous weight of the earlier style are present, but the calm face no longer looks out at the world; rather, the vision is turned within, the mood being one of serene contemplation. The style, which consistently uses the local red sandstone, undergoes further refinement, seen in a series of magnificent life-size Buddha images of the 5th century (now scattered in museums throughout the world). The more delicate face radiates a feeling of calm inner bliss, and the body is most subtly modelled by smoothly flowing planes that both suggest the swelling force of life and subordinate it to the spiritual vision of the whole. Mathurā images generally show the Buddha wearing a diaphanous robe, the folds of which are rendered by stringlike ridges in a reinterpretation of a Gandhāra convention. The gestures of the hand are delicate and varied. The hair is usually rendered by rows of small curls that conceal the conical protuberance. These Mathurā images established an iconographical type that became the norm for the Buddha image.

Mathurā Hindu images

In addition to the Buddha figure, Mathurā has yielded large numbers of images of the various Hindu divinities, particularly Viṣṇu-Kṛṣṇa. This is in keeping with the increasing strength of the various Hindu cults and the intimate association of Mathurā with the god Kṛṣṇa. The famous image of Viṣṇu from Kaṭrā Keśavadeva in Mathurā is one of the finest (National Museum, New Delhi). The god is conceived as a royal figure, wearing a crown and appropriate jewelry, his features imbued with a dignified calm that is suitable to his function as the preserver and is also characteristic of most Gupta art.

Gupta period: Sārnāth. This famous centre of Indian art developed a sweeter and more elegant version of the Buddha image than Mathurā's. Instead of the rather strict frontal posture, the weight of the body is thrown more on one leg, resulting in a very subtle contrapposto position, in which the hips, shoulders, and head are turned in different directions. This lends a certain movement to the figure, so that it does not quite possess the static, steadfast quality of Mathurā. The robes are no longer ridged with folds but are plain, and the surface of the stone is even more abstractly handled than is the Mathurā. The faces are heart-shaped, the transitions from one part of the body to another smoother, so that the images have great



Buddha, red sandstone sculpture from Mathurā, Uttar Pradesh, India, 5th century AD. In the Indian Museum, Calcutta.
P. Chandra

refinement even if they do not possess the strength of Mathurā. The characteristic Sārnāth style, the preferred material of which is the local buff Chunār sandstone, seems to have developed in the late 5th century, the few earlier works being closer to the Mathurā school. The most famous image from the site and one of the masterpieces of Indian art is that of the seated Buddha preaching (Sārnāth Museum). It is exceptionally well preserved and delicately carved. The face, with serene features and a gentle smile playing on the lips, suggests the joy of supreme spiritual achievement. The halo behind the Buddha is also very beautifully carved, with exquisite floral patterns. Large numbers of Buddha and *bodhisattva* images have been excavated at Sārnāth and are to be found in the museum at the site and in major collections throughout the world.

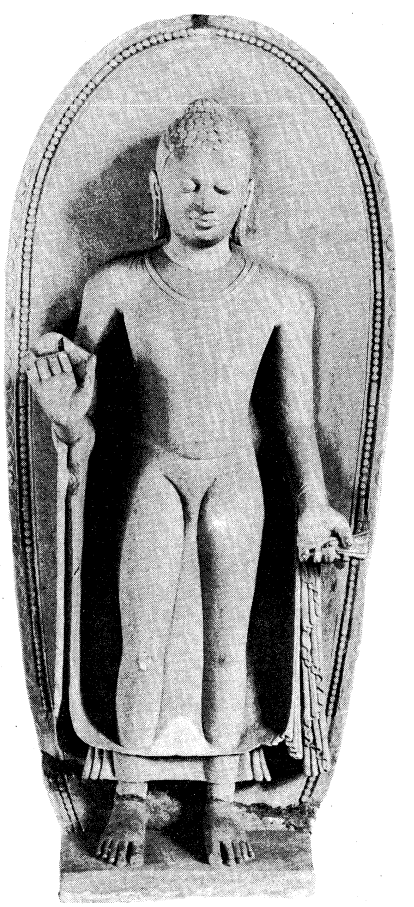
Gupta period: central India. In addition to the major schools of Sārnāth and Mathurā, important sculpture of the 5th and 6th centuries is found at several sites in central India. The sculptures here are often in their original locations, surviving not as isolated images torn from their architectural context but in association with the temples of which they formed a part. At Udayagiri, near Vidiśā, are a series of simple rock-cut caves of the opening years of the 5th century. The sculpture, made of soft stone, has suffered greatly, but whatever has survived reveals a style that stresses strength and power. Perhaps the most magnificent work is a great relief panel depicting the boar incarnation of Viṣṇu lifting the earth goddess from the watery deeps into which she had been dragged by a demon. The massive figure of the god, with the body of a man and the head of a boar, is carved in a surging movement across the face of the rock, the goddess resting easily on his shoulder, while a host of beings, human and divine, celebrate this great triumph.

The Śiva temple at Bhumarā has also yielded some sculpture of fine quality. The stone is carved with great precision and skill, nowhere more evident than in the handling of exuberant floral ornament. Little in Indian decorative sculpture can match the brilliance of the large panels filled with lotus stems and floriated scrolls discovered at this site and at Nāchnā Kutharā.

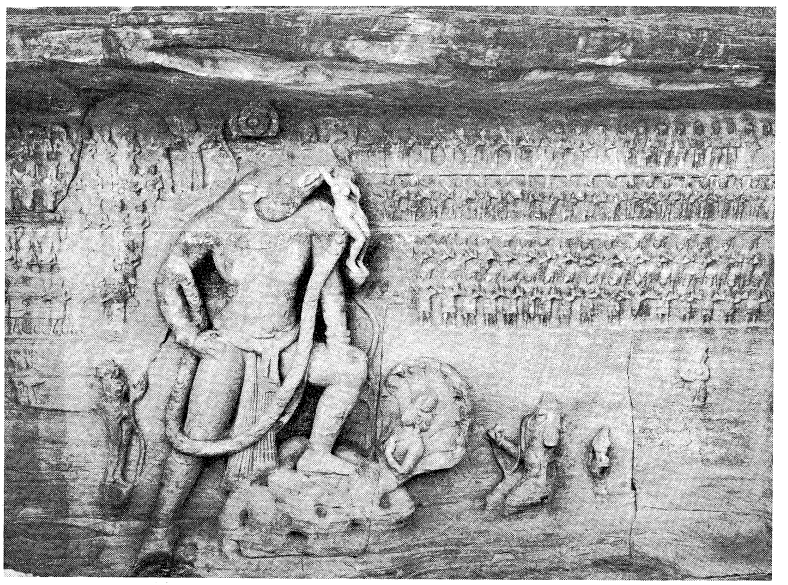
Some of the finest Gupta sculpture adorns the walls of the Viṣṇu temple at Deogarh. Particularly striking are three large relief panels depicting Viṣṇu lying on the serpent Śeṣa, the elephant's rescue, and the penance of Nara-Nārāyaṇa. The compositions tend to be dramatic; the carving and decoration, sumptuous, the sturdy forms recalling Mathurā rather than the attenuated grace of Sārnāth. The doorframe of the sanctum of this temple is an especially fine example of architectural decoration popular in this period. Bands of floral scrolls, amorous couples, and flying angels of great elegance are carved around the entrance. Particularly impressive are groups of worshippers at the base, their swaying bodies related to each other with an easy rhythm.

Gupta period: Mahārāshtra. A great revival of artistic activity seems to have taken place in this region during the reign of the Vākāṭaka dynasty and its successors, best expressed in the splendid sculpture decorating the cave temples of Ajantā and Elephanta. The idioms established in the North were adapted here to the needs of a style that conceived figures on a massive scale, as determined by the demands of the great expanses of rock out of which they

P. Chandra



Buddha, Chunār sandstone sculpture from Sārnāth, Uttar Pradesh, India, 5th century AD. In the Indian Museum, Calcutta.



Viṣṇu rescuing the Earth goddess, sandstone relief panel from a cave at Udayagiri, Madhya Pradesh, India, early 5th century AD.
P. Chandra

Sculpture of the Ajantā and Elephanta caves

were carved. Although the sculpture at Ajantā (mostly of the late 5th century) combines the old weightiness with the new restraint and elegance, the style finds its supreme expression in the magnificent cave temple at Elephanta. The central image of this great temple is of immense size and in deep relief. It represents Śiva in his cosmic aspect, the central head calm, introspective, self-sufficient, and transcending time, the heads to the sides, in their sensuous beauty and awesome terror, reflecting the creative and the destructive aspects of the supreme divinity.

Gupta period: other regions. The impact of the Gupta style of the 5th and 6th centuries was felt in many parts of India, though actual remains thus far discovered are more abundant in some parts than in others. There appears to have been, in Bihār, a distinct school characterized by rather heavy, compact forms; and Gujarāt and southern Rājasthān developed an individual style of considerable voluptuousness and plasticity. Among the notable sculpture of the Idar region are groups of mother goddesses whose massive forms are rendered with an easy grace and intimacy. In the Karṇāṭa country, to the south, the cave temples of Bādāmi reveal yet another distinct idiom, somewhat direct and elemental but nevertheless belonging to the same general style, with local variations, that prevailed over the greater part of India.

Gupta period: terra-cotta. Terra-cotta sculpture, like art in other mediums, was greatly developed. Fairly large and elaborate plaques were used to adorn brick *stūpa*s and Hindu temples from Sind to Bengal. The polychrome relief images of the Buddha from Mīrpur Khās are delicate and slender, with traces of Gandhāra feeling. Representations of divinities and mythological scenes from temples in Bīkaner, Ahichhatrā, Bhītargaon, and Śrāvastī are works on a more popular level, possessing an earthy ponderousness. A large number of figurines, particularly fragments of heads with elaborate coiffures and delicate, smiling features, have been found at Rājghāt in Vārānasi and at other sites.

Medieval Indian sculpture. Indian sculpture from the 7th century onward developed, broadly speaking, into two styles that flourished in northern and southern India, respectively. In each of these regions there also developed additional local idioms, so that there was a wide variety of schools. All, however, evolved in a consistent manner, the earlier phase marked by relatively plastic forms, the later phase by a style that emphasizes a more linear rendering. The sculpture was used mainly as a part of the architectural decor, and the quantity required was vast. This often entailed a mechanical production, with the result that works of quality are few in proportion to the numbers.

Besides the two main idioms, the local schools of Mahārāshtra and Karṇāṭa are of particular interest because they possess considerable individuality and often show both northern and southern features.

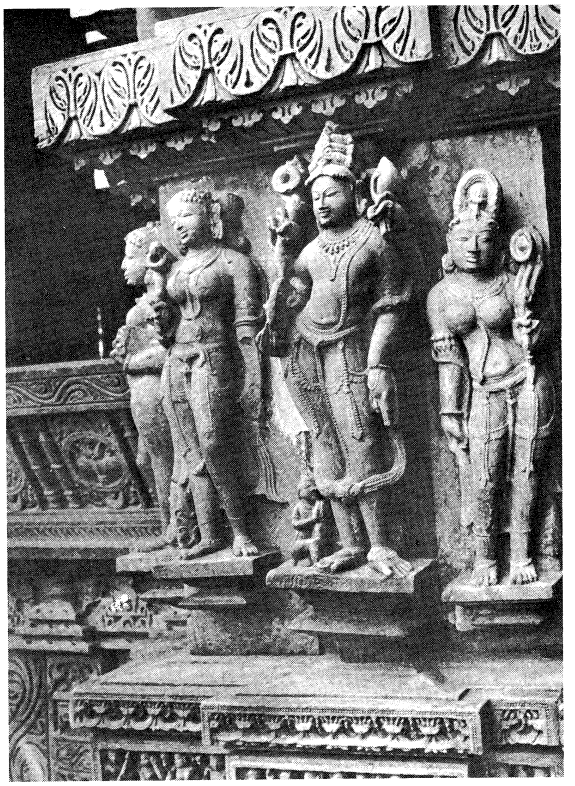
Sculpture in bronze was also produced in fairly large quantities in this period. Again, several local schools can be distinguished, the most important of which are those of eastern and southern India.

Medieval Indian sculpture: North India. The history of North Indian sculpture from the 7th to the 9th centuries is one of the more obscure periods in Indian art. Two trends, however, are clear: one exhibits the decline and disintegration of classical forms established during the 5th and 6th centuries; and the other, the evolution of new styles that began to possess overall unity and stability only in the 10th century.

A breakdown of the Gupta formula is observable from at least the 7th century onward, if not a little earlier: harmonious proportion, graceful movement, and supple modelling begin to yield to squat proportions, a halting movement, and a more congealed form. Toward the 8th century, signs of a new movement become evident in a group of sculptures that departs from the progressively lifeless working out of the Gupta idiom. The modelling emphasizes breadth but with a pronounced feeling for rhythm, and the delineation of decorative detail is fairly restrained. In the 9th century, particularly during the second half, a distinct change came over the styles of all of northern India. A new elegance, a richer decorativeness, and a staccato rhythm so characteristic of the medieval styles of the 10th and 11th centuries begin to be clearly seen and felt. Sculpture of this period reaches a standard of elegance never surpassed in the medieval period: the grace and voluminousness of earlier work are modified but not lost; the harsh angularity of later work, avoided. An idea of the style can be formed from an important group of sculptures at Ābānerī, the Śiva temple at Indore, and the Telī-kā-Mandir temple at Gwalior, as well as from individual works in various North Indian museums.

Signs of a new movement

With the 10th century, the conventions of North Indian sculpture became fairly well established. The style is represented by examples from such monuments as the Lakṣmaṇa temple at Khajurāho (dated 941), the Harasnāth temple at Mt. Harsha (*c.* mid-10th century), in Rājasthān, and numerous other sites scattered all over northern India. These works are executed in a style that has become harder and more angular, the figures covered with a profusion of jewelry that tends to obscure the forms it decorates. These features are further accentuated in the 11th century, when many temples of great size, adorned



Detail of a wall of the Laksmana temple at Khajurāho, Madhya Pradesh, India, sandstone, *c.* AD 941.
P. Chandra

with prodigious amounts of sculpture, were erected all over northern India. There is a decline in the general level of workmanship: the carving is often entirely conventional and lifeless, the features rigid and masklike, and the contours stiff and unyielding. The ornamentation, consisting of a profusion of beaded jewelry, is for the most part as dull, repetitive, and lifeless as the rest of the sculpture. This phase of artistic activity is represented at important centres from Gujarāt to Orissa; one of them is Khajurāho, with a vast amount of sculpture, all in a good state of preservation but conceived and executed as perfunctory architectural ornamentation. Not all sculpture, however, is of inferior quality; the hard, metallic carving and angular, stylized line sometimes result in works possessing a cold brilliance.

The 12th century marks the end of traditional sculpture all over northern India, except for a few pockets not yet penetrated by the Islāmic invasions. A rigid line imposed itself on the forms, which in turn became desiccated and hard, so that whatever unity of surface may have existed was entirely shattered. A brief revival took place in parts of Gujarāt and Rājasthān in the 15th century, but the sculpture merely imitated the work of the late medieval phase. The pure geometry of their forms, however, sometimes results in works possessing a curious archaistic power.

Sculpture in eastern India

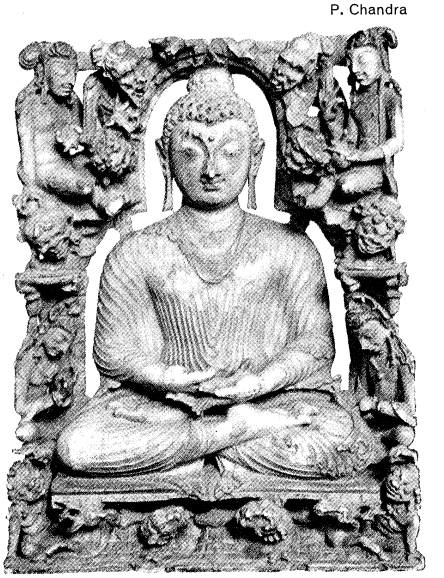
Sculpture in eastern India (consisting of Bangladesh and the modern Indian states of Bihār, West Bengal, and Orissa), though sharing in the broad pattern of development of the rest of northern India, nevertheless represents a distinct idiom. The flatness of planes and angularity of contours are less pronounced, the figures retaining a sense of mass and weight for a greater period of time and to a greater degree. This can be clearly seen in sculpture from Konārak, in Orissa. Dating to the 13th century, the style retains a considerable semblance of plasticity at a period when sculpture in other parts of northern India had assumed a very wooden appearance. In Bihār and Bengal a flourishing school of bronze sculpture also developed, as evidenced by the large number of finds, notably from the sites of Nālandā and Kurkihār. The style generally parallels works in stone, emphasizing plastic values to a great degree. The most flourishing period was the 9th century, when a series of magnificent images representing the gods and goddesses of the Buddhist pantheon were made at Kurkihār and Nālandā. The work of the 10th and 11th centuries is more decorative and often very skillfully and elaborately cast. Of relatively small size and therefore easily transportable, bronze sculpture from this area played an important part in the diffusion of Indian influence in Southeast Asia.

Kashmir sculpture tends to be weightier and more massive than works in other parts of India. Some Gandhāra memories survive, particularly in the fleshy rendering of the body and the drapery, but the sculpture is very much a part of the stylistic developments in northern India. Representative examples of the style, dating to around the mid-9th century, have been found from Avantipura. A flourishing school of bronze sculpture also existed, numerous examples having come to light in recent years. One of the finest, discovered at Devsar (Sir Pratap Singh Museum, Srinagar), is a large 9th-century ornamental frame, 6½ feet (two metres) high, decorated with various incarnations of Viṣṇu, all filled with great energy and movement. A good number of ivory images of Kashmir workmanship have also been preserved. These are generally of miniature size, polychromed, and of extremely fine and delicate workmanship. Influences of the Kashmir style of sculpture were strongly felt in the neighbouring Himalayan region, including both Tibet and Nepal.

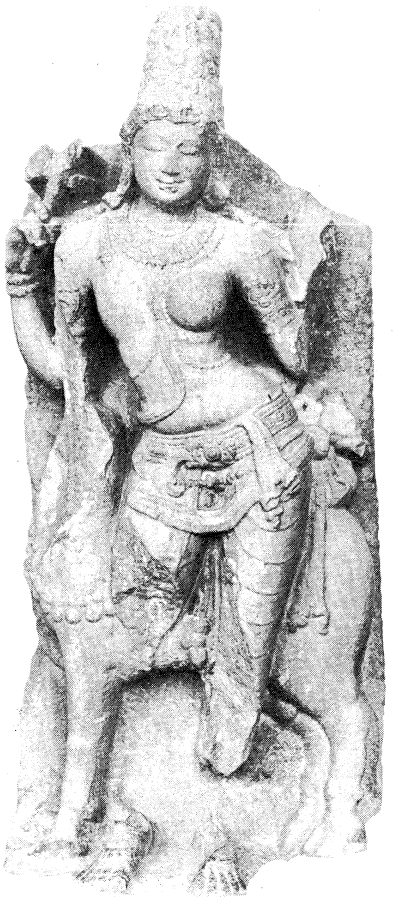
Medieval Indian sculptures: southern India. The medieval phase in southern India opened with elegant 7th-century sculptures at Mahābalipuram, by far the most impressive of which is a large relief depicting the penance of Arjuna (previously identified as an illustration of the mythical descent of the Ganges). It is carved on the face of a granite boulder with a deep cleft in the centre, representing a river, down which water actually flowed from a reservoir situated above. On both sides are carved numerous figures of divinities, human beings, and animals that crowd the hermitage where Arjuna, practicing penance, is visited by Śiva. The tall, slender figures, with supple tubular limbs, remotely recall the proportions of Amarāvatī, now greatly transformed; and the numerous animals, including the elephant herd with its young, show the same intimate feeling for animal life that characterizes all Indian sculpture, but in a manner that has seldom been surpassed.

The light, aerial forms gained stability and strength in subsequent centuries, culminating in superb sculptures adorning small, elegant shrines built during the late 9th century when the Cōḷa dynasty was consolidating its pow-

P. Chandra



Seated Buddha with attendants, carved ivory sculpture from Kashmir, *c.* 8th century AD. In the Prince of Wales Museum of Western India, Bombay. Height 10 cm.



Siva-Ardhanārīsvara, granite sculpture from the Ugrakaliamman temple, Thanjāvūr, Tamil Nadu, India. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu, India. Height 1.14 m.
P. Chandra

er. The temples at Tiruvalīśvaram, Kodumbālūr, Kilaiyur, Śrīnivāsanālūr, Kumbakonam, and a host of other sites of this period are only sparingly adorned with sculpture, but it is of superb quality. With the 10th and 11th centuries, South Indian sculpture, like its counterpart in the north though to a lesser degree, was carved in flatter planes and more angular forms, and the fresh, blooming life of earlier work is gradually lost. This can be seen, for example, in the sculpture of the numerous temples of Thanjāvūr and Gaṅgaikoṇḍacōḷapuram. The subsequent phase, extending up to the 13th century, is represented by work at Dārāsuram and Tribhuvanam; although the forms become increasingly congealed, brittle works of fine quality—often capturing outer movement with great skill—continue to be produced. Sculpture in southern India continued when artistic activity was interrupted in the north by the Islāmic invasions but, in spite of technical virtuosity, became progressively lifeless. Artistic activity continued in the south into the 17th century, the elaborately sculptured halls at Madurā and the masses of stucco sculpture adorning the immense entrances, or *gopuras*, testifying to the prodigious output and the undistinguished quality of the work produced.

South Indian bronze sculpture

South Indian bronze sculpture has a special place in the history of Indian art. A large number of images were made (some of them still in worship in the mid-20th century and others unearthed from the ground by chance), but examples before the 8th century are quite rare. In bronze, as in stone, the 9th and 10th centuries were periods of high achievement, and many images of excellent quality have survived. They are all cast by the lost-wax, or cire perdue, process (in which a wax model is used) and technically are very accomplished. In the early stages the forms were smooth and flowing, with a fine balance maintained between the body and the complex jewelry, the lines of which follow and reinforce every movement of the plastic surface. The bronzes of the later period lose this cohesiveness, the ornament, by virtue of its hardness, tending to divide and fragment the body it covers. The modelling also became flatter and sharper, though not quite as rapidly in bronze sculpture as in stone. Ancient traditions of workmanship survive to the present day, and a few guilds of craftsmen continue to make competent if somewhat lifeless images.

Most South Indian bronze images are representations of Hindu divinities, notably Viṣṇu and Śiva. One particular form deserves special notice as a striking southern contribution to Indian iconography. It is that of a four-armed Śiva as Lord of the Dance (Naṭarāja), shown within a flaming halo, or aureole, one hand holding the double-headed drum symbolizing sound, or creation, and the other holding the fire that puts an end to all that is created. The palm of the third hand faces the devotee, assuring him of freedom from fear, while the fourth hand points to the raised foot, the place of refuge from ignorance and delusion, which are symbolized by the dwarf demon crushed beneath the other foot. Several splendid images are known, the finest being, perhaps, the great image still worshipped in the Bṛhadīśvara temple at Thanjāvūr.

Medieval Indian sculpture: Mahārāshtra and Karṇāṭa. The Karṇāṭa country possessed a flourishing school of sculpture in the 7th and 8th centuries, as seen in examples from Aihole, Pattadkal, and Ālampur. As in architecture, influences from the north are discernible, but the style is basically southern, emphasizing rugged strength and power compared to the more elegant and delicate forms of the Tamil country. In Mahārāshtra, cave temples at Ellora carry the most important examples of this phase of sculpture. Here the tradition is continued

P. Chandra



Siva slaying the elephant demon, granite sculpture from Dārāsuram, Tamil Nadu, India, 13th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu, India.



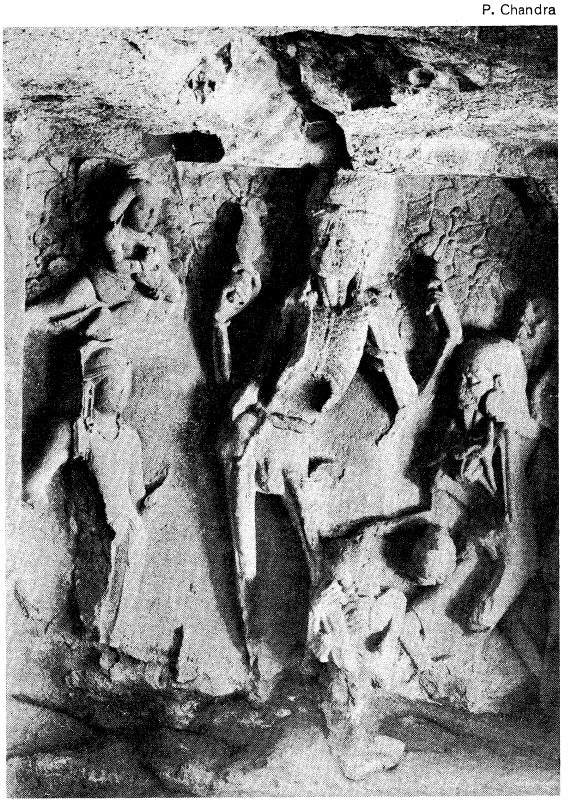
Siva-Naṭarāja, bronze sculpture from Kivalur, Tamil Nadu, India, 11th century AD. In the Thanjāvūr Museum and Art Gallery, Tamil Nadu, India. Height 81 cm.
P. Chandra

of images of great size that, in their primitive strength, partake of the nature of the rock out of which they are carved. A series of large, splendid panels (6th century AD) depicting incidents from Hindu mythology in high relief are to be found in the Rameśvara cave; notable among them is a fearsome representation of the dancing Kālī, goddess of death. The Kailāsa temple (*c.* 757–783) has a remarkable group of elephants struggling with lions all around the plinth. Of the several large reliefs, also at Kailāsa, the depiction of Rāvana shaking Kailāsa is a composition of considerable grace and charm.

Distinctive style of Karṇāṭa

Toward the 13th and 14th centuries, a very distinctive style developed in the Karṇāṭa country, which was then largely ruled by kings of the Hoysaḷa dynasty. The materials employed are varieties of stone that are soft when freshly quarried but harden on exposure, which may account partially for the extreme richness of the work. The sculpture is in very high relief, often undercut and literally covered with the most elaborate ornaments and jewelry from top to toe. This unrestrained extravaganza is unique even for Indian art, which shows a preference for intricate and elaborate ornament at all stages of its history.

P. Chandra



Dancing Kālī, relief from the Ramesvara cave, Ellora, Mahārāshtra, India, *c.* 7th century AD.

Painting. Literary works testify to the eminence of painting as an art form in India, particularly in the decoration of walls, but climate has taken a devastating toll, leaving behind only a few tantalizing examples. By far the bulk of the preserved material consists of miniature painting, initially done on palm leaf but later on paper. The subject matter is generally religious (illustrating divinities, myths, and legends) and literary (illustrating poetry and romances, for example), though the Mughal school was also concerned with historical and secular themes. The styles were rich and varied, often closely connected with one another and sometimes developing and changing rapidly, particularly from the 16th century onward. The work also shows a surprising vitality under strained circumstances, surviving up to the very eve of the modern period when the other arts had deteriorated greatly.

Prehistoric and protohistoric periods. Painting in India should have a history stretching as far back as any of the other arts but, because of its perishable nature, little has survived. None of the examples found in rock shelters over almost all of India, and chiefly representing scenes of hunting and war, appears to be earlier than the 8th century BC, and all may be as late as the 10th century AD. A faint idea of the painter's art in the Indus Valley civilization can be had from the pottery, elaborately decorated with leaf designs and geometrical patterns.

Paintings at Ajantā

Ancient wall painting. The earliest substantial remains are those found in rock-cut cave temples at Ajantā, in western India. They belong to the 2nd or 1st century BC and are in a style reminiscent of the relief sculpture at Sāñcī. Also found at Ajantā are the most substantial remains of Indian painting of about the 5th century AD and a little later, when ancient Indian civilization was in full flower. The paintings, the work of several ateliers, decorate the walls and ceilings of the numerous cave temples and monasteries at the site. They are executed in the tempera technique on smooth surfaces, prepared by application of plaster. The themes, nominally Buddhist, illustrate the major events of the Buddha's life, the *Jātaka* tales, and the various divinities of the expanding Buddhist pantheon. The ceilings are covered with rich motifs, based generally upon the lotus stem and the world of animals and birds. The style is unlike anything seen in later Indian art, expansive, free, and dynamic. The graceful figures are painted by a sweeping and accomplished brush; and they are given body and substance by modelling in colour and by a schematic distribution of light and shade that has little to do with scientific chiaroscuro. The narrative compositions, handled with utmost dexterity, are a natural outgrowth of the long traditions of relief sculpture and reflect the splendour and maturity of contemporary sculpture. The large images of the *bodhisattvas* in Cave 1, combining rich elegance with spiritual serenity, reflect a vision that sees the shifting world of matter and the transcendental calm of Nirvāṇa as essentially one.

Except for a large and magnificent painting of a dance scene found at the rock-cut cave at Bāgh—a painting executed in a style closely resembling Ajantā—hardly any other work of this great period survives. Cave temples at Bādāmi, in the Karṇāṭa country, and Sittānavāsal, in Tamil Nadu, probably of the late 6th and 7th centuries AD are already but echoes of the style of the 5th century, which appears to have died out around this time.

Eastern Indian style. Small illustrations on palm leaf, chiefly painted at the great Buddhist establishments of eastern India, appear to have conserved some elements of this ancient style; but they have lost its dramatic impact, which is replaced by a studied preciosity and an inhibited meticulousness. The surviving paintings date from the

11th and 12th centuries and are conventional icons of the numerous Buddhist gods and goddesses, narrative representations having largely disappeared. With the destruction of these Buddhist centres by the Islāmic invader, the east Indian style seems to have come to an end.

Western Indian style. The style of Ajantā is succeeded in western India by what has been appropriately named the western Indian style. Among the earliest examples are a few surviving wall paintings of the Kailāsa temple (mid-8th century) at Ellora and the Jaina temples, built at the same site a hundred years later. The plastic sense of form, so evident at Ajantā, is emphatically replaced by a style that even at this early stage is heavily dependent on line. The contours of the figures are sharp and angular, the forms dry and abstract; and the fluent, stately rhythms of Ajantā have become laboured and halting.

Jaina manuscripts

The most copious examples of this style, however, have survived not on the walls of temples but in the large number of illustrated manuscripts commissioned by members of the Jaina community. The earliest of these are contemporary with eastern Indian manuscripts and are also painted on palm leaf; but the style, instead of attempting to cling to ancient traditions, moves steadily in the direction already established at Ellora. It is a perfect counterpart of contemporary sculpture in western India, relying for its effect on line, which progressively becomes more angular and wiry until all naturalism has been deliberately erased. The figures are almost always shown in profile, the full-face view generally reserved for representations of the *tīrthaṅkaras*, or the Jaina saviours. A convention that appears unfailingly for the duration of the western Indian style is the eye projecting beyond the face shown in profile, meant to represent the second eye, which would not be visible in this posture. The colours are few and pure: yellow, green, blue, black, and red, which was preferred for the background. In the beginning, the illustrations are simple icons in small panels; but gradually they become more elaborate, with scenes from the lives of the various Jaina saviours as told in the *Kalpa-sūtra* and from the adventures of the monk Kālaka as related in the *Kālakācāryakathā* the most favoured. Even greater elaboration was possible with the increasing availability of paper from the late 14th century; with larger surfaces to paint on, by the middle of the 15th century artists were producing opulent manuscripts, such as the *Kalpa-sūtra* in the Devasanopaḍā library, Ahmadābād. The text is written in gold on coloured ground, the margins gorgeously illuminated with richest decorative and figural patterns, and the main paintings often occupying the entire page. Blue and gold, in addition to red, are used with increasing lavishness, testifying to the prosperity of the patron. The use of such costly materials, however, did not necessarily produce works of quality, and one is often left with the impression of cursive and hasty workmanship. With some variations—but hardly any substantial departures from the bounds that it had set for itself—the style endured throughout the 16th century and even extended into the 17th. The political subjugation of the country by the forces of Islām may have contributed to the conservatism of the style but did not result in its total elimination, as seems to have been the case in eastern India. Indeed, in the course of its long life, the western Indian school became a national style, painting at other centres in India interpreting and elaborating its forms in their own individual manner. In the province of Orissa, painting on palm leaf and in a manner entirely dependent on the western Indian style has continued up to the present day.

Transition to the Mughal and Rajasthani styles. The belief held earlier by scholars that the new Islāmic rulers of India did not patronize any painting until the rise of the Mughal dynasty in the 16th century is being abandoned in the face of the literary testimony and the discovery or recognition of illustrated manuscripts that were painted at Indian courts. Nor should this be surprising, as the Muslim kings of India had before them the example of other rulers of the Islāmic world who were great patrons of painting in spite of the injunctions of orthodox Islām against the portrayal of living beings. The taste of these Indian rulers, however, did not turn to the western Indian style but to the flourishing traditions of Islāmic painting abroad, notably neighbouring Iran. As many painters as architects had in all probability been invited from foreign countries; and illustrated manuscripts, handily transported, must have been easily available. As a result there appears to have developed what can only be called an Indo-Persian style, based essentially on the schools of Iran but affected to a greater or lesser extent by the individual tastes of the Indian rulers and by the local styles. The earliest known examples are paintings dating from the 15th century onward. The most important are the *Khamseh* ("Quintet") of Amīr Khosrow of Delhi (Freer Gallery of Art, Washington, D.C.), a *Bostān* painted in Mandu (National Museum, New Delhi), and, most interesting of all, a manuscript of the *Neʿmat-nāmeh* (India Office Library, London) painted for a sultan of Mālwa in the opening years of the 16th century. Its illustrations are derived from the Turkmen style of Shīrāz but show clear Indian features adapted from the local version of the western Indian style.

Fifteenth century changes in western Indian style

Though the western Indian style was essentially conservative, it was not unfailingly so. It began to show signs of an inner change most notably in two manuscripts from Mandu, a *Kalpa-sūtra* and a *Kālakācāryakathā* of about 1439, and a *Kalpa-sūtra* painted at Jaunpur in 1465. These works were done in the opulent manner of the 15th century, but for the first time the quality of the line is different, and the uncompromisingly abstract expression begins to make way for a more human and emotional mood. By the opening years of the 16th century, a new and vigorous style had come into being. Although derived from the western Indian style, it is clearly independent, full of the most vital energy, deeply felt, and profoundly moving. The earliest dated example is an *Āraṇyaka Parva* of the *Mahābhārata* (1516; The Asiatic Society of Bombay), and among the finest are series illustrating the *Bhāgavata-Purāṇa* and the *Caura-*

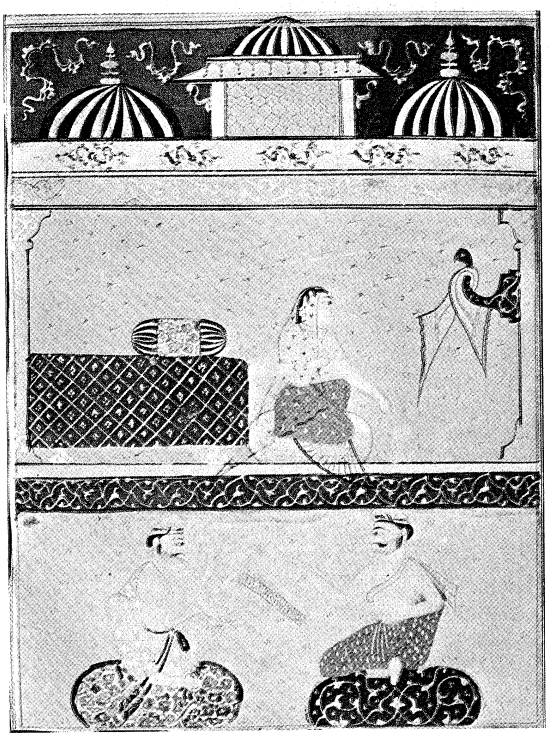
P. Chandra



Folio from a series illustrating the *Caurapañcāśikā* of Bilhaṇa, c. mid-16th century AD. In the Municipal Museum, Ahmadābād, Gujarāt, India.

pañcāśikā of Bilhaṇa, scattered in collections all over the world. A technically more refined variant of this style, preferring the pale, cool colours of Persian derivation, a fine line, and meticulous ornamentation, exists contemporaneously and is best illustrated by a manuscript of the ballad *Candāyana* by Mullā Dāūd (*c.* first half of the 16th century; Prince of Wales Museum of Western India, Bombay). The early 16th century thus appears to have been a period of inventiveness and set the stage for the development of the Mughal and Rājput schools, which thrived from the 16th to the 19th century.

Mughal style: Akbar period (1556–1605). Although the Mughal dynasty came to power in India with the great victory won by Bābar at the Battle of Pānīpat in 1526, the Mughal style was almost exclusively the crea-



Folio from an illustrated manuscript of *Candāyana*, c. first half of the 16th century. In the Prince of Wales Museum of Western India, Bombay.
P. Chandra

tion of Akbar. Trained in painting at an early age by a Persian master, Khwāja ʿAbd-uṣ-Ṣamad, who was employed by his father, Humāyūn, Akbar created a large atelier, which he staffed with artists recruited from all parts of India. The atelier, at least in the initial stages, was under the superintendence of Akbar's teacher and another great Persian master, Mīr Sayyid ʿAlī; but the distinctive style that evolved here owed not a little to the highly individual tastes of Akbar himself, who took an interest in the work, inspecting the atelier frequently and rewarding painters whose work was pleasing.

The work of the Mughal atelier in this early formative stage was largely confined to the illustration of books on a wide variety of subjects: histories, romances, poetic works, myths, legends, and fables, of both Indian and Persian origin. The manuscripts were first written by calligraphers, with blank spaces left for the illustrations. These were executed largely by groups of painters, including a colourist, who did most of the actual painting, and specialists in portraiture and in the mixing of colours. Chief of the group was the designer, generally an artist of top quality, who formulated the composition and sketched in the rough outline. A thin wash of white, through which the initial drawing was visible, was then applied and the colours filled in. The colourist's work proceeded slowly, the colour being applied in several thin layers and frequently rubbed down with an agate burnisher, a process that resulted in the glowing, enamel-like finish. The colours used were mostly mineral but sometimes consisted of vegetable dyes; and the brushes, many of them exceedingly fine, were made from squirrel's tail or camel hair.

The technique of Mughal painting

The earliest paintings (*c.* 1560–70) of the school of Akbar are illustrations of *Ṭūṭī nāmeh* ("Parrot Book; Cleveland Museum of Art) and the stupendous illustrations of the *Dāstān-e Amīr-Ḥamzeh* ("Stories of Amīr Ḥamza"; Österreichisches Museum für Angewandte Kunst, Vienna), which originally consisted of 1,400 paintings of an unusually large size (approximately 25 inches by 16 inches [65 by 40 centimetres]), of which only about 200 have survived. The *Ṭūṭī-nāmeh* shows the Mughal style in the process of formation: the hand of artists belonging to the various non-Mughal traditions is clearly recognizable, but the style also reveals an intense effort to cope with the demands of a new patron. The transition is achieved in the *Dāstān-e Amīr Ḥamzeh*, in which the uncertainties are overcome in a homogeneous style, quite unlike Persian work in its leaning toward naturalism and filled with swift, vigorous movement and bold colour. The forms are individually modelled, except for the geometrical ornament used as architectural decor; the figures are superbly interrelated in closely unified compositions, in which depth is indicated by a preference for diagonals; and much attention is paid to the expression of emotion. One of the last manifestations of this bold and vigorous early manner is the *Dārāb-nameh* (*c.* 1580) in the British Museum.

Immediately following were some very important historical manuscripts, including the *Tārīkh-e Khāndān-e Tīmūrīyeh* ("History of the House of Timūr," *c.* 1580–85; Khuda Baksh Library, Patna) and other works concerned with the affairs of the Tīmūrid dynasty, to which the Mughals belonged. Each of these manuscripts contains several hundred illustrations, the prolific output of the atelier made possible by the division of labour that was in effect. Historical events are recreated with remarkable inventiveness, though the explosive and almost frantic energy of the *Dāstān-e Amīr-Ḥamzeh* has begun to subside. The scale was smaller and the work began to acquire a studied richness. The narrative method employed by these Mughal paintings, like that of traditional literature, is infinitely discursive; and the painter did not hesitate to provide a fairly detailed picture of contemporary life—both of the people and of the court—and of the rich fauna and flora of India. Like Indian artists of all periods, the Mughal painter showed a remarkable empathy for animals, for through them flows the same life that flows through human beings. This sense of kinship allowed him to achieve unqualified success in the illustration of animal fables such as the *Anwār-e Suhaylī* ("Lights of Caropus"), of which several copies were painted, the earliest dated 1570 (School of Oriental and African Studies, London). It was in the illustrations to Persian translations of the Hindu epics, the *Mahābhārata* and the *Rāmāyaṇa*, that the Mughal painter revealed to the full the richness of his imagination and his unending resourcefulness. With little precedent to rely on, he was nevertheless seldom dismayed by the subject and created a whole series of convincing compositions. Because most of the painters of the atelier were Hindus, the subjects must have been close to their hearts; and, given the opportunity by a tolerant and sympathetic patron, they rose to great heights. It is no wonder, therefore, that the *Razm-nāmeh* (City Palace Museum, Jaipur), as the *Mahābhārata* is known in Persian, is one of the outstanding masterpieces of the age.

Illustrations for Persian translations of Hindu epics

In addition to large books containing numerous illustrations, which were the products of the combined efforts of many artists, the imperial atelier also cultivated a more intimate manner that specialized in the illustration of books, generally poetic works, with a smaller number of illustrations. The paintings were done by a single master artist who, working alone, had ample scope to display his virtuosity. In style the works tend to be very refined and delicate, with infinite attention to detail and filled with exquisite colour. A *Dīvān* ("Anthology") of Anwarī (Fogg Art Museum, Cambridge, Massachusetts), dated 1589, is a relatively early example of this manner. The paintings are very small, none larger than five inches by 2½ inches (12 by six centimetres) and most delicately executed. Very similar in size and quality are the miniatures illustrating the *Dīvān* of Ḥāfeẓ (Reza Library, Rāmpur). On a larger scale but in the same mood are the manuscripts that represent the most delicate and refined works of the reign of Akbar: the *Bahāristān* of Jāmī (1595; Bodleian Library, Oxford), a *Khamseh* of Neẓāmī (1593; British Museum, London), a *Khamseh* of Amīr Khosrow (1598; Walters Art Gallery, Baltimore and Metropolitan Museum of Art, New York), and an *Anwār-e Suhaylī* (1595–96; Bharat Kala Bhavan, Vārānasī).

Also prepared in the late 1590s were magnificent copies of the *Akbar-nāmeh* ("History of Akbar"; Victoria and Albert Museum, London) and the *Kitāb-e Changīz-*

nāmeh ("History of Genghis Khan"; Gulistan Library, Tehrān). These copiously illustrated volumes were produced by artists working jointly, but the quality of refinement is similar to that of the poetic manuscripts.

Of the large number of painters who worked in the imperial atelier, the most outstanding were Dasvant and Basāvan. The former played the leading part in the illustration of the *Razm-nāmeh*. Basāvan, who is preferred by some to Dasvant, painted in a very distinctive style, which delighted in the tactile and the plastic, and with an unerring grasp of psychological relationships.

Mughal style: Jahāngīr period (1605–27). The emperor Jahāngīr, even as a prince, showed a keen interest in painting and maintained an atelier of his own. His tastes, however, were not the same as those of his father, and this is reflected in the painting, which underwent a significant change. The tradition of illustrating books began to die out, though a few manuscripts, in continuation of the old style, were produced. For Jahāngīr much preferred portraiture; and this tradition, also initiated in the reign of his father, was greatly developed. Among the most elaborate works of his reign are the great court scenes, several of which have survived, showing Jahāngīr surrounded by his numerous courtiers. These are essentially large-scale exercises in portraiture, the artist taking great pains to reproduce the likeness of every figure.

Portraiture

The compositions of these paintings have lost entirely the bustle and movement so evident in the works of Akbar's reign. The figures are more formally ordered, their comportment in keeping with the strict rules of etiquette enforced in the Mughal court. The colours are subdued and harmonious, the bright glowing palette of the Akbarī artist having been quickly abandoned. The brushwork is exceedingly fine. Technical virtuosity, however, is not all that was attained, for beneath the surface of the great portraits of the reign there is a deep and often spiritual understanding of the character of the person and the drama of human life.

Many of the paintings produced at the imperial atelier are preserved in the albums assembled for Jahāngīr and his son Shāh Jahān. The Muraqqah-e Gulshan is the most spectacular. (Most surviving folios from this album are in the Gulistan Library in Tehrān and the Staatliche Museen Preussischer Kulturbesitz, Berlin; a section is temporarily housed in Tūbīngen.) There are assembled masterpieces from Iran, curiosities from Europe, works produced in the reign of Akbar, and many of the finest paintings of Jahāngīr's master painters, all surrounded by the most magnificent borders decorated with a wide variety of floral and geometrical designs. The album gives a fairly complete idea of Jahāngīr as a patron, collector, and connoisseur of the arts, revealing a person with a wide range of taste and a curious, enquiring mind.

P. Chandra



Painter at work, detail from a folio of the Muraqqah-e Gulshan, Mughal style, early 17th century AD. In the Staatliche Museen Preussischer Kulturbesitz, Berlin.

Jahāngīr esteemed the art of painting and honoured his painters. His favourite was Abū al-Hasan, who was designated Nādir-uz-Zamān ("Wonder of the Age"). Several pictures by the master are known, among them a perceptive study of Jahāngīr looking at a portrait of his father. Also much admired was Ustād Manṣūr, designated Nādir-ul-ʿAṣr ("Wonder of the Time"), whose studies of birds and animals are unparalleled. Bishandās was singled out by the emperor as unique in the art of portraiture. Manohar, the son of Basāvan Govardhan, and Daulat are other important painters of this reign.

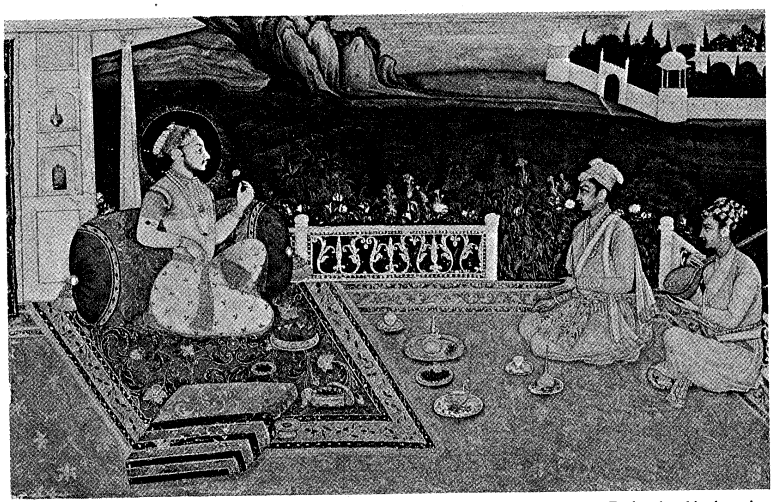
Mughal style: Shāh Jahān period (1628–58). Under Shāh Jahān, attention seems to have shifted to architecture, but painting in the tradition of Jahāngīr continued. The style, however, becomes noticeably rigid. The portraits resemble hieratic effigies, lacking the breath of life so evident in the work of Jahāngīr's time. The colouring is jewel-like in its brilliance, and the outward splendour quite dazzling. The best work is found in the *Shāhjahān-nāmeh* ("History of Shāh Jahān") of the Windsor Castle Library and in several albums assembled for the emperor. Govardhan and Bichitra, who had begun their careers in the reign of Jahāngīr, were among the outstanding painters; several works by them are quite above the general level produced in this reign.

Mughal style: Aurangzeb and the later Mughals (1659–1806). From the reign of Aurangzeb (1659–1707), a few pictures have survived that essentially continue the cold style of Shāh Jahān; but the rest of the work is nondescript, consisting chiefly of an array of lifeless portraits, most of them the output of workshops other than the imperial atelier. Genre scenes, showing gatherings of ascetics and holy men, lovers in a garden or on a terrace, musical parties, carousals, and the like, which had grown in number from the reign of Shāh Jahān, became quite abundant. They sometimes show touches of genuine quality, particularly in the reign of Muḥammad Shāh (1719–48), who was passionately devoted to the arts. This brief revival, however, was momentary, and Mughal painting essentially came to an end during the reign of Shāh ʿĀlam II (1759–1806). The artists of this disintegrated court were chiefly occupied in reveries of the past, the best work, for whatever it is worth, being confined to copies of old masterpieces still in the imperial library. This great library was dispersed and destroyed during the uprising of 1857 against the British.

Genre scenes

Company school. Rising British power, which assumed political supremacy in the 19th century, resulted in a radical change of taste brought about by the Westernization of important segments of the population. Heavily influenced by Western ideas, a style emerged that represented the adjustment of traditional artists to new fashions and demands. Rooted at Delhi and the erstwhile provincial Mughal capitals of Murshidābād, Lucknow, and Patna, it ultimately spread all over India. Most of the works produced were singularly impoverished, but occasionally there were some fine studies of natural life.

Deccani style. In mood and manner, Deccani painting, which flourished over much of the Deccan Plateau from at least the last quarter of the 16th century, is reminiscent of the contemporary Mughal school. Again, a homogeneous style evolved from a combination of foreign (Persian and Turkish) and Indian elements, but with a distinct local flavour. Of the early schools, the style patronized by the sultans of Bijāpur—notably the tolerant and art-loving Ibrāhīm ʿĀdil Shāh II of Bijāpur, famous for his love of music—is particularly distinguished. Some splendid portraits of him, more lyrical and poetic in concept than contemporary Mughal portraits, are to be found. A wonderful series depicting symbolically the musical modes (*rāgamālā*) also survives. Of illustrated manuscripts, the



Nobleman seated on a terrace, Mughal style painting, mid-18th century AD. In the National Museum of India, New Delhi.
P. Chandra

most important are the *Nujūm-ul-ʿulūm* ("The Stars of the Sciences," 1590; Chester Beatty Library, Dublin) and the *Tārīf-e Ḥuseyn-Shāhī* (Bharata Itihasa Samshodhaka Mandala, Pune), painted around 1565 in the neighbouring state of Ahmadnagar. The sultanate of Golconda also produced work of high quality—for example, a manuscript of the *Dīvān* of Muḥammad-Qulī Quṭb Shāh in the Salar Jang Library, Hyderābād, and a series of distinguished portraits up to the end of the 17th century (dispersed in various collections). The state of Hyderābād, founded in the early 18th century and headed by a grandee of the Mughal Empire, was a great centre of painting. The work that was produced there reflects both Golconda traditions and increasing Mughal and Rajasthani influences.

Rajasthani style. This style appears to have come into being in the 16th century, about the same time the Mughal school was evolving under the patronage of Akbar; but, rather than a sharp break from the indigenous traditions, it represented a direct and natural evolution. Throughout the early phase, almost up to the end of the 17th century, it retained its essentially hieratic and abstract character, as opposed to the naturalistic tendencies cultivated by the Mughal atelier. The subject matter of this style is essentially Hindu, devoted mainly to the illustration of myths and legends, the epics, and above all the life of Kṛṣṇa; particularly favoured were depictions of his early life as the cowherd of Vraja, and the mystical love of Vraja's maidens for him, as celebrated in the *Bhāgavata-Purāṇa*, the *Gītagovinda* of Jayadeva, and the Braj Bhasa verses written by Sūrdās and other poets. The style of the painting, no less than the literature, is a product of the new religious movements, all of which stressed personal devotion to Kṛṣṇa as the way to salvation. Related popular themes were pictorial representations of the musical modes (*rāgamālā*) and illustrations of poetical works such as the *Rasikapriyā* of Keśavadāsa, which dealt with the sentiment of love, analyzing its varieties and endlessly classifying the types of lovers and beloveds and their emotions. Portraits, seldom found in the early phase, became increasingly common in the 18th century—as did court scenes, scenes of sporting and hunting events, and other scenes concerned with the courtly life of the great chiefs and feudal lords of Rājasthān.

The Hindu basis of Rajasthani style

P. Chandra



The musical mode *vasanta*, Deccani school painting, Bijāpur, late 16th century AD. In the National Museum of India, New Delhi.

The Rajasthani style developed various distinct schools, most of them centring in the several states of Rājasthān, namely Mewār, Būndi, Kotah, Mārwār, Bīkaner, Kishangarh, and Jaipur (Amber). It also had centres outside the geographical limits of present-day Rājasthān, notably Gujarāt, Mālwa, and Bundel Khand. The study of Rajasthani painting is still in its infancy, for most of the material has been available for study only since the mid-1940s.

The Mughal and Rajasthani styles were always in contact with each other, but in general the Rajasthani schools were not essentially affected by the work produced at the Mughal court during the greater part of the 17th century. This became less true in the 18th century, when the sharp distinction between the two became progressively obscured, though each retained its distinctive features right up to the end.

Rajasthani style: Mewār. The Mewār school is among the most important. The earliest dated examples are represented by a *rāgamālā* series painted at Chawand in 1605 (Gopi Krishna Kanoria Collection, Patna). These simple paintings, filled with bright colour, are only a step removed from the pre-Rajasthani phase. The style became more elaborate in the first quarter of the 17th century when another *rāgamālā*, painted at Udaipur in 1628 (formerly in the Khajanchi Collection, Bīkaner; now dispersed in various collections), showed some superficial acquaintance with the Mughal manner. This phase, lasting until around 1660, was one of the most important for the development of painting all over Rājasthān. Ambitious and extensive illustrations of the *Bhāgavata*, the *Rāmāyaṇa*, the poems of Sūrdās, and the *Gītagovinda* were completed, all full of strength and vitality. The

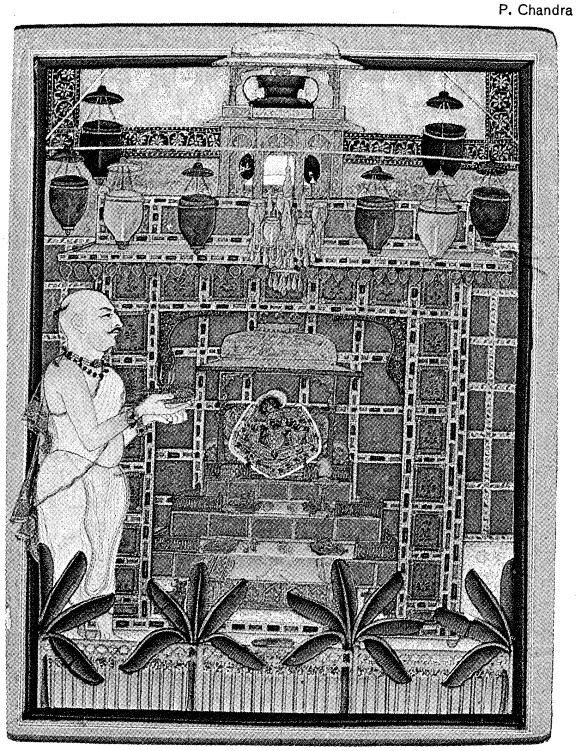
name of Sāhabadī is intimately connected with this phase; another well-known painter is Manohar. The intensity and richness associated with their atelier began to fade toward the close of the 17th century, and a wave of Mughal influence began to affect the school in the opening years of the 18th century. Portraits, court scenes, and events in the everyday world of the ruling classes are increasingly found. Although the emotional fervour of the 17th century was never again attained, this work is often of considerable charm. The 19th century continued to create work in the tradition of the 18th, one of the most important centres being Nāthdwāra (Rājasthān), the seat of the Vallabha sect. Large numbers of pictures, produced here for the pilgrim trade, were spread over all parts of Rājasthān, northern India, Gujarāt, and the Deccan.

Rajasthani style: Būndi and Kotah. A school as important as that of Mewār developed at Būndi and later at Kotah, which was formed by a partition of the parent state and ruled by a junior branch of the Bundi family. The earliest examples are represented by a *rāgamālā* series of extraordinarily rich quality, probably dating from the end of the 16th century. From the very beginning the Būndi style seemed to have found Mughal painting an inspiring source, but its workmanship was as distinctively Rajasthani as the work of Mewār. The artists of this school always displayed a pronounced preference for vivid movement, which is unique in all of Rājasthān. Toward the second half of the 17th century, work at Būndi came unmistakably under the influence of Mewār; many miniatures, including several series illustrating the *Rasikapriyā*, indicate that this was a period of prolific activity. The sister state of Kotah also appears to have become an important centre of painting at this time, developing a great fondness for hunting and sport scenes, all filled with great vigour and surging strength. This kind of work continued well into the 19th century, and if the workmanship is not always of the highest quality, the style maintained its integrity against the rapidly increasing Western influence right up to the end.

Earliest examples of Būndi school

Rajasthani style: Mālwa. It has been suggested but not definitely determined that the school itself does not belong to Mālwa but to some other area, probably Bundelkhand. In contrast to the Būndi school, miniatures generally thought to have been painted in Mālwa are quite

P. Chandra



A priest at worship, Rajasthani style painting, Nāthdwāra, mid-19th century AD. In a private collection.



Lion hunt, Rajasthani style painting, Kotah, late 18th century AD. In a private collection.
P. Chandra

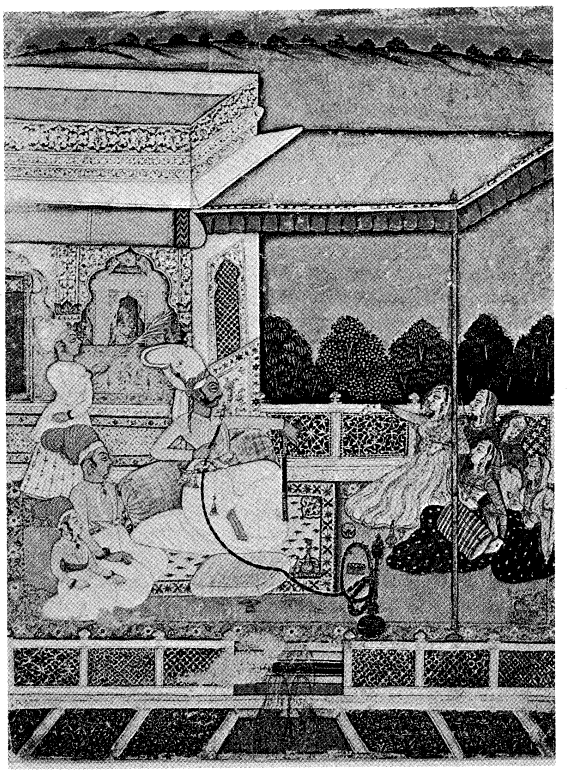
archaistic, with mannerisms inherited from the 16th century retained until the end of the 17th. The earliest work is an illustrated version of the *Rasikapriyā* (1634), followed by a series illustrating a Sanskrit poem called the *Amaru Śataka* (1652). There are also illustrations of the musical modes (*rāgamālā*), the *Bhāgavata-Purāṇa*, and other Hindu devotional and literary works. The compositions of all of these pictures is uncompromisingly flat, the space divided into registers and panels, each filled with a patch of colour and occupied by figures that convey the action. This conservative style disappeared after the close of the 17th century. The course of Mālwa painting in the 18th century and later is not known.

Rajasthani style: Mārwār. A *rāgamālā* series dated 1623 reveals that painting in this state shared features common to other schools of Rājasthān. Miniatures of the second half of the 17th century are distinguished by some splendid portraits that owe much to the Mughal school. A very large amount of work was done in the 19th century, all of which is highly stylized but strong in colour and often of great charm.

Rajasthani style: Bīkaner. Of all Rajasthani schools, the Bīkaner style, from its very inception in the mid-17th century, shows the greatest indebtedness to the Mughal style. This is due to the presence in the Bīkaner atelier of artists who had previously worked in the Mughal manner at Delhi. They and their descendants continued to paint in a style that could only be classed as a provincial Mughal manner had it not been for the quick absorption of influences from the Rajasthani environment and a sympathy for the religious and literary themes favoured by the royal Hindu patrons. Delicacy of line and colour are consistent characteristics of Bīkaner painting even when, toward the end of the 18th century, it assumed stylistic features associated with the more orthodox Rajasthani schools.

Characteristics of the Bīkaner style

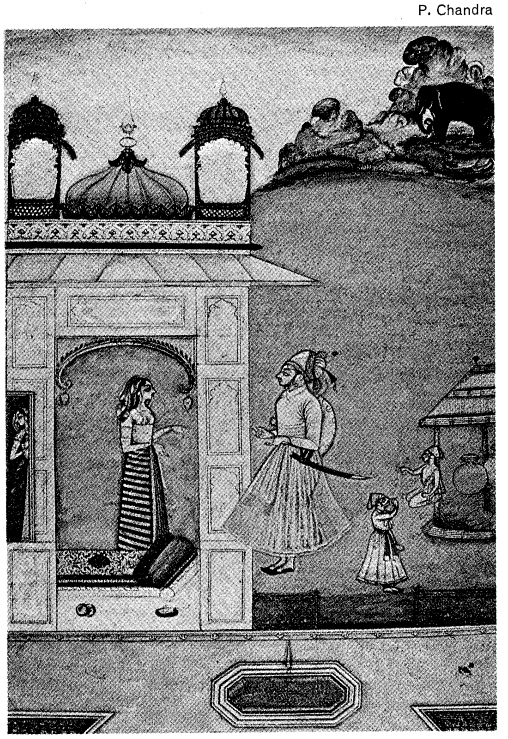
Rajasthani style: Kishangarh. The Kishangarh school, which came into being toward the mid-18th century, was also indebted to the contemporary Mughal style but combined a rich and refined technique with deeply moving religious fervour. Its inspiring patron in the formative phase was Sāvant Singh, more of a devotee and a poet than a king. The style established by him, characterized by pronounced mystical leanings and a distinctive facial type, continued to the middle of the 19th century, though at a clearly lower level of achievement.



A Rājput king listening to music, Rajasthani style painting, Mārwār, mid-18th century AD. In a private collection.
P. Chandra

Rajasthani style: Jaipur (Amber). The rulers of the state were closely allied to the Mughal dynasty, but paintings of the late 16th and early 17th centuries possessed all of the elements of the Rajasthani style. Little is known about the school until the opening years of the 18th century, when stiff, formal examples appear in the reign of Savāī Jai Singh. The finest works, dating from the reign of Pratāp Singh, are sumptuous in effect and include some splendid portraits and some large paintings of the sports of Kṛṣṇa. Although the entire 19th century was extremely productive, the work was rather undistinguished and increasingly affected by Western influences. Of the Rajasthani styles of this period, the Jaipur school was the most popular, examples having been found all over northern India.

P. Chandra



Month of summer, Rajasthani-style painting, Bīkaner, early 18th century AD. In a private collection.

Pahari style. Closely allied to the Rajasthani schools both in subject matter and technique is the Pahari style, so-named because of its prevalence in the erstwhile hill states of the Himalayas, stretching roughly from Jammu to Gaṛhwāl. It can be divided into two main schools, the Basohlī and the Kāngra, but it must be understood that these schools were not confined to the centres after which they are named but extended all over the area. Unlike Rājasthān, the area covered by the Pahari style is small, and the probability of artists travelling from one area to another in search of livelihood was much greater. Thus, attempts to distinguish regional schools are fraught with controversy, and it has been suggested that a classification based upon ateliers and families is likely to be more tenable than those presently current among scholars. Because the Basohlī and the Kāngra schools show considerable divergences, scholars have postulated the existence of a transitional phase, named the pre-Kāngra style.

Pahari style: Basohlī school. The origins of this remarkable style are not yet understood, but it is clear that the style was flourishing toward the close of the 17th century. The earliest dated paintings are illustrations to the *Rasamañjarī* of Bhānudatta (a Sanskrit work on poetics), executed for a ruler of Basohlī (1690; Boston Museum of Fine Arts). Bold colour, vigorous drawing, and primitive intensity of feeling are outstanding qualities in these paintings, quite surpassing the work of the plains. In addition to other Hindu works such as the *Gītagovinda* and the *Bhāgavata-Purāṇa*, a fairly large number of idealized portraits have also been discovered.

Pahari style: Kāngra school. The Basohlī style began to fade by the mid-18th century, being gradually replaced by the Kāngra style, named after the state of Kāngra but, like the Basohlī style, of much wider prevalence. A curvilinear line, easy flowing rhythms, calmer colours, and a mood of sweet lyricism easily distinguish the work from that of the Basohlī style. The reasons for this change are to be sought in strong influences from the plains, notably the Mughal styles of Delhi and Lucknow. These influences account for the more refined technique; but whatever was borrowed was transmuted and given a fresh and tender aspect. Among the greatest works are large series illustrating the *Bhāgavata-Purāṇa* (National Museum, New Delhi), the *Gītagovinda*, and the *Satsaī* of Bihārī (both in the collection of the maharaja of Tehrī-Garhwāl), all painted in 1775–80. The corpus of work produced is very large and, although it seldom fails to please, works of high achievement are rare. The school flourished from about 1770 to almost the end of the 19th century, but the finest work was produced largely around 1775–1820.

Modern period. Toward the late 19th century traditional Indian painting was rapidly dying out, being replaced by feeble works in a variety of idioms, all strongly influenced by the West. A reaction set in during the early 20th century, symbolized by what is called the Bengal school. The glories of Indian art were rediscovered, and the school consciously tried to produce what it considered a truly Indian art inspired by the creations of the past. Its leading artist was Rabindranath Tagore and its theoretician was E.B. Havell, the principal of the Calcutta School of Art. Nostalgic in mood, the work was mainly sentimental though often of considerable charm. The Bengal school, whatever its other merits, did a great deal to reshape contemporary taste and to make Indian artists aware of their own heritage. Amrita Sher-Gil, who was inspired by the Postimpressionists, made Indian painters aware of new directions. Mid-20th-century Indian painting is very much a part of the international scene, the artists painting in a variety of idioms, often attempting to come to terms with their heritage and with the emergence of India as a modern culture.

The Bengal school

Decorative arts. Fragmentary ivory furniture (*c.* 1st century AD) excavated at Begrām is one of the few indica-

tions of the existence in ancient India of a secular art concerned with the production of luxurious and richly decorated objects meant for daily use. Objects that can be clearly designated as works of decorative art become much more extensive for the later periods, during which Islāmic traditions were having a profound effect on Indian artistic traditions. The reign of the Mughal emperors, in particular, produced works of the most elaborate and exquisite craftsmanship; the decorative tradition is clearly preserved in architectural ornament, though surviving decorative objects themselves, particularly before the 17th century, are far fewer than might be expected. Economic conditions, including competition with machine-made goods imported from English factories, and a change in taste from increasing European influence had disastrous consequences for traditional craftsmanship, especially in the late 19th and 20th centuries.

Pre-Islāmic period. Of the very few objects surviving from the pre-Islāmic period, the most important are fragments of ivory caskets, chairs, and footstools found at Begrām, in eastern Afghanistan, but obviously of Indian origin and strongly reminiscent of the school of Mathurā in the 1st century AD. The work is profusely decorated with carved panels and confirms the wide reputation for superb ivories that India had in ancient times. Nothing as spectacular has come down from the succeeding periods, but stray examples such as the so-called Charlemagne chessman (*c.* 8th century; Cabinet des Medailles, Paris) and two magnificent throne legs, of Orissan workmanship, carved in the shape of griffins with elephant heads (13th century; Freer Gallery of Art, Washington D.C., and Philadelphia Museum of Art), indicate that ivory craftsmanship was always vital. Ancient traditions, relatively unaffected by Islāmic influence, continued in southern India up to modern times. An exquisitely carved box from Vijayanagar (16th century; Prince of Wales Museum of Western India, Bombay) is representative; many other exquisite objects of this period and later are among the treasures of South Indian temples.

Work in ivory

There is even less evidence of what the decorative work in metal was like. The practice of re-using the metal by melting unserviceable items may account for the paucity of objects, for there is little doubt that the craft was always flourishing. A hoard found at Kolhāpur, consisting of plates, various kinds of vessels, lamps and *objets d'art*, including a superb bronze elephant with riders, constitutes the most important surviving group of metal objects and is datable to about the 2nd century AD. Some fine examples of ritual utensils, notably elaborate incense burners, of the 8th–9th century have been excavated at Nālandā; and a large number of 14th-century ceremonial vessels of complex design and excellent workmanship, and apparently belonging to the local temple, were discovered at Kollur, in Mysore state.

Gold played an extremely important role in the manufacture of jewelry, but once again the finds are hardly commensurate with tradition. Small amounts of gold jewelry have been excavated at Mohenjo-daro and Harappā (3rd millennium BC); and, in the historical period, a very important group, of delicate workmanship, has been excavated at Taxila (*c.* 2nd century AD).

From earliest times, India has been famous for the variety and magnificence of its textiles. In this case, however, the Indian climate has been particularly destructive; virtually nothing has survived the heat and moisture. Besides the testimony of literature and the evidence of figural sculpture, only a few fragments of printed textiles are preserved—at Fusṭāṭ in Egypt, where they had been exported. These date approximately to the 14th century.

Islāmic period. Traditions of craftsmanship established during the Islāmic period came to full flower during the reign of the Mughal dynasty. Surviving works of decorative art are more abundant, though once again there are hardly as many examples as might be expected, particularly from the 16th and 17th centuries. According to literary testimony and the few available examples, the finest objects were undoubtedly made in the imperial workshops set up in large number at the capital and in the great cities of the empire, where they were nourished by



Ritual vessel from Kollur, Mysore, India, copper, *c.* 14th century AD. In the Prince of Wales Museum of Western India, Bombay.
P. Chandra

local traditions. Well-organized, these shops specialized in particular items, such as textiles, carpets, jewelry, ornamental arms and armour, metalware, and jade. Textile manufacture must have been enormous, considering the demands of court and social etiquette and ritual. Contributing to the popularity of tapestries, curtains, draperies, canopies, and carpets in contemporary architecture were the nomadic tenting traditions of the Mughal rulers.

Popularity of textiles

The variety of techniques employed in the manufacture of textiles was infinite, ranging from printed and painted patterns to the exquisite embroidery decoration of woolen shawls and the costly figured brocading of silk. An important contribution to carpet weaving was the landscape carpet that reproduced pictorial themes inspired by miniature painting. Much of the surviving textile work dates from the 18th century or later, though the 16th and 17th centuries produced works of the most outstanding quality.

In response to growing European trade, a considerable amount of furniture (chairs, cabinets, chests of drawers, and the like) was produced, mostly wood inlaid with ivory. Many of these pieces have been preserved in the kinder European climate. Although the furniture made for export gives some idea of the craft in India, it must be emphasized that only the ornamental and figural work was Indian, while the form was European. Also in a hybrid Indo-European style were the Christian objects produced by a local school of ivory carvers at Goa.

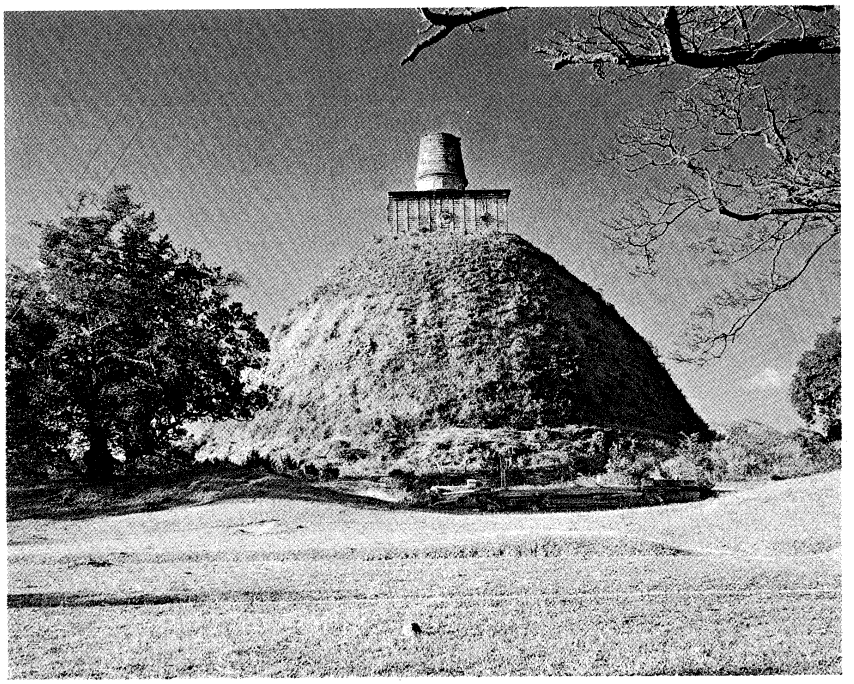
Metal objects of sumptuous quality were also made, a unique example of which is a splendid, elaborately chiselled 16th-century cup in the Prince of Wales Museum of Western India in Bombay. This tradition was continued in the 17th and particularly the 18th century, when vessels made of a variety of metals and adorned with engraved, chiselled, inlaid, and enamelled designs were very popular. Arms and armour, in particular, were decorated with the skill of a jeweler. Particularly striking are the carved hilts, often done in animal shapes.

Jade or jadeite was much fancied by the rich and was

P. Chandra



Detail of a Kashmir shawl, wool embroidery, 19th century AD. In the Prince of Wales Museum of Western India, Bombay.



Abhayagiri *dagāba* in the Jetavana at Anurādhapura, Sri Lanka, *c.* 4th century AD.
ZEFA

used together with crystal to make precious vessels as well as sword and dagger hilts. A rather large number of 18th- and 19th-century objects have survived, but they are often of nondescript quality. The greatest period for jade carving seems to have been the 17th century; a few outstanding examples associated with the emperors Jahāngīr and Shāh Jahān are of singular delicacy and perfection. The practice of inlaying jade, and also stone, with precious or semiprecious stones became more popular with the reign of Shāh Jahān and increasingly characteristic of Indian jade craftsmanship from that time on.

Architectural decoration provides a clear idea of the range of ornamental patterns used by the Mughal artist. They consisted mainly of arabesques (intricate interlaced patterns made up of flower, foliage, fruit, and sometimes animal and figural outlines) and infinitely varied geometric patterns—motifs shared with the rest of the Muslim world—together with floral scrolls and other designs adapted from Indian traditions. As a whole, the Mughal decorative style tends to endow ornamental patterns with a distinctive plasticity not seen in the more truly two-dimensional Iranian and Arab work. From the 17th century, a type of floral spray became the most favoured motif and was found on almost every decorated object. The motif, symmetrical but relatively naturalistic at the beginning, became progressively stiff and stylized, but never lost its importance in the ornamental vocabulary.

VISUAL ARTS OF SRI LANKA (CEYLON)

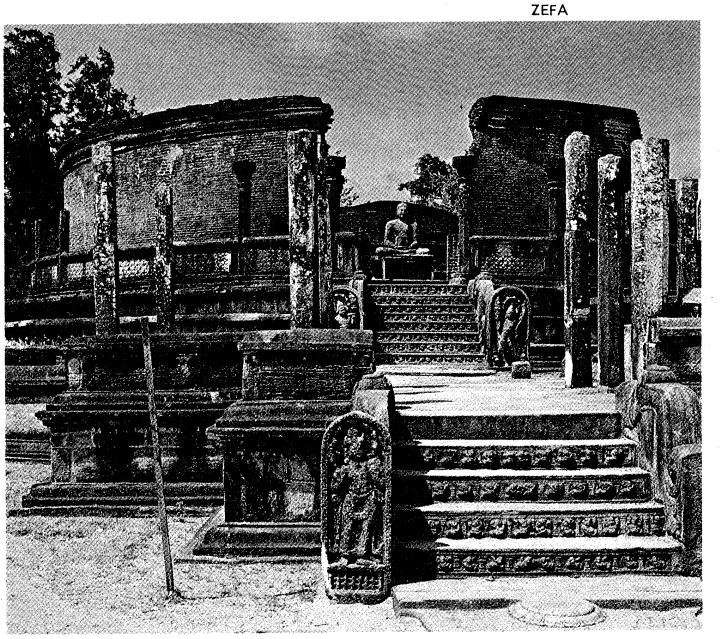
Distinction between art of Sri Lanka and that of India

The art of Sri Lanka is closely allied to that of India but presents several distinctive features that make a separate treatment convenient. There is, first, the considerable transformation of Indian influences, resulting in an idiom of great power and individuality. Sri Lanka also often served as a geographical pocket in which styles that had disappeared in India were preserved, which accounts for the anachronistic features of some phases of Sinhalese art. It also appears that although predominant influences were from neighbouring southern India, this was not exclusively so, and styles flourishing in western and northern India, too, contributed to the formulation of Sinhalese art. The difficulties in the study of the art are considerable: the long, unbroken Buddhist traditions and the piety of the rulers and the people have led to the successive renovation of monuments; and in the absence of firmly dated monuments, one of the few relatively reliable tools of study is comparison with Indian art, an approach full of pitfalls and shortcomings.

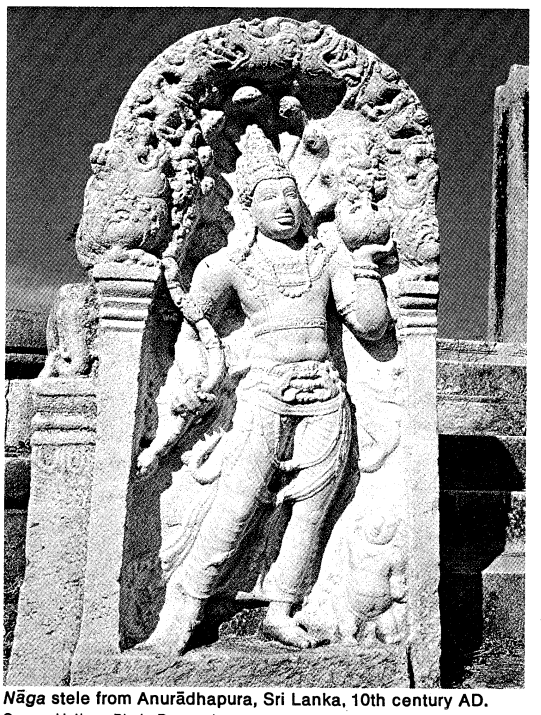
Architecture. The most impressive monuments are the great *stūpa*s, some of gigantic size and considerable antiquity but often reconstructed in the course of the centuries. They generally have a triple circular base, and as in early Indian *stūpa*s, a hemispherical dome with a miniature railing on top, and a multiple parasol that tends to solidify into a conical structure in the course of time. The material is brick, sometimes covered with plaster and white paint. An important feature are the platforms (*vāhalakaḍa*s) at the cardinal points, often adorned with sculpture. There are many *stūpa*s at the ancient capital of Anurādhapura, at Polonnaruva, and at other sites; of these the Jetavana at Anurādhapura is the largest, though now largely ruined.

Small *stūpa*s were often placed in a circular building with a domical metal and timber roof supported by concentric rows of stone pillars. This type of building, known in ancient India as the *caityagṛha*, was very popular in Sri Lanka, though it had disappeared at a fairly early pe-

ZEFA



The *vaṭadāgē* at Polonnaruva, Sri Lanka, 12th century AD.



Nāga stele from Anurādhapura, Sri Lanka, 10th century AD.
George Holton—Photo Researchers

riod in the country of its origin. A famous example is the *vaṭadāgē* at Polonnaruva, a structure of great elegance. The dome itself, being of perishable material, has not survived. The *geḍigē*, or large rectangular hall with a corbelled brick vault, housing a Buddha image, is first found in Sri Lanka from the 8th century AD; the most impressive example is the Laṅkātilaka at Polonnaruva built by Parākramabāhu I in the 12th century.

Literature testifies to the existence of elaborate royal and priestly residences of wood, which have largely disappeared. The Lohapāsāda at Anurādhapura, traditionally ascribed to Duṭṭhagāmaṇi (101–77), was originally a nine-story building, now destroyed except for the large number of stone pillars that supported the upper floors. Sigiriya, a 6th-century fortress city with extensive remains, is another notable example of secular architecture.

Sculpture. The earliest sculpture, perhaps, is from the platforms, or *vāhalakaḍas*, of the Kaṇṭaka Cetiya, at Mihintalē, and reveals an archaistic style indebted to 1st-century-BC Indian sculpture of Sāñcī and Amarāvatī regions. A certain simplicity and restraint characteristic of most Sinhalese work is present even at this early stage. The first Buddha images show a pronounced relationship to examples from Andhradeśa of the 2nd–3rd century AD but often possess considerable vigour, revealing the contribution of the local sculptor. Several fine images are known, one of the best of which is at Ruanveli, Anurādhapura, now very badly restored.

Dated monuments are absent from the 5th to the 12th centuries, but an approximate idea of stylistic development can be obtained by a comparative study of Indian examples. An outstanding image, rather hideously repaired in recent years, is a great seated Buddha in Anurādhapura, the smooth and abstract modelling of which recalls the school of Sārnāth of the 5th–6th century. At Isurumuni, near Anurādhapura, are some marvellous reliefs carved on rocks. One of these depicts elephants at play, and another, a seated man with the head of a horse carved in the background. These fine sculptures recall the South Indian style of the 7th century. A radiant amorous couple carved in relief on a stone slab, also at Isurumuni, represents Sinhalese sculpture at its most joyous.

Sculptured staircases

Of about the same period or a little later, are exquisitely sculptured staircases decorated with moonstones, and stelae, or commemorative pillars, carved with a guardian *nāga*, a spirit with combined superhuman and serpent qualities. The latter are among the finest examples of Sinhalese sculpture, the full and weighty modelling relieved by the skillful movement of clearly chiselled ornament. The Ratnapāsāda at Anurādhapura and the eastern staircase of the *vaṭadāgē* at Polonnaruva possess particularly superb specimens. Moonstones—decorated with bands of floral motifs, geese, and a row of animals consisting of a lion, bull, elephant, and horse—placed at the bottom of the staircase, testify to the great tase and elegance that mark Sinhalese decorative carving. At Anurādhapura and related sites a certain freedom characterizes the work, while the slightly later examples at Polonnaruva are stiffer but technically brilliant.

A colossal Buddha, 42 feet (13 metres) high, at Avukana, testifies to the increasing hardness of the Sinhalese style, which, even so, never ceases to be moving. Large images of the Buddha at the Gal Vihāra and a figure supposedly representing Parākramabāhu at Potgal Vihāra, both at Polonnaruva, are of the 12th century. They are figures of great majesty and surpass contemporary work in southern India. After the 13th century, Sinhalese sculpture began to decline, though work of some decorative value was produced up to the 19th century.

Painting. The rock at Sigiriya is adorned with a series of exquisitely painted *apsarases* (nymphs) showering flowers, their torsos emerging from clouds. The paintings

P. Chandra



Painted figure of an *apsaras* from Sigiriya, Sri Lanka, 6th century AD.

are dated to the 6th century AD; in their plastic resiliency they are reminiscent of contemporary work in India. The next important group of wall paintings come from Tivaṁka-patimā-ghara at Polonnaruva. Although dated to the 12th or 13th century, the figures continue to be modelled, relatively unaffected by the linear distortions of the western Indian style that was flourishing in India. Eighteenth-century paintings, with their flat figures arranged in horizontal rows, reflect contemporary styles of southern India.

(P.Ch.)

BIBLIOGRAPHY

Literature: (*Sanskrit, Pāli, and Prākrit*): The old literature of Southeast Asia has been more extensively described than any of the modern literatures. Though antiquated, MORIZ WINTERNITZ, *Geschichte der indischen litteratur*, 3 vol. (1908–22; Eng. trans., *A History of Indian Literature*, 2nd ed. 1959–67), is still very useful. So is ARTHUR B. KEITH, *A History of Sanskrit Literature* (1928, reprinted 1956), which, however, does not include the Sanskrit theatre. For the old literature of the Veda, ARTHUR A. MACDONNELL, *A History of Sanskrit Literature* (1900), remains very helpful as a survey. A full inventory of the literature is given by S.N. DAS GUPTA and S.K. DE, *A History of Sanskrit Literature, Classical Period*, 2nd ed. (1962). The epochal material is best treated by EDWARD W. HOPKINS, *The Great Epic of India* (1901). On the

Sanskrit play specifically, the study by the French scholar SYLVAIN LEVI, *Le Théâtre indien* (1890), remains an important contribution. The best introduction to the narrative literature is to be found in CHARLES H. TAWNEY (trans.), *The Ocean of Story*, ed. by NORMAN M. PENZER, 10 vol. (1923–28, reprinted 1968).

Modern Indian literatures: The literatures in the modern Indian languages are bibliographically underrepresented in English. For some of them the best sources are to be found written in those languages themselves. The following bibliography must confine itself to works written in more accessible languages. (*Hindi*): A good though incomplete inventory of the older literature in Hindi and Hindustani is found in JOSEPH GARCIN DE TASSY, *Histoire de la littérature hindouie et hindoustanie*, 2nd ed., 3 vol. (1870–71, reprinted 1968). A useful guide up to its date is EDWIN GREAVES, *A Sketch of Hindi Literature* (1918). A survey of the more modern literature in Hindi is R.A. DWIVELDI, *A Critical Survey of Hindi Literature* (1966). (*Assamese*): Among the few works available is B.K. BARUA, *Assamese Literature* (1941) and *A History of Assamese Literature* (1965). (*Bengali*): The best extensive introduction to the literature in Bengali is that of SUKUMAR SEN, *History of Bengali Literature* (1960). Another useful source is J.C. GHOSH, *Bengali Literature* (1948). (*Marathi*): Many works on Marathi literature are written in Marathi itself. A good guide is G.C. BHATE, *History of Modern Marathi Literature, 1800–1938* (1939). (*Gujarati*): Little is written in English on Gujarati literature. To be recommended is K.M. MUNSHI, *Gujarāt and Its Literature from Early Times to 1852*, 3rd ed. (1967). A useful older study is K.M. JHAVERI, *Milestones in Gujarati Literature* (1914). (*Punjabi*): A good survey of Punjabi is given by MOHAN SINGH, *An Introduction to Panjabi Literature* (1951). For information about literature in the smaller Indo-Aryan languages, see the symposium of the All-India Writers' Conference, *Writers in Free India* (1950). (*Muslim contributions*): The best short introduction to the cultural and intellectual life of the Moslems of the subcontinent is AZIZ AHMAD, *An Intellectual History of Islam in India* (1969). A comprehensive survey of the literature in Urdu produced in South India, especially in the medieval kingdoms of Bijāpur and Golconda and in the later state Hyderābād is the Urdu work *Dekan mem Urdu* by NASIR-RUDDIN HASHMI (1963). A study of the convention of classical Urdu poetry in the light of the writings of three Urdu poets of 18th-century Delhi is RALPH RUSSELL and KHURSHIDUL ISLAM, *Three Mughal Poets: Mir, Sauda, Mir Hasan* (1968). The latest and most comprehensive survey of literature produced in Persian in various countries, including India and Pakistan, is JAN RYPKA, *History of Iranian Literature* (1968). A useful introduction is MUHAMMAD SADIQ, *A History of Urdu Literature* (1964). (*Tamil*): Much useful information is found in XAVIER S. THANI NAYAGAM, *A Reference Guide to Tamil Studies* (1966). Fuller treatment is given in C. and H. JESUDASAN, *A History of Tamil Literature* (1961). Especially recommended are also T.P. MEENAKSHISUNDARAM, *A History of Tamil Literature* (1965); and J.M. SOMASUNDARAM PILLAI, *A History of Tamil Literature with Texts and Translations from the Earliest Times to 600 A.D.* (1968). A good account of Telugu literature is P. CHENCHAYYA and R.M. BHUJANGA RAO BHADUR, *A History of Telugu Literature* (1928). More recent is P.T. RAJU, *Telugu Literature* (1944); and GIDUGU VENKAJA SITAPATI, *History of Telugu Literature* (1968). Very little has been written on Kannada literature. Mention can be made of the older EDWARD P. RICE, *A History of Kanarese Literature*, 2nd ed. rev. and enl. (1921); and of H. THIPPERUDA SWAMY, *The Vīraśaiva Saints* (1968). The literature in Malayalam is sketched in K.M. GEORGE, *A Survey of Malayalam Literature* (1968). Further mention can be made of K.K. NAIR, *A History of Malayalam Literature* (1971); and of P.K. PARAMESWARAN NAIR, *History of Malayalam Literature* (Eng. trans. 1967). Finally, among the very few books in English on the literature of Ceylon (Sri Lanka), prominent mention can be made of CHARLES EDMUND GODAKUMBURA, *The Literature of Ceylon* (1963).

Music: ARNOLD A. BAKE, "The Music of India," in *The New Oxford History of Music*, vol. 1 (1957), a general chapter on Indian music dealing with the philosophical background, Vedic chant, the ancient musical system, the modern classical system, and musical instruments; ELISE B. BARNETT, "Special Bibliography: The Art Music of India," *Ethnomusicology*, 14:278–312 (1970), a listing of books and articles on Indian music, published since 1959, which includes some publications on folk and religious music, dance, and drama; SUDHIBHUSHAN BHATTACHARYA, *Ethnomusicology and India* (1968), a synchronic approach attempting to relate folk, tribal, religious, and classical music in terms of the cultural background—includes outline notations in Indian *sargam*; ALAIN DANIELOU, *The Rāga-s of Northern Indian Music* (1968), an individualistic interpretation of modern North Indian *ragas*, based partly on ancient theory; B.C. DEVA, *Psychoacoustics of Music and Speech* (1967), a collection of articles on various aspects of music and speech, with emphasis on acoustics and the scientific study of Indian music; ARTHUR H. FOX-STRANGWAYS, *The Music of Hindostan* (1914, reprinted 1965), a work of wide scope including discussion of Vedic chant, folk music, and modern Indian classical music (includes notations in Western staff and analogies with Western music); O.C. GANGOLY, *Rāgas and Rāginīs*, 2 vol. (1934–35, reprinted 1948), a historical study that traces the systems of classifying *ragas* in Indian musical treatises, including a discussion of the time-theory of *ragas* as well as their iconography; NAZIR A. JAIRAZBHOY, *The Rāgas of North Indian Music* (1971), a technical work dealing with the structure and evolution of North Indian *ragas* (includes a 45 r.p.m. record of *raga* demonstrations performed on the *sitar* by VILAYAT KHAN, and extensive notations in Western staff and Indian *sargam*); BABURAO JOSHI and A. LOBO, *Introducing Indian Music* (n.d.), a series of four long-playing records, including musical examples and commentary, illustrating the main features of North Indian classical music, with accompanying booklet; WALTER KAUFMANN, *The Rāgas of North India* (1968), description and notations, in Western staff, of about 230 *ragas* of modern North Indian music; ALLEN KEESE, *The Sitar Book* (1968), an elementary guide to the *sitar*, with some description of playing techniques, musical exercises, and *gats* in ten *ragas;* HERBERT A. POPLEY, *The Music of India*, 3rd ed. (1970), a general work including discussion of historical background, scale, *raga*, *tala*, musical form, and instruments, with notations in Western staff and Indian *sargam;* HAROLD S. POWERS, "Indian Music and the English Language: A Review Essay," *Ethnomusicology*, 9:1–12 (1965), a survey of the most important literature on Indian music written in the English language; and "An Historical and Comparative Approach to the Classification of Rāgas (with an Appendix on Ancient Indian Tunings)," in *Selected Reports* of the Institute of Ethnomusicology, UCLA, 1:1–78 (1970), a scholarly monograph that attempts to show the relationship between North and South Indian *ragas* that bear the same name but now differ in scale—includes material on ancient Indian music and gives many musical examples in Western staff; SWAMI PRAJNANANDA, *A History of Indian Music*, vol. 1 (1963), a technical work dealing with the origins and the music of the ancient period; P. SAMBAMOORTHY, *South Indian Music*, 5 vol. (1958–69), a comprehensive work covering many aspects of South Indian musical theory, both synchronically and diachronically; RAVI SHANKAR, *My Music, My Life* (1969), a general book combining autobiographical and biographical material with a discussion of Indian music history, theory, and instruments, including a manual for the *sitar*.

Dance and theatre: The main source book of Indian classical dance and theatre is the *Nātya-śāstra*, ascribed to BHARATA MUNI, trans. by MANMOHAN GHOSH, 2 vol. (1950–61), which deals with ceremonies, gesture language, architecture, production styles, makeup, and costumes. A.K. COOMARASWAMY, *The Dance of Shiva*, rev. ed. (1957), brings alive the philosophy and aesthetics of Hindu dance. For general understanding of classical dance forms and techniques, see RINA SINGHA and REGINALD MASSEY, *Indian Dances* (1967), which includes semiclassical styles, modern ballet, and biographical notes on dancers and gurus. FAUBION BOWERS, *The Dance in India* (1953), is a fascinating description of the four major classical dances. K. BHARATHA IYER, *Kathakali* (1955), is perhaps the best work on this dance-drama, with detailed descriptions of characters, historical background, and interpretation of dramatic symbols, with photographs and line drawings. KAPILA VATSYAYAN, *Classical Indian Dance in Literature and the Arts* (1968), surveys dance as found in temple sculpture, the plastic arts, and literature—a scholarly work with photos. For tribal dances and ceremonies of Central India, see VERRIER ELWIN, *The Muria and Their Ghotul* (1947; abridged ed., *The Kingdom of the Young*, 1968). For dramatic forms, production techniques, and classical rules the best source book is also the *Nātya-śāstra*, ascribed to BHARATA MUNI, *op. cit.* ARTHUR B. KEITH, *Sanskrit Drama in Its Origin, Development, Theory and Practice* (1924, reprinted 1959), remains a standard scholarly critical work for comparative analysis of Sanskrit plays. P. LAL, *Great Sanskrit Plays* (1964) are actable transcreations in crisp modern English. A delightfully-written survey is FAUBION BOWER, *Theatre in the East* (1956), which puts Indian dance and theatre in the larger perspective of South Asia with vivid and perceptive descriptions. For a general survey of classical, folk, and modern drama with sidelights on opera and ballet, illustrated by photographs and sketches, see BALWANT GARGI, *Theatre in India* (1962). HEMENDRA NATH DAS GUPTA, *The Indian*

Stage, 4 vol. (1934–44), deals mainly with the growth of Bengali theatres, actors, and productions in exhaustive detail with reproductions of period notebooks and papers. BALWANT GARGI, *Folk Theatre in India* (1966), is an eyewitness account of religious, secular, and masked dramas in villages, with over 100 photographs and line sketches. See also J.C. MATHUR, *Drama in Rural India* (1964). BERYL DE ZOETE, *Dance and Magic Drama in Ceylon* (1957), is a vivid account of rituals and magical masked dances in a diary form of day-to-day performances. E.R. SARACHCHANDRA, *The Folk Drama of Ceylon*, 2nd ed. (1966), is a scholarly work on the development and background of Sinhalese folk dramas and cults of exorcism. For puppets and masks, see two small pamphlets: J. TILAKASIR, "Puppetry in Ceylon" and SIRI GUNASINGHE, "Masks of Ceylon" (both published by the Department of Cultural Affairs, Ceylon, 1962). For the history of Kolam and description of various characters see O. PERTOLD, "The Ceremonial Dances of the Sinhalese," *Archiv Orientální*, vol. 2 (1930). For Devil Dances and their interpretation in the light of witchcraft rituals, see DANDRIS DE SILVA GUNARATNA, "Demonology and Witchcraft in Ceylon," *J. Ceylon Brch. R. Asiat. Soc.*, vol. 4, no. 13 (1861); and PAUL WIRZ, *Exorzismus und Heilkunde auf Ceylon* (1941).

Visual Arts (General works): VINCENT A. SMITH, *A History of Fine Art in India and Ceylon*, 3rd ed. rev. and enl. (1962); A.K. COOMARASWAMY, *History of Indian and Indonesian Art* (1927, reprinted 1965); and BENJAMIN ROWLAND, *Art and Architecture of India*, 3rd ed. rev. (1967), are general introductions with good bibliographies. A.K. COOMARASWAMY, *Figures of Speech or Figures of Thought* (1946), and *Tranformation of Nature in Art* (1934, reprinted 1956), contain important essays which discuss Indian aesthetic theories from the traditional point of view. A clear classification of Hindu images with particular reference to south India has been made in T.A. GOPINATHA RAO, *Elements of Hindu Iconography*, 2 vol. in 4 (1914–16, reprinted 1968); and J.N. BANERJEA, *The Development of Hindu Iconography*, 2nd ed. rev. and enl. (1956), is an analytical introduction to the subject. N.K. BHATTASALI, *Iconography of Buddhist and Brahmanical Sculptures in the Dacca Museum* (1929); BENOYTOSH BHATTACHARYYA, *The Indian Buddhist Iconography*, 2nd ed. (1958); and ALFRED FOUCHER, *Étude sur l'iconographie bouddhique de l'Inde*, 2 vol. (1900–05), are important studies of Buddhist iconography. A.K. COOMARASWAMY, *Yaksas*, 2 vol. (1928–31), is a masterly study of early iconography. HEINRICH ZIMMER, *Myths and Symbols in Indian Art and Civilization* (1946, reprinted 1963), discusses some important symbols of Indian art. Good collections of photographs are in JAMES BURGESS, *The Ancient Monuments, Temples and Sculptures of India*, 2 vol. (1897–1911); A.K. COOMARASWAMY, *Viśvakarma* (1912); STELLA KRAMRISCH, *The Art of India* (1954); and HEINRICH ZIMMER, *The Art of Indian Asia*, 2 vol. (1955).

Architecture: JAMES FERGUSSON, *History of Indian and Eastern Architecture*, rev. ed., 2 vol. (1910, reprinted 1967); PERCY BROWN, *Indian Architecture*, 2 vol., 5th ed. (1965); and S.K. SARASWATI, "Architecture," in R.C. MASUMDAR (ed.), *History and Culture of the Indian People*, vol. 3 and 5 (1954–57), are standard works that survey the entire history of Indian architecture. JAMES FERGUSSON and JAMES BURGESS, *The Cave Temples of India* (1880, reprinted 1969), is a comprehensive account of rock-cut architecture. STELLA KRAMRISCH, *The Hindu Temple*, 2 vol. (1946), is concerned with the principles and symbolism of Hindu architecture, and KRISHNA DEVA, *Temples of North India* (1969), presents a synoptic view of the various north Indian styles. G. JOUVEAU-DUBREUIL, *Archéologie du sud de l'Inde*, 2 vol. (1914), analyses the south Indian style. K.R. SRINVASAN, *Cave Temples of the Pallavas* (1964); and ARTHUR H. LONGHURST, *Pallava Architecture*, 3 vol. (1924–30), are important studies of early south Indian architecture. SIR JOHN MARSHALL, "The Monuments of Muslim India," in *The Cambridge History of India*, vol. 3, pp. 568–640 (1928, reprinted 1965); and PERCY BROWN, "Monuments of the Mughal Period," *ibid.*, vol. 4, pp. 523–576 (1937, reprinted 1963), are important essays on Islamic architecture in India.

Sculpture: S.K. SARASVATI, *A Survey of Indian Sculpture* (1957); and STELLA KRAMRISCH, *Indian Sculpture* (1933), discuss the broad historical and stylistic trends. LUDWIG BACHHOFER, *Early Indian Sculpture* (1929), is a fine stylistic analysis of sculptures from the third century B.C. to the third century A.D. The classic work on Gandhāra art is ALFRED FOUCHER, *L'Art gréco–bouddhique du Gandhâra*, 2 vol. (1905–41), though several of its conclusions are no longer tenable. SIR JOHN MARSHALL, *Taxila*, 3 vol. (1951), discusses works recovered from that site; and HAROLD INGHOLT, *Gandharan Art in Pakistan* (1957, reprinted 1971), provides excellent photographic documentation. R.D. BANERJI, *The Age of the Imperial Guptas* (1933), has a general discussion of the Gupta style; D.R. SAHNI, *Catalogue of the Museum of Archaeology at Sārnāth* (1914), contains information of interest on the school of Sārnāth. ELIKY ZANNAS, *Khajurāho* (1960); K.C. PANIGRAHI, *Archaeological Remains at Bhubaneswar* (1961); and R.D. BANERJI, *Eastern Indian School of Medieval Sculpture* (1933), are monographs on schools of north Indian medieval sculpture. A.H. LONGHURST, *op. cit.;* and K.A. NILAKANTA SASTRI, *The Cōlas*, 2nd ed. rev. (1955), contain information on the south Indian medieval styles. C. SIVARAMAMURTI, *South Indian Bronzes* (1963); P.R. SRINIVASAN, *Bronzes of South India* (1963); and DOUGLAS E. BARRETT, *Early Cōla Bronzes* (1965), are important studies of south Indian bronzes.

Painting: In considering the published literature it is important to remember that the study of Indian painting, confined to a limited number of scholars, is of comparatively recent growth, and is therefore full of controversies and uncertainties which keep shifting with the discovery of fresh materials. The standard work on Ajanta is GHULAM YAZDANI, *Ajanta*, 4 vol. (1930–55); and on the western Indian style, MOTI CHANDRA, *Jain Miniature Paintings from Western India* (1949). Much interesting information on the period of transition to the Rajasthani and Mughal styles has been brought together in KARL J. KHANDALAVALA and MOTI CHANDRA, *New Documents of Indian Painting* (1969). The Mughal school has been ably presented in PERCY BROWN, *Indian Painting under the Mughals, A.D. 1550 to A.D. 1750* (1924); and STUART C. WELCH, *The Art of Mughal India* (1964). DOUGLAS E. BARRETT, *Painting of the Deccan, XVI–XVII Century* (1958), is a brief introduction to the subject. The main publication on the Company style is MILDRED and WILLIAM G. ARCHER, *Indian Painting for the British, 1770–1880* (1955). The classic work on Pahari and Rajasthani painting is A.K. COOMARASWAMY's pioneering *Rajput Painting*, 2 vol. (1916). Fresh discoveries which have considerably changed the understanding of its history are summarized in KARL KHANDALAVALA, MOTI CHANDRA, and PRAMOD CHANDRA, *Miniature Painting: A Catalogue of the Exhibition of the Sri Motichand Khajanchi Collection* (1960). MOTI CHANDRA, *Mewar Painting in the Seventeenth Century* (1957); WILLIAM G. ARCHER, *Indian Painting in Bundi and Kotah* (1959); PRAMOD CHANDRA, *Bundi Painting* (1959); ERIC DICKINSON and KARL KHANDALAVALA, *Kishangarh Painting* (1959); WILLIAM G. ARCHER, *Central Indian Painting* (1958); and ANAND KRISHNA, *Malwa Painting* (1963), are informative summaries of the growing knowledge of the various schools of Rājasthān. The standard work, illustrated copiously, on the Pahari style is KARL KHANDALAVALA, *Pahārī Miniature Painting* (1958), and different in its account from WILLIAM G. ARCHER, *Indian Painting in the Punjab Hills* (1952). Books which cover most of the schools and are profusely illustrated include N.C. MEHTA, *Studies in Indian Painting* (1926); WILLIAM G. ARCHER, *Indian Miniatures* (1960); ROBERT SKELTON, *Indian Miniatures from the XVth to the XIXth Centuries* (1961); DOUGLAS E. BARRETT and BASIL GRAY, *Painting of India* (1963); and STUART C. WELCH and MILO C. BEACH, *Gods, Thrones and Peacocks: Northern Indian Painting from Two Traditions, Fifteenth to Nineteenth Centuries* (1965). *(Ceylon):* Standard works on Sinhalese art are SENERAT PARANAVITANA, *The Stūpa in Ceylon* (1946), *Art and Architecture of Ceylon: Polonnaruva Period* (1954), and *Ceylon: Paintings from Temple, Shrine and Rock* (1957).

Decorative Arts: Scholarly literature on the decorative arts in India is scanty and mainly in learned journals. The *Journal of Indian Art and Industries* (1886–1916) is the most important and contains numerous pioneering studies. SIR GEORGE WATT, *Indian Art at Delhi* (1903); and SIR GEORGE BIRDWOOD, *The Industrial Arts of India*, 2 vol. (1880), are for the most part descriptive texts emphasizing the technical aspects of the decorative arts as they had survived up to the closing years of the 19th century. JOHN IRWIN, "Textiles and the Minor Arts," in LEIGH ASHTON (ed.), *The Art of India and Pakistan*, pp. 201–237 (1950), is a brief historical survey of the subject. MOTI CHANDRA, "Ancient Indian Ivories," *Bulletin of the Prince of Wales Museum of Western India*, no. 6, pp. 4–63 (1957–59), is a monograph on the history of Indian ivory carving. THOMAS H. HENDLEY, *Asian Carpets* (1905), treats Indian examples in the important collections of the Maharaja of Jaipur. GEORGE P. BAKER, *Calico Painting and Printing in the East Indies in the XVIIth and XVIIIth Centuries* (1921), is the most important work on the subject; WILBRAHAM EGERTON, *An Illustrated Handbook of Indian Arms* (1880), catalogs and describes the wide range and achievement of the armourer's craft.

(C.S./J.A.B.v.B./E.C.D./
C.M.N./A.K.R./N.A.J./B.Ga./P.Ch.)

South Australia

One of the six federated states of the Commonwealth of Australia, South Australia lies on the south central coast of the nation. Western Australia lies to the west, the Northern Territory to the north, Queensland to the north and east, and New South Wales and Victoria to the east. The great majority of its people (whose total number, at 1,173,000, constituted almost 10 percent of the national total by the early 1970s) cling to the seacoasts, 70 percent of them living in Adelaide, the capital. The rainfall of two-thirds of the vast interior is insufficient to support significant human or animal populations. In the remainder there are scattered towns, mostly small, linked with the coasts by rail, road, or air. The aridity of the interior is only slightly compensated for by the presence, in the northeast, of a portion of the Great Artesian Basin, the world's largest region of natural springs. In two reserves in the northwest and southwest live remnants of the Aboriginal peoples of Australia. (For information on related topics, see AUSTRALIA; AUSTRALIA, HISTORY OF; AUSTRALIA, COMMONWEALTH OF; EYRE, LAKE.)

The physical environment. Most of South Australia is flat, with large tracts of sand and of the stony plains known as gibber desert. Over 80 percent of the land is less than 1,000 feet (300 metres) above sea level.

Geology and surface features. The state is divided into five major regions. The western shield, a vast barren region that also takes in much of Western Australia, is underlain primarily by Precambrian rock, over 570,000,000 years old. It swoops down from the north and west to engulf the entire Eyre Peninsula and much of the Yorke Peninsula on the coast. Bordering it in the northeast and reaching nearly to the coast, great layers of sandstone, lying 4,000 to 5,000 feet underground atop the Precambrian Shield, act as the water bed for the Great Artesian Basin. The several ranges of the Precambrian highland chain stretch mainly north from the Adelaide region and east into New South Wales.

The Eucla and Murray basins

In the extreme southwest and southeast, the Eucla and Murray basins reach into Western Australia and Victoria, respectively. Most of the land in these areas is of Mesozoic origin, between 65,000,000 and 225,000,000 years old. Near the town of Mt. Gambier, close to the border of Victoria, a few lake-filled craters have become tourist attractions, but the state is geologically inactive in terms of volcanic activity or surface glaciation.

The coastline is indented by two large, navigable gulfs separated by Yorke Peninsula, which is about 30 miles wide. Gulf St. Vincent, the more easterly of the two, is about 100 miles in length; Adelaide lies on its eastern shore. Spencer Gulf, twice as long, has near its northern limits the industrial towns of Port Pirie, Port Augusta, and Whyalla. The only large stream is the Murray River, which enters South Australia after forming the New South Wales–Victoria boundary. The rivers shown on maps as lying in the north rarely contain any water. To the south and east of Lake Eyre is a striking group of "mound springs," mineral deposits looking much like miniature volcanic cones and rising as much as 130 feet in the air. They have been formed by subterranean artesian action, which forces the water to the surface. The highest point in the state, Mt. Woodroffe, rises to 4,723 feet in the far northwest.

Climate. More than 80 percent of the state receives less than 10 inches (250 millimetres) of rain annually, all of it sometimes falling within a few months to be followed by months of almost complete drought. Only 0.3 percent gets 30 inches or more. The driest known area on the continent occurs around Lake Eyre. A few areas in the southeast receive more than 35 inches, but the average is between 20 and 30 inches. During the winter months of June, July, and August, most of the state has mean temperatures between 50° and 60° F (10° and 16° C), whereas summer figures stand at 65° F (18° C) in the south and between 80° and 85° F (27° and 29° C) in the north.

Land usage. Land suitable for farming other than extensive grazing is confined to the south, excluding the far west. For the analysis of rural statistics, the commonwealth statistician divides the state into nine divisions. Seven of these, representing about one-eighth of the state's area, contained in 1970–71 some 28,300 rural holdings with a combined area of about 36,000,000 acres. The remaining divisions had 750 holdings totalling 127,000,000 acres. About two-thirds of the state was thus unused for either agriculture or pasture. Some of this area is set aside for Aboriginal reserves.

Plant and animal life. No vegetation grows at all in the stretches of sand and gibber desert, and only scrub grasses and saltbush are hardy enough to survive in the dry regions of the north and west. Eucalyptus and native pines are scattered in the Flinders Ranges; the Mt. Lofty Ranges, near Adelaide, are more heavily wooded, mainly with eucalyptus. In the southeast, more than 170,000 acres of man-made forest, 90 percent of it in softwood, have been created since the late 19th century to assure forest preserves. These have softened considerably the landscape of the entire region.

Man-made forests

Most Australian species are represented, and some species are unique to the state. However, in comparison with the eastern states, South Australia has a sparse population of native fauna, partly because of the climate and soil but also because of the use of better land for farming. Some species have become extinct. Kangaroos and wallabies are common, as are wombats and bush rats. There are many kinds of birds, lizards, and snakes. The coastal sea has a wide variety of fish, but few fish are to be found in inland waters except in the Murray River.

Water resources. Maintaining and transporting supplies of fresh water for human, animal, and agricultural use is a constant concern in nearly all of Australia, the world's most arid continent. A few small streams rising in the Mt. Lofty Ranges near Adelaide have been dammed, but supplementary supplies must be pumped over these mountains from the Murray River. Another pipeline to the industrial town of Whyalla stretches for 223 miles from the Murray, whose water is also used for irrigation. The need to ensure a regular flow of high-quality water in the Murray led in 1970 to disputes between South Australia and its partners in the River Murray Agreement—the commonwealth, New South Wales, and Victoria—on the location of additional dams. In the interior, supplies are even scarcer, and private water catchment devices are still widely used.

The people. Slightly more than three-fourths of South Australia's population was born in Australia, and virtually all others were born in the British Isles or continental Europe. At the 1966 census 5,500 persons were reported as having 50 percent or more Aboriginal blood, and few of these had been assimilated into the processes of modern Western society prevalent throughout the state. In 1971 it was estimated that nearly 9,450 persons had some Aboriginal blood. The Anglican Church has the largest number of adherents, followed by the Methodist and Roman Catholic churches, with approximately equal numbers.

Since World War II the state's population has grown at a faster rate than that of Australia as a whole, mainly because of a disproportionately large influx of immigrants. After 1965, however, the rate of growth fell sharply and the immigration rate fell below the Australian average because of a deteriorating employment picture. From a postwar peak of 25.2 per 1,000 in 1947 the birth rate fell to 18.3 in 1967, but rose to 19.5 in 1970. The death rate, following a slowly declining trend, was 8.8 per 1,000 in 1970.

Population distribution

South Australia's population distribution is extremely skewed, plunging from Adelaide's more than 800,000 inhabitants to the approximately 32,000 in Whyalla, the next largest urban centre. The only other centres with 10,000 or more are Mt. Gambier, with nearly 18,000; Port Pirie, 16,000; and Port Augusta, 12,000. In 1971, 16 percent of the state's population lived in urban centres other than the capital, compared with 28 percent in all Australia. The 15 percent of the state's people classified as rural, the same as the average for all of Austrialia, has declined since World War II.

The economy. South Australia's rate of economic growth exceeded that of the nation as a whole during the 1950s and 1960s, and the labour market was strong, though a lower rate of growth in demand for motor vehicles and consumer durables after 1964 slowed the pace. Although the mean income is below the national average, so are living costs, putting the standard of living on a par with the rest of Australia. There were more automobiles and television sets per capita than in other states. In Adelaide the cost of land and building is below that of other Australian capitals.

Mineral resources. More than 10 percent of the state's work force is engaged in primary production, including mining and quarrying. The state is not rich in minerals, but the Middleback Ranges produce the bulk of Australia's iron ore. Some of this is processed at the Whyalla blast furnaces, but most is carried by water to Newcastle and Wollongong in New South Wales. The state is a major world source of opals. Most of the nation's salt and gypsum come from South Australia, but the open-pit coal mines in the Flinders Ranges are not extensive. A pipeline to Adelaide from the large natural gas fields discovered in the north was completed in 1969, and another to Sydney, New South Wales, was planned.

Manufacturing. Motor-vehicle manufacture is a large component of secondary industry, which employs about 30 percent of the state's labour force. General Motors-Holden Pty. Ltd., the nation's largest manufacturer, divides its operations between Melbourne and Adelaide, and Chrysler Australia Ltd. has its two main plants in Adelaide. Consumer durables, metal pipes and tubes, chemicals, and textiles are produced in or near Adelaide. A soda-ash plant at Osborne, near Port Adelaide, is the major producer in the Australian alkali industry. Whyalla, in addition to its iron and steel industry, has Australia's largest shipyards. Sawmills and factories producing particle boards and paper are located near the softwood plantations in the southeast. Port Augusta has a large power station using the low-grade coal transported by rail from the Flinders Ranges. Smelters at Port Pirie treat ores sent by rail from Broken Hill in New South Wales.

Agriculture. South Australia is a major producer of wheat and barley for export abroad, but it excels in its vineyards. Growing about a third of Australia's grapes, it produces over two-thirds of the wine and nine-tenths of the brandy for the tables of the commonwealth. Tuna and crayfish are pulled from the coastal waters for export to other states and abroad, and there are nearly 20,000,000 sheep producing wool, meat, and hides. Dairy production is largely for local consumption, but a considerable quantity of vegetables is sent to other states.

Role of the vineyards

Transportation. Transport services within the state centre upon Adelaide, and there has been little planning to coordinate the transportation systems as a whole with the state's economic development in terms of geography. The railways are almost entirely owned by the state and commonwealth governments, most of the commonwealth operations being related to the transcontinental line linking Sydney and Perth. A direct service operates between Adelaide and Melbourne, the Victorian capital, and a line of five-foot three-inch gauge from Adelaide meets the commonwealth railway at Port Pirie, the latter a standard-gauge line. The total length of operating railways is a little over 2,400 miles. Operating losses have caused the closure of many rural lines and a diversion of traffic to road transport. Over 9,000 miles of roads are hard-surfaced and another 19,000 are of crushed stone. Port Adelaide and Whyalla are major harbours for trade with other states and foreign nations; Port Stanvac is an oil-discharging port near Adelaide. Direct air services operate from Adelaide to Sydney, Melbourne, and Perth and to several towns within the state.

Administrative and social conditions. *Structure of the government.* The division of power between commonwealth and state parliaments results in commonwealth predominance in defense, foreign affairs, economic policy, welfare payments, and shipping and aviation. The main matters under state authority are education, hospitals, housing, police and prisons, and railways and roads. South Australia returns 12 members to the House of Representatives and 10 to the Senate in the Commonwealth Parliament. The two houses have total memberships of 125 and 60 respectively. The state parliament comprises a House of Assembly elected by universal and compulsory suffrage, and a Legislative Council elected by a restricted and voluntary franchise. Legislation must be passed by both houses.

The governor, the official representative of the crown, accepts the advice of ministers who, by constitutional convention, are collectively responsible to the House of Assembly. The principal minister, the premier, is usually the leader of the majority party in that house. Local councils are elected by property owners and tenants in 137 local-government areas.

The two main political parties are the Australian Labor Party (ALP) and the Liberal and Country League (LCL). The ALP usually receives a majority of the votes, but because the distribution of representatives favours rural areas the party is not always assured of a majority in the House.

Political parties

The governor appoints the eight members of the state's Supreme Court, from which appeals lie to the High Court of Australia and the Privy Council. Among the lower courts are those dealing with such specific matters as licenses for the sale of liquor. The Industrial Commission is responsible for labour arbitration, but most workers are covered by decisions of the Commonwealth Conciliation and Arbitration Commission.

Education. Although education is compulsory from six through 15 years of age, most children enter at five, and the rates of retention in the later years of secondary education are rising. All state schools are tuition free and administered by the Department of Education with a cabinet-level minister. The secondary system comprises high schools offering general academic training, vocationally oriented schools, and the "area schools" that serve rural populations in both primary and secondary grades. Private schools enroll 13 percent of primary pupils, 17 percent of secondary pupils. Apprenticeships are controlled by a state commission.

The institutions of higher learning in the state include the University of Adelaide, founded in 1876; the Flinders University of South Australia, which opened on the outskirts of the capital in 1966; and the South Australian Institute of Technology, dating from 1889. There are, in addition, an agricultural college, a school of art, and five teacher-training colleges.

Health and welfare. South Australia's very high health standards are reflected in the figures for infant mortality, which are among the lowest in the world (although the Aboriginal infant mortality rate is high). Over 250 public and private hospitals are in service throughout the state. There are few slums, and a large proportion of housing is built by the South Australian Housing Trust.

The Aborigines and the state. In 1967 the Australian constitution was amended so as to allow the commonwealth to legislate with respect to Aborigines. However, commonwealth activity so far has been limited to providing a small amount of money for projects to assist them. Most responsibility remains with the states. South Australia has a Department of Aboriginal Affairs headed by a minister. Assistance is given by way of housing, grants for education, loans for business purposes, and relief to those in need. Eight widely dispersed centres provide assistance in adapting living habits to Western ways within the security of the Aborigines' own communities. Discrimination against persons by reason of race or colour is prohibited.

Department of Aboriginal Affairs

Culture and recreation. *The arts.* South Australia's contributions to the arts are few, but it provides facilities for its citizens to participate in them. The Art Gallery of South Australia has an interesting collection of Australian art. A professional symphony orchestra is maintained by the Australian Broadcasting Commission and several theatres provide amateur and professional drama. Visiting artists present frequent concerts and recitals, and every second year a Festival of Arts is held, modelled on the Edinburgh International Festival of Music

and Drama in Scotland. In addition to the art school, there is a conservatorium of music at the University of Adelaide, and Flinders University offers courses in theatre, film, and television.

Sports. The main spectator sport is Australian-rules football, played during the winter from April to September. Cricket takes over during the summer. Year-round horse racing is popular, and in the summer the Adelaide showgrounds are given over to night trotting races.

Prospects. South Australia enjoys the advantages of a high material standard of living, a comfortable climate and, in the areas of settlement, a pleasant, if not spectacularly beautiful, landscape. Its major disadvantages are those of cultural isolation. The continuing growth of population may make some of the state's industries more productive and allow greater cultural diversity. It will, however, accentuate the urban sprawl already apparent in Adelaide, though to a much less degree than in Sydney and Melbourne. Moreover, it will create an increasing scarcity of water unless economically feasible methods of desalination are developed.

BIBLIOGRAPHY. The best general sources of information are the annual *South Australian Year Book* and the *Official Year Book of the Commonwealth of Australia*, both published by the Commonwealth Bureau of Census and Statistics. More detailed statistical information is contained in the *Statistical Register of South Australia* (South Australian Parliamentary Paper, annual). For geographical aspects, see R.J. BEST (ed.), *Introducing South Australia*, pt. 2 (1958); and M. WILLIAMS (ed.), *South Australia from the Air* (1969). A popular survey of the economy is *Business Guide to South Australia*, published by the Premier's Department (1969).

For a comprehensive study of education see *Education in South Australia* (1971), the report of an official enquiry published by the Government of South Australia.

(K.J.H.)

South Carolina

One of the original 13 states of the United States, South Carolina lies on the southern Atlantic Seaboard of the nation. Its 31,055-square-mile (80,432-square-kilometre) area, roughly triangular in shape, is bounded by North Carolina on the north and northeast, the Atlantic Ocean along its southeastern leg, and by Georgia along the southwestern leg, where the long boundary is formed entirely by the Savannah River. Columbia is its largest city and the capital.

Location and general character

South Carolina generally is looked upon as a state of the Deep South, a designation suggesting the traditions that have marked it since its early years in the 17th century. The coastal enclave from around Charleston to Savannah, Georgia, produced an aristocratic culture that was and remains unique, even in the Deep South. The long-dominant agricultural economy and the institution of slavery upon which this culture was based was shattered by the Civil War and its aftermath of intersectional discord. For a century and more, South Carolina, like its regional neighbouring states to the west, experienced a political and social turmoil that was intensified beginning in the 1950s by the nationwide civil-rights movement, which touched upon many of the state's institutionalized ways of life.

By the early 1970s, South Carolina had closed much of the gap separating it from the mainstreams of American life. The majority, however, of its almost 2,600,000 citizens, nearly one-third of whom were black, remained fiercely proud of the solid accomplishments of the state's past and sought to conserve the finer elements of that tradition. The atmosphere and architecture of Charleston continued to make it one of the more notable cities in the nation, symptomatic of the special heritage and history that pervades many parts of the state. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; CIVIL WAR, U.S.; and NORTH AMERICA.)

THE HISTORY OF SOUTH CAROLINA

The colony. The popularly elected assembly of South Carolina was in almost constant political tugs-of-war with, first, the proprietors and, second, after these men had been thrown out, the British crown, the latter turmoil leading directly into the American Revolution. Permanent European settlement of the region—which comprised present-day North as well as South Carolina, Georgia, and parts of Florida and extended the width of the continent—had been accomplished in 1670 by a grant of Charles II to several of his favourites after a 1629 grant by Charles I was never followed through. In 1680 the settlers moved their town to the site of Charleston. By 1693 the assembly gained the right to initiate legislation, but reactionary rule by the proprietors in the early 1700s and the imposed establishment of the Church of England were among the factors leading to the overthrow of proprietory rule in 1719.

Under crown rule from 1729, the Carolinas prospered as an outpost of empire in which the European wars of the 18th century could be fought and from which a healthy trade in pelts, rice, and indigo could be expected. Charleston became the commercial centre of the southern coast, and its citizens evolved a social structure that stressed a refinement and cultural attainment rare even in such older cities as Boston, New York, and Philadelphia.

The demand of the assembly for the right to control finances was perhaps primary among the factors leading Carolinians into revolutionary league with their fellow colonials; in September 1775 the royal governor fled. After a British fleet captured Charleston in 1780, the state was the scene of more engagements than any other, though most were small.

Statehood, Civil War, and aftermath. Well before the Revolution, a fundamental rivalry had developed between the older seaboard settlements and those inland, the latter producing a breed of tough frontier people whose numbers increased dramatically after statehood. By 1808 the Up Country had four-fifths of the population, but the Low Country had four-fifths of the wealth. Among several political compromises that helped to maintain internal harmony was the establishment by the General Assembly in 1786 of Columbia, in the interior, as the capital.

The unity of white South Carolinians was achieved largely by the antislavery agitation and other external threats. In 1832 a federal tariff harmful to Southern commerce led the state to attempt nullification of the law within its boundaries; the chief spokesman for the South was the fiery South Carolina senator John C. Calhoun; and from 1847 to 1852 its secession from the Union was blocked only by the non-support of the other Southern states. The state's act of secession in December 1860 and the engagement with federal troops at Ft. Sumter in Charleston harbour in April 1861 led the nation into the Civil War.

Nullification and secession

During the war, South Carolina lost nearly one-fourth of its 63,000-man army contingent and suffered the burning of Columbia in 1865. But the humiliation of the Reconstruction years, filled with rule by corruption and bribery, left a greater scar on the people of the state. By retaliatory fraud and intimidation of blacks, white South Carolinians regained state rule in 1876. Blacks were effectively disfranchised by constitutional revision in 1895.

Thereafter, internal politics involved mainly the struggle between the poorer masses and the propertied "Bourbon" classes for control of the totally dominant Democratic Party. Beginning in 1948, the state's voters virtually abandoned Democratic presidential candidates, and, as in most states of the Deep South, the state and national party increasingly were split philosophically. Governor, later Senator, Strom Thurmond became a nationally recognized spokesman for the Southern conservatism ranked against civil-rights movements.

In spite of South Carolina's forceful reaction against this movement and occassional outbreaks of violence, by the early 1970s it had led the way in compliance with the letter of the federal laws against racial discrimination in several fields.

THE NATURAL AND HUMAN LANDSCAPE

Geographical divisions and patterns of settlement. Included within South Carolina are parts of two large natural regions, the Coastal Plain and the Appalachian

Highlands, with elevations ranging from sea level to about 3,560 feet.

The Coastal Plain. Encompassing roughly two-thirds of the state's area, the Coastal Plain extends from the Atlantic Ocean inland 120–150 miles to the Piedmont. Comprised of loosely consolidated to unconsolidated marine sediments, the Coastal Plain may be further divided into the outer Coastal Plain, less than 100 feet above sea level, and the inner Coastal Plain, which rises gradually toward the Piedmont, reaching elevations of nearly 300 feet. The outer Coastal Plain has the longest history of human settlement in the state and is known locally as the Low Country. From the border of North Carolina to Garden City Beach, the coast is a smooth, unbroken crescent beach. Midway along this stretch of coast is Myrtle Beach, centre of a favourite tourist attraction. Southwest of Garden City Beach, the coast is broken by numerous islands, which continue into Georgia. These are the Sea Islands, where the excellent long-staple cotton was grown until ravaged by the boll weevil. The islands are climatically the most tropical part of South Carolina; some, such as Hilton Head, rarely experience freezing temperatures. There is much tidal and freshwater marsh. Huge pines, gums, live oaks, cypresses, and magnolias are draped with Spanish moss. Alligators and every poisonous snake known in the United States inhabit these swamps. Deer, wild hogs, and wild turkey are among the game animals hunted on the Coastal Plain.

The subtropical Sea Islands

Separating the Coastal Plain and the Piedmont is a strip of land as much as ten miles wide that geographers call the fall zone. The area is easily recognized by the rapids in the streams, which have a gradient sufficiently strong to permit hydroelectric production.

The state has no large natural lakes within its borders, although on the Coastal Plain there are hundreds of small, shallow, egg-shaped depressions, many of which are filled with water. Geomorphologists have discovered no satisfactory explanation for these curious natural features, known as Carolina Bays.

The Piedmont. While the Piedmont is frequently referred to as a plateau, the description is in fact unfortunate, for it is really an erosional plain, or peneplain. Numerous monadnocks (protuberances of resistant rock) are evidence of a once much higher surface. Comprising about one-third of the total area of the state, the Piedmont is known locally as the Up Country and, economically and politically, has always differed greatly from the Low Country. Everywhere on the Piedmont, rivers have cut valleys, leaving a surface rolling to steeply sloping. Most of the highways, cities, towns, railroads, and farm fields are on the flat, broad tops of the hills and ridges of the Piedmont.

The Blue Ridge. Encompassing only a small part of the state is a rugged mountainous portion of the Blue Ridge Mountains. Elevations here are 1,000 to 2,000 feet higher than those of the highest portions of the Piedmont. Sassafras Mountain, the highest point in the state (3,560 feet, or 1,085 metres), is on a crest in this region.

Climate. Generally, South Carolina enjoys a mild climate. Summers are long, hot, and humid; winters are short, cool, and drier. Average January temperatures vary from 45° to 51° F (7° to 11° C). Mean July temperatures vary from 79° to 82° F (26° to 28° C). Most of the state receives about 48 inches (1,200 millimetres) of rain per year, but the mountains may receive as much as 70 inches. Average snowfall amounts vary from a trace along the coast to seven inches in the mountains. The growing season varies from 200 days in the mountains, to about 260 days in the southern, lower portions, and 300 days or more on some of the Sea Islands.

THE PEOPLE OF SOUTH CAROLINA

National origins. English and Irish settlers were the first people to arrive directly from Europe. A group of Dutch from New York moved to South Carolina in 1671, and 45 French Protestants from England arrived there in 1680. After 1685 large numbers landed directly from France following the revocation of the Edict of Nantes and the persecution of the Protestant Huguenots. Negotiations were begun in 1682 to bring in Scots, and during the 18th century many Germans arrived. Place-names in the state reflect the national origins of those early arrivals. By 1970, however, South Carolina had the lowest proportion of foreign-born residents of any state of the nation. Of a total population of over 2,500,000, only 11,140 (0.5 percent) were foreign-born. The vast majority of South Carolinians are of English-speaking ancestry, immigrations from central and southern Europe having largely bypassed the state.

The ethnic mix

Five Negroes were among the 148 people who arrived in 1670 to establish the first permanent settlement, and by 1708 the Negro population exceeded the Caucasian population. Slave population increased principally from importation, and the natural increase was very low. The first decennial census of 1790 showed 43.7 percent of the total population to be Negro; by 1880 the figure was 60.7 percent. By 1970, however, the percentage of blacks had declined to 30.

Demography. The 1970 census revealed that South Carolina's population continued to be predominantly rural, unlike the nation as a whole, although the rural–urban mixture was changing. Since the end of World War II, urbanization has proceeded rapidly. In 1950, about 37 percent of the population was classed as urban; in 1970, the figure was nearly 48. The four largest urban areas—Charleston, Columbia, Greenville, and Spartanburg—had the greatest growth. There is a continuing stream of emigration to other states—mainly North Carolina, New York, and Pennsylvania.

Rural and urban settlement

Distribution of population is very uneven. Large areas of the outer Coastal Plain are only sparsely populated, and a substantial portion of the state is in fact unpopulated. Three main belts of dense population may be noted: along the coast, with centres of population at Beaufort, Charleston, Georgetown, and the Conway–Myrtle Beach district; along the fall zone, with concentrations at North Augusta, Aiken, Columbia, West Columbia, Cayce, Camden, and Cheraw-Bennettsville; and an area centred on Greenwood, Anderson, Greenville, and Spartanburg, in the upper portion of the Piedmont.

THE STATE'S ECONOMY

In 1970 the state's annual average work force was more than 1,000,000, of which 94 percent were in nonagricultural employment. South Carolina ranked 48th among the states and the District of Columbia with a per capita income some $1,000 below the national average. While food prices were essentially the same in South Carolina as anywhere in the United States, the cost of living generally was lower.

Agriculture and forestry. Contrary to popular opinion, South Carolina is not primarily an agricultural state. The most recent census of agriculture showed less than half the total land area of the state to be in farms and less than half the farmland to be under crops. From 1930 to 1970, the number of farms declined by two-thirds, land in farms decreased nearly one-third and average farm size increased from 70 to 161 acres, all figures well above the national averages of change.

Cash receipts from farm marketings slightly exceeded $400,000,000, exclusive of government payments; 59 percent was from crops and 41 percent from livestock. Tobacco, soybeans, cotton, and peaches were the principal crops, while poultry and eggs, cattle and calves, and dairy products were the principal sources of income from livestock.

Farm products

A 1970 forest survey classified 62 percent (nearly 12,500,000 acres) of the total land area in South Carolina as woodland, an increase of over 550,000 acres since 1958. Dense hardwood forests formed a principal source of supply for the state's furniture industry. Though the outlook for industries dependent upon large high-quality timber was bleak, for those able to shift their use to smaller, lower quality timber the outlook was much brighter, since the supply and growth of small softwoods and hardwoods was increasing rapidly.

Manufacturing. The 1967 Census of Manufactures showed that 17 of 20 standard industrial classification

manufacturing groups were found in South Carolina. In terms of value added by manufacturing, these ranged from $1,236,600,000 for textile-mill products to $16,100,000 for primary metal manufactures. Eight industrial groups added values in excess of $100,000,000.

From a modest beginning in 1820, the textile industry has grown to be the overwhelming giant of South Carolina's manufacturing industry. Textile manufacturing in 1971 was carried on in some 425 establishments, which employed over 40 percent of all of the state's manufacturing employees.

Chemicals and allied products industries ranked second in value added but third in number of employees. The apparel and related industry group ranked third in value added and second in number of employees, while paper and allied product manufacturing ranked fourth in value added and eighth in number of employees.

Stretching along the banks of the Savannah River lies one of the United States' largest atomic-energy plants: the 200,000-acre Savannah River Plant of the Atomic Energy Commission. The plant's main purpose is to manufacture plutonium for use in nuclear weapons, but it also produces heavy water and irradiates cobalt for use in treating cancer.

Transportation. Among South Carolina's prime assets is the vast network of rail lines and highways that blanket the state. The combined facilities of 17 railroad companies make readily accessible supplies of raw materials and the means of reaching markets.

Although ranking 40th among the states in area, South Carolina ranks 32nd in highway mileage, with 59,726 miles of highway, 39,466 miles of which are surfaced. South Carolina also has portions of five routes of a national system of interstate and defense highways, totalling over 750 miles. More than 400 trucklines serve the state.

The air-transportation network comprises five commercial airlines serving eight cities and 31 general and 40 limited airports.

Port facilities

More than 20 steamship lines with regular service operate out of Charleston. Georgetown is the only other South Carolina port with facilities for handling oceangoing freighters. Along the entire coast of South Carolina is the Intracoastal Waterway, providing storm-free passage for pleasure boats moving between Miami and New York City, as well as for commercial vessels.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *The state level.* Governmental authority in South Carolina derives from the constitution of 1895.

The highest executive authority of the state is the governor, who is elected for a four-year term and cannot succeed himself. The lieutenant governor is president of the Senate; in the event the governor's office becomes vacant, the lieutenant governor takes that position. The secretary of state, comptroller general, attorney general, treasurer, adjutant and inspector general, superintendent of education, and commissioner of agriculture are also elected for terms of four years. The General Assembly has the power to increase the duties of all these executive officers and to establish new departments when necessary for carrying on the government's business. Some departments already thus established include health, agriculture, education, highway, and labour.

Legislative power is vested in a General Assembly, composed of the Senate and the House of Representatives. The Senate is composed of 46 members apportioned among 20 districts according to population; members serve four-year terms. The 124 members of the House are elected for two years.

The power of the judicial department is vested in a Supreme Court consisting of a chief justice and four associate justices elected by the General Assembly for a term of ten years, the terms of one justice expiring every two years so that there are always experienced men serving. There is also a Court of Common Pleas, which has civil jurisdiction, and a Court of General Sessions, which deals with criminal cases only. For these courts there is a judge, elected by the General Assembly, and a solicitor, elected by the people, in each of the 16 judicial circuits into which the state is divided.

Local government. The unit of local government is the county, which is in turn divided into townships that exist largely for tax-assessment purposes. Counties are administered by commissions. County commissions are appointed by the governor upon recommendation of the county legislative delegation. With the exception of the auditor and treasurer, who are appointed by the governor with Senate confirmation, county officers are elected.

Since there is no legislative body on the county level, the state legislature is oppressed with a mass of local legislation. In practice, the county delegation to the state legislature governs the county and forms in effect a county legislature.

Party activity. A two-party system is gradually evolving in the state: in 1970, for the first time, Republicans held a primary election to choose candidates for national and state offices. Efforts on the part of concerned South Carolina citizens, as well as efforts from outside the state, have resulted in a large increase in registered black voters. For the first time since the 1930s, three black members were elected to the State House of Representatives. Though the state remains basically conservative, continued immigration is diluting the degree of conservatism. In the early 1970s, the state's representation in Congress included a Republican in both House and Senate.

The social milieu. Some indication of the conservatism of South Carolinians as regards state and local government involvement in social and economic affairs may be gauged from the fact that in the fiscal year 1970–71, for example, only $327 per capita was spent by state and local government on such functions as education, public welfare, hospitals, and sanitation, compared with a national median of $416.

Problems in education and literacy

The average age of South Carolina's population is relatively low, and the fact that it has a higher percentage of its population in public schools than any other state has contributed to major problems in public education. By every statistical measure of educational attainment, South Carolina ranks near the bottom of all the states. In the fiscal year 1971, over 30 cents of every dollar spent by the state government was for education. With one of the lowest literacy rates of any state, and with an alarming school dropout rate, South Carolina, at the beginning of the 1970s, faced a monumental task in raising the educational level of its citizens to equal the national average.

The state supports eight institutions of higher education, the chief of which is the University of South Carolina in Columbia, chartered as South Carolina College in 1801 and opened for instruction in 1805. Enrollment in the early 1970s was over 17,000, including some 3,500 students enrolled at five two-year regional campuses. Clemson University, a land-grant college incorporated in 1896, has an enrollment of over 8,000 students at its main campus in Clemson and regional centres. Winthrop College at Rock Hill (1886), with an enrollment of some 4,000, was long known as the South Carolina College for Women and is becoming coeducational. One of the few state-supported military colleges, The Citadel, is located in Charleston, as is The Medical University of South Carolina. South Carolina State College, for years a Negro college, chartered in 1896, enrolled 3,179 students and is now integrated. The College of Charleston, the oldest publicly supported institution in the United States, chartered in 1770, officially became a part of the state-supported system of colleges and universities in 1971. Numerous private institutions supported by the major Protestant denominations dot the state.

Health and welfare

Although the cities and larger towns have excellent medical and health facilities, some of the purely rural counties are woefully lacking in such facilities. Most of the state's health problems exist in the areas of rural poverty. Malnutrition and certain parasitic infestations, such as hookworm, continue to be problems in some areas; respiratory ailments are on the decline, as are other contagious diseases, as a result of educational programs and free inoculations at county medical clinics.

Cultural life and institutions. South Carolina antebellum society was said to have been based on the plantation, slavery, and romanticism. Interest in education—most early families sent their sons to England or the Continent for their university education if they could afford to do so—and the arts and the love of entertainment and gracious living began early and persist to the present. Almost from the earliest days, families who could afford to do so hired tutors to educate their children or sent them to private schools. Musical societies, such as the St. Cecilia Society in Charleston, still existing in the 1970s, were founded. These ultimately became the organizations that sponsored elaborate debutante balls for introducing the daughters of member families to society.

It is thought that the first play produced in America was performed in Charleston. The Dock Street Theatre was built in 1736. It was rebuilt in 1927 and is used regularly by the Dock Street Players. Columbia, Greenville, and other cities have theatre groups of professional calibre performing current plays. Columbia's Town Theatre claims the distinction of being the oldest continuously performing theatre group in the United States.

The prolific novelist William Gilmore Simms was South Carolina's greatest antebellum literary figure and the focus of a literary circle that founded *The Southern Review* in 1828 and otherwise maintained Charleston's status as the cultural centre of the South. Other notable literary figures include Dubose Heyward, who wrote the novel *Porgy* (published in 1927) upon which George Gershwin based his musical *Porgy and Bess*, although the latter was not seen in Charleston until 1970, when a local cast performed it as a part of South Carolina's tricentennial celebration. Jasper Johns, who had paintings selected for hanging in the art pavilions at the Seattle World's Fair (1962), Expo 67 in Montreal (1967), and in Ōsaka, Japan, for Expo 70 (1970), is probably South Carolina's most famous painter. The Columbia Philharmonic Orchestra is developing into an orchestra of major stature. Columbia, Charleston, Greenville, and Spartanburg have local musical associations that sponsor annual musical programs performed by internationally known orchestras, opera stars, ballet companies, and solo instrumentalists.

Early American architecture

Charleston is noted for its splendid, well-preserved examples of late 18th- and early 19th-century architecture in both domestic and public buildings. Beaufort, Georgetown, and Columbia also provide examples of Early American architectural designs. Local historical societies have been successful in obtaining and restoring old houses of historic architectural interest and in preserving public buildings of equal interest for future generations.

PROSPECTS

In 1970 South Carolina celebrated the tricentennial of the establishment of the first permanent English settlement. Its people are proud, steeped in its history and aggressively planning for the future. Industrialization continues at a rapid pace, and each year thousands of new manufacturing jobs are created. South Carolinians are aware of problems associated with manufacturing and have been successful in blocking new industry that does not guarantee avoidance of atmospheric and water pollution. New manufacturing plants hasten the migration of people from rural to urban or urbanized areas. One of the most exciting developments in South Carolina has been the relatively peaceful integration of the races. Public institutions are open to all people whatever their race or religious beliefs. In 1970 the public-school system was totally unified for the first time. Problems associated with total integration have been minimal, and prospects for the future have never been more promising.

BIBLIOGRAPHY. DAVID DUNCAN WALLACE, *The History of South Carolina*, 4 vol. (1934); and *South Carolina: A Short History, 1520–1948* (1951), are the standard exhaustive accounts of South Carolina's development. NELL S. GRAYDON, *Tales of Columbia* (1964), is a collection of materials of the opulent atmosphere of the antebellum to the present. DORIS E. HARLESS, *South Carolina: An Economic Profile* (1964), provides a summary view of South Carolina's economy in 1963 but also includes past data. GEORGE C. ROGERS, JR., *Charleston in the Age of the Pinckneys* (1969), is a history of Charleston during its greatest years. GEORGE C. ROGERS, JR., *Evolution of a Federalist: William Loughton Smith of Charleston (1758–1812)* (1962), shows the relationship between the changing economic fortunes and political interests of the South Carolinians. BEN ROBERTSON, *Red Hills and Cotton, An Up Country Memory* (1942, reprinted 1960), consists of reminiscences by a well-known journalist of Up Country life in the first half of the 20th century. JULIAN J. PETTY, *20th Century Changes in South Carolina Population* (1962), is an exhaustive treatment of population changes and their significance. J. DOUGLAS CONKLIN, *Selected South Carolina Economic Data* (1964), is an analysis of South Carolina's economic growth and development.

(D.O.B.)

South Dakota

South Dakota is one of the plains states of the United States, its westward-rising topography providing a transition from the prairies of the Middle West to the Rocky Mountains farther to the west. Bounded on the north by North Dakota, on the east by Minnesota and Iowa, on the south by Nebraska, and on the west by Wyoming and Montana, it comprises 77,047 square miles (199,551 square kilometres), which are split into the two nearly equal segments of "east-river region" and "west-river region" by the broad Upper Missouri River Basin. Pierre, at the centre of the state, is among the smallest of state capitals, having only 10,800 of the state's more than 665,000 inhabitants. The state's name is derived from the Dakota or Sioux Indians who roamed the plains before the arrival of the white man.

In the late 20th century, South Dakota was among the few states still predominantly rural in population. This condition reflects, in part, the 19th-century settlement of the land by hardy immigrants from northern and central Europe who bypassed the more developed states to the east to pioneer the new frontier. It recalls also the rapid influx of prospectors in the 1870s and 1880s following the discovery of gold in the Black Hills, in the southwestern part of the Dakota Territory. The town of Deadwood became a symbol of the flamboyant era of the bonanza seekers and a part of Western lore. Although South Dakota remains the nation's major gold producer—its Homestake Mine is the largest in the Western Hemisphere—the economy of the modern state is more closely tied to its agriculture and industry and to the millions of tourists who visit the exotic, vividly eroded Badlands and other points of interest in the state. (For information on related topics, see the articles UNITED STATES; UNITED STATES, HISTORY OF THE; NORTH AMERICA; and GREAT PLAINS.)

Frontier heritage

THE HISTORY OF SOUTH DAKOTA

Settlement and gold rush. A lead plate, discovered at Ft. Pierre in 1913, records the presence of French explorers during 1742–43. Twenty years later Spain acquired sovereignty over the region, and its rule led to United States ownership in 1803. Trappers and fur traders were virtually the only Europeans in the area until 1856, when Ft. Randall was built on the Missouri. The permanent settlements of Vermillion and Yankton were established in 1859. The Dakota Territory was created in 1861, but for several years settlement was confined to the southeast between the Big Sioux and Missouri rivers.

The discovery of gold in the Black Hills in 1874 attracted settlers to the western part of the territory, which had been closed to white entry under an earlier treaty with the Teton Sioux tribes. In the year following the Battle of the Little Bighorn in Montana, in 1876, the Indians were forced to accept a reduction of their reservation and to cede the Black Hills. Despite the major attention that Deadwood commanded, Rapid City emerged as the main gateway city to the region. Freight and stage lines connected the mining population with the east until railroads were constructed across the Indian country in 1905.

Statehood and homesteading. The gold rush was followed by a flood of settlers into the east-river region,

swelling its population from about 80,000 to 325,000 between 1878 and 1887. This rapid expansion led to calls for division of the territory at the 46th parallel and separate statehood for the southern half. In the north, and in Congress, a single state was favoured. The southern section held constitutional conventions in 1883 and 1885; at the latter the state of Dakota was established. Dual statehood based on a division below the 46th parallel received congressional approval in 1889, and South Dakota was admitted as the 40th state in the Union.

The Black Hills agreement with the Sioux opened up not only the mining region but also the contiguous grazing lands, and by a new agreement in 1889 additional Indian lands became available in the west-river region. Cattlemen remained in control until the early 1900s, when homesteaders broke up the range and received for settlement about 4,000,000 more acres of the Indians' lands. Severe droughts during 1910 and 1911 brought the land rush to a halt, but in the next decade the semi-arid west-river region grew from about 44,000 to 138,000 inhabitants. Further droughts, onslaughts of grasshoppers, and dust storms from increasingly eroded soils seriously depressed the farm economy of the entire state in the 1930s.

Political traditions

Conservatism and progressivism. South Dakota's political history is similar to that of its neighbouring states. During the 1890s the appeal of the Populist Movement led it temporarily and briefly away from the Republican Party. A four-year fusion administration, however, produced only the nation's first initiative and referendum law. At the turn of the century the state was swept along by the tide of progressivism. An insurgent, progressive wing captured the Republican Party in 1906 and enacted legislation calling for direct primary elections and railroad regulation. In 1917-19 the state embraced a still more radical program, establishing a rural credits plan, a system of state hail insurance, a state coal mine, and a cement plant. Of these state-operated enterprises, only the cement plant, at Rapid City, remained in operation early in the 1970s, and since the 1920s the voters have shown a more conservative bent.

THE NATURAL AND HUMAN LANDSCAPE

The physical environment. Although it is divided into three distinct regions, South Dakota, save at isolated places, does not exhibit marked variations in its physical environment.

Geological character. Eastern South Dakota and some river valleys in the west near the Missouri lie within the glaciated physiographic province known as the Prairie Plains. The western half, except for the Black Hills in the southwestern corner, has the rolling topography of the unglaciated Great Plains, characterized by high buttes, rough canyons, and wide expanses of nearly level tablelands. It includes the Badlands, which extend along the White and Cheyenne rivers for a distance of more than 100 miles. This eroded landscape has been a rich source of fossilized prehistoric animals.

The Black Hills—two-thirds of which lie in South Dakota, the remainder in Wyoming—are a dome-shaped uplift rising 3,500 feet (1,100 metres) above the surrounding plain. Harney Peak, 7,242 feet above sea level, is the highest point in North America east of the Rocky Mountains.

Rivers and drainage. The Missouri River drains the entire state except for the Minnesota Valley lowland, which is drained by Big Stone Lake and Lake Traverse in the northeastern corner. In the east-river area the principal tributaries of the Missouri, the Big Sioux, Vermillion, and James, flow southward into valleys formed by glaciation. The main rivers of the west-river region, the Grand, Moreau, Cheyenne, Bad, and White, flow eastward.

Climate. The climate is characterized by extremes in temperature, persistent winds of moderate velocity, low precipitation, and relatively low humidity. The skies are generally clear. Cyclonic storms occur frequently in the east-river section during the summer. The mean annual temperature for the entire state is about 45° F (7° C). A weather station at east central Huron reported average lows of 2° F (−17° C) and highs of 25° F (−4° C) in January, average lows of 61° F (16° C) and highs of 90° F (32° C) in July. Extreme temperatures recorded were −34° F (−37° C) and 110° F (43° C). The average number of frost-free days ranges from 160 in the southeast to 110 in the Black Hills.

The average annual precipitation for the state is 19.12 inches (485.6 millimetres), but it ranges from about 24 inches along the eastern border to 14 inches or less in the northwestern corner. In the centre of the state, the transitional zone from the Prairie Plains to the Great Plains, precipitation drops from the 20-inch to the 18-inch level. About three-fourths of the rain falls during the growing season, and snowfall ranges from about 22 to 60 inches. The Black Hills region receives more moisture, especially in winter, than the surrounding plains.

Soils. The soil types vary. The chernozem, or black, soils, formed mostly from glacial drift and well adapted for wheat and corn crops, cover the east-river area. Within the Great Plains province are the chestnut and Pierre, or gumbo, soils, distinguished by their heavy, sticky texture. They are well adapted to grassland regions.

Virgin prairie grasslands

Vegetation. At the time of settlement, the east-river region was a prairie grassland with a thick matte of tall grasses growing to a height of three feet or more. These tall grasses have a deep root system, adapted to subhumid conditions. The shortgrass species, comprising chiefly the grama, buffalo, and western wheatgrasses, are endemic to the Great Plains. They are drought resistant and have a shallow root system and the capacity to mature quickly. Aside from its forest, the ecology of the Black Hills is much like that of the Great Plains.

The wooded areas lie mostly in the Black Hills and along the buttes in the northwestern part of the state. Most wooded acreage is incorporated into the Black Hills National Forest and the Custer National Forest. The western yellow, or ponderosa, pine is the chief commercial tree.

Animal life. Custer State Park in the Black Hills maintains a herd of about 2,000 bison. Other Black Hills species include antelope, deer, elk, beavers, bobcats, and porcupines. Coyotes, jackrabbits, and prairie dogs are plentiful in other sections. The state has nearly 300 species of birds. Bald and golden eagles are found in diminishing numbers along the Missouri Valley and in the Badlands. The Missouri is an important route for the north–south migration of waterfowl, mostly ducks and Canadian geese. For many years South Dakota has been a hunter's paradise because of its plentiful supply of ring-necked pheasants, a game bird introduced into the state early in the 20th century.

West-river and east-river areas

Patterns of human settlement. Early in South Dakota's history, the west-river region was set apart from the farming settlements of the east-river area by the large Indian population, the open-range cattle industry, and the physiographic characteristics of the Great Plains, together with the Black Hills mining economy. The distinction between the two sections remains real and forms the basis for differing views in the legislative halls.

Except for the Black Hills region, the west-river area has only about two persons per square mile, a sharp contrast with the northeastern and southeastern regions, in which the density is generally four times greater. Of the rural inhabitants, only about a fifth reside in the west-river region. Almost 10 percent of the state's area consists of tribal and individually owned Indian land. Most of it is located in the west-river section in the remaining portions of the old Great Sioux Reservation created in 1868.

The east-river section is divided into two areas approximately north and south of a line running eastward from Pierre. General farming practices prevail in the northern part, with the main reliance on wheat and other small-grain crops. The southern part represents a more diversified farming economy, with feed grains and livestock production its specialties.

THE SOUTH DAKOTANS

European ethnic composition. The earliest settlers generally came from the Upper Mississippi Valley, many

of them with a New England background. Immigrants arriving directly from the northern and central countries of Europe also played an important part in the colonizing process. In 1890 a third of the white residents were foreign-born. Although a half century later only one out of every ten was foreign-born, nearly half the population was only two or three generations removed from Europe. The Americanization process generally had run its course by the end of World War I, but Old World customs and speech still prevail in many communities.

Immigrant pioneers

Those of Scandinavian descent—Danes, Norwegians, and Swedes—make up about a third of the foreign stock. The Norwegians, located mostly in the east-river region, have exerted an especially strong influence. Compact Danish and Swedish communities are found in the southeastern part of the state.

Persons of German descent make up about a sixth of the population. They include German-Russians, who are heavily concentrated near Yankton in the southeast and in three north central counties. The Bohemians, or Czechs, nearly 2 percent of the population, generally live in the south central counties. Other significant ethnic groups are the Dutch, the Finns, and the Welsh.

The Indians. The descendants of the original Sioux inhabitants form a major segment of the population, numbering slightly more than 33,000 in 1970. The old tribal system has largely disappeared, but about four-fifths still live in reservation areas scratching out a living by farming and stock raising. The Teton Sioux live mainly in the west central part of the state between the Black Hills and the Missouri; the Yankton Sioux, on the Yankton Reservation; and the Sissetons, on the Sisseton Reservation in the northeastern corner of the state.

Low living standards in reservation areas have led to a substantial migration of younger Indians to nearby communities. The resulting racial tension amid charges of social and economic discrimination led, in 1963, to a civil rights law that prohibited exclusion of any person from a public place because of race, colour, religion, or national origin.

Religious organizations. Immigrant groups established about half of the state's churches. Their religious establishments not only promoted social solidarity but also played an important part in the acculturation process. In 1890 there were 12 religious sects in the state, the Catholics being most numerous and the Lutherans a close second. In 1970 these two were still the most numerous, while other leading organizations, in order of rank, were the Methodist, United Church of Christ, Presbyterian, Episcopal, and Baptist churches.

The German-Russian groups include a few thousand Mennonites and members of the Hutterian Brethren. Although the Hutterites, in some of their religious beliefs, are related closely to the Mennonites, they practice a form of religious communism, living in colonies and holding all goods and property in common, in accordance with a strict interpretation of early Christian teachings. There were 16 such religious communities, most of them in the James River Valley, in 1955, when a state law prohibited the formation of additional colonies or the expansion of existing ones.

Mennonite and Hutterite sects

Some parts of the Indian community still adhere to ancient Indian faiths. The majority, however, profess Christianity, most of them identifying themselves as Roman Catholics and Episcopalians.

Demography. The state reached a population peak of 692,849 in 1930, but drought and depression had caused a drop by 1940. This declining trend continued despite improved conditions, and the 1970 census figures indicated a further loss of 2.2 percent during the 1960s. All but 14 counties sustained losses. There were 17 births per thousand and 10 deaths per thousand in 1970.

In addition to the outward flow of people, there was a pronounced shift from the rural communities to the city, though the state remained more than 55 percent rural in 1970. Sioux Falls and Rapid City represent metropolitan areas with populations, respectively, of more than 70,000 and 43,000. The population of Rapid City more than trebled between 1940 and 1970.

THE STATE'S ECONOMY

South Dakota had a per capita income of $3,446 in 1971, ranking 36th in the nation. This was below the average of the nation and of all neighbouring states except North Dakota. Total personal income was $2,108,000,000 in 1970, of which 19 percent was derived from agriculture, 13 percent from all levels of government, 13 percent from wholesale and retail trade, 9.5 percent from services, and 8 percent from manufacturing. The mining industry accounted for less than 2 percent of the income.

Agriculture. Cash income from livestock and livestock products is generally about four times that from crops. The state ranks high in the production of rye and flaxseed as well as spring wheat. In 1970 it held ninth place in corn and seventh in alfalfa-seed production. In many years it produces half the nation's supply of Kentucky bluegrass seed. Approximately 44,500 farming units were in operation in 1972, averaging a little over 1,000 acres in size. Normally, 150,000 acres are under irrigation, mostly in the west-river region.

Manufacturing. The major industrial groups are food products, lumber and wood products, and stone and glass products. Meat processing is the largest single industry, representing about 35 percent of the value added by manufacturing in the state.

Mineral production. In 1970, South Dakota ranked 41st among the states in mineral production. Leading products are sand and gravel, gold, stone, and cement. Nonmetallic commodities account for more than half the total mineral value of $62,000,000. The entire yearly output of gold, valued at about $17,000,000, is produced by the Homestake Mine. Large lignite beds containing low-grade uranium deposits await exploitation.

Hydroelectric power. The multipurpose Big Bend, Fort Randall, Gavins Point, and Oahe dams, with a combined power output of 1,483,035 kilowatts, have made South Dakota a major producer of hydroelectric power. Seven small hydroelectric plants, all but one located in the Black Hills, are privately owned. Only a small portion of the total production of electric energy is generated by coal or gas. More than 96 percent of South Dakota's farms are electrified, most of them served by rural electric cooperatives.

Electrification

The labour movement. Organized labour has not found a congenial climate in South Dakota, and there have been few major disturbances. A "right to work" amendment was adopted in 1946, making it illegal to require union membership as a condition of employment. It was followed by legislation placing restrictions on boycotts and picketing. There are about 200 unions with a total membership of about 24,000, mostly affiliated with the national AFL-CIO. About half of the union membership lives in Sioux Falls.

Transportation. South Dakota has a network of roads totalling more than 84,000 miles. The state trunk system of nearly 9,000 miles carries the bulk of the highway traffic. The Missouri River, not bridged until 1924, can be crossed at ten different points, including the crests of the four dams. Passenger bus and motor freight lines have kept pace with highway improvements.

As motorized transportation grew, rail services declined. About 3,800 miles of railroads remain in operation, but passenger trains have been discontinued and many small-town freight depots abandoned. With the aid of federal funds, air transportation has increased steadily since World War II. At the start of the 1970s there were 82 approved airports and four commercial airlines operating from nine certified terminals.

ADMINISTRATION AND SOCIAL CONDITIONS

Governmental structure. Adopted in 1889, the constitution has been amended more than 80 times to reflect changing conditions and revised thinking about the state's basic law.

Executive, legislative, and judicial branches. Besides the governor and lieutenant governor, six administrative officials are chosen by the electorate, all for two-year terms. The superintendent of public instruction is chosen by nonpolitical ballot. The legislature comprises a Senate

of 35 members and a House of Representatives of 70. The membership is reapportioned every ten years, following the federal census. In 1951 a Legislative Research Council was created to provide continuity between the annual sessions. The judicial system comprises a Supreme Court, consisting of five judges elected for six-year terms; a Circuit Court, consisting of 21 judges operating within ten judicial districts; and a local system of 19 county-court districts and a municipal court in each of the three largest cities. All judges are chosen by nonpartisan ballot. There are also provisions for justices of the peace in rural areas and police magistrates in the smaller cities and towns.

Law enforcement

The machinery for law enforcement includes the state's attorney and sheriff at the county level and the office of attorney general at the state capitol. Special enforcement agencies include a state highway patrol and a force of game wardens. There are correctional and penal institutions at Plankinton and Sioux Falls.

Local government. The state maintains more than 3,500 units of government below the state level. These include 64 organized and three unorganized counties, 306 towns and cities, and 1,050 organized township governments. There are also more than 100 special districts, most of them concerned with soil conservation and drainage.

Public finances. South Dakota has no bonded indebtedness. The major sources of income for the state government are a sales tax, revenue from licenses and other fees, and federal aid, the latter in recent years accounting for more than a third of the income. The general property tax is reserved for the local governments. Most of the state's expenditures are for highways, education, and welfare.

Politics. South Dakota holds frequent elections and sets short terms for its officials. A primary election occurs in June during even-numbered years, followed by general elections in November. Special elections often occur to decide local issues. The Republicans, the dominant party since early settlement, have elected all but five of the state's 25 governors, and they have been the minority party in the legislature only three times. In presidential elections the Republican candidate failed to carry the state only in 1896, 1932, 1936, and 1964.

Education. The public school system is under the control of local and county boards but subject to policies formulated by a Department of Public Instruction. Since 1943, direct legislative appropriations for the support of the common schools have supplemented the general property tax and revenues from the common school lands. School district reorganization remained on a voluntary basis until 1968. Public school enrollment in 1971–72 was more than 163,000, and private schools accommodated nearly 15,000 others.

The state supports seven institutions of higher learning: the University of South Dakota in Vermillion (founded 1882); South Dakota State University in Brookings (1883); South Dakota School of Mines and Technology in Rapid City (1885); and four state teachers' colleges. The oldest of the several private colleges and junior colleges is Augustana College at Sioux Falls (1860).

Health and welfare services. The Department of Health is responsible for programs dealing with communicable diseases and sanitation as well as with the inspection and licensing of hospitals and nursing homes and the collection of vital statistics. Sixty-six licensed hospitals were in operation during 1970. The elderly were served by more than 120 nursing homes and 20 extended-care facilities.

The welfare needs of the state are the responsibility of the Department of Public Welfare. More than two-thirds of its funds were received from federal grants in 1969–70. Greater welfare and delinquency problems prevail among the Indian population because of cultural differences and the unfavourable economic conditions. Institutions maintained by the state include a school for the blind at Aberdeen, a school for the deaf at Sioux Falls, a home for the feeble-minded at Redfield, a hospital for the insane at Yankton, a tuberculosis sanitarium at Sanator, and a soldiers' home at Hot Springs.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. The early settlers were not oblivious to cultural refinements. They attended musical and dramatic performances by traveling artists in local opera houses and heard national lectures booked by lyceum bureaus. Schoolhouses were centres of community social life, and local reading circles and library associations grew into modern libraries. South Dakota today has 86 municipal libraries. Eighteen libraries are maintained by state and private institutions, including the library of the South Dakota State Historical Society in Pierre. A traveling library system, operated under the Free Library Commission, was established in 1913.

The arts. The dramatic arts manifest their appeal at community and summer theatres and departments of dramatic art on college and university campuses. Most notable among the summer theatres is the Black Hills Playhouse, operated in Custer State Park since 1946. More spectacular is the Black Hills Passion Play, presented during the summer at Spearfish since 1939. The musical arts are represented by seven symphony orchestras.

In the graphic arts two native-born artists, Harvey Dunn and Oscar Howe, gained wide recognition. Harvey Dunn, reared on a pioneer Dakota homestead and one of the nation's leading illustrators, became well-known for his paintings of pioneer life. Oscar Howe, a Yanktonai Sioux, has made use of the motifs and symbolism of his Indian heritage. Another outstanding artist to identify himself with the state was Gutzon Borglum, whose stone carvings of four American presidents on Mt. Rushmore in the Black Hills became a major tourist attraction.

Native artists

In literature many native writers have dealt with the South Dakota scene. Among them are the writer of frontier stories Hamlin Garland, who launched his career in the state, and Norwegian-born Ole Edvart Rölvaag, who spent his early immigrant life near Sioux Falls, the locale portrayed in his *Giants in the Earth* (1927). The *South Dakota Review,* a literary quarterly published by the University of South Dakota since 1963, affords an outlet for regional as well as local talent.

Communications. South Dakota at the start of the 1970s supported 13 daily newspapers, the largest of which were the *Aberdeen American News, Rapid City Journal,* and the *Sioux Falls Argus-Leader,* and 146 weeklies. Radio broadcasting was represented by 37 licensed stations, six of which were educational. Ten commercial and four educational television stations were in operation in 1971.

Prospects. Notwithstanding a full century of growth, South Dakota still bears many aspects of a frontier state. There are, however, encouraging signs of change. Through soil conservation and proper patterns of management, agriculture has attained a greater degree of stability. The multipurpose Missouri Basin program holds great promise for the future, and flood control and hydroelectric development have been successful. The recreational development of the reservoirs and their shores lures an increasing number of sportsmen and vacationers to the region. Planning is under way for the irrigation of 500,000 acres in the James River Valley with water diverted from the Oahe Dam above Pierre. There are plans for the construction of a data bank at Sioux Falls, which will contain computerized documentation obtained from space-satellite observations as part of the Earth Resources Observation System.

The population changes, attributable both to mechanical improvements in farming and lack of employment opportunities, pose serious problems for the state. They particularly point up a pressing need for a more diversified economy in order to maintain existing social institutions and to provide adequate governmental services. The campaign to attract industry is making progress, and small plants have been drawn to the smaller cities in increasing numbers. Oil has been produced in the state in small quantities since 1954, but prospects are good for a major oil industry in the northwestern section of the state.

BIBLIOGRAPHY. E.P. ROTHROCK, *A Geology of South Dakota* (1943), a description of the land surface; LYLE M. BEN-

DER, *The Rural Economy of South Dakota* (1956), an analysis and survey by a farm management specialist; SOUTH DAKOTA STATE GEOLOGICAL SURVEY, *Mineral and Water Resources of South Dakota* (1964), descriptive and historical compilation. HERBERT S. SCHELL, *History of South Dakota*, 2nd rev. ed. (1968), a standard account based on recent research; DOANE ROBINSON, *History of South Dakota*, 2 vol. (1904), a general account; GEORGE W. KINGSBURY, *History of Dakota Territory*, 2 vol. (1915), a detailed work; HOWARD R. LAMAR, *Dakota Territory, 1861–1889: A Study of Frontier Politics* (1956), a well-researched study concerned with an interpretation of political behaviour; HAROLD E. BRIGGS, *Frontiers of the Northwest: A History of the Upper Missouri Valley* (1950), various aspects of early South Dakota history; J. LEONARD JANNEWEIN and JANE BOORMAN (eds.), *Dakota Panorama* (1961), a publication of the Dakota Territory Centennial Commission, covering various topics, with a bibliography of more than 1,000 titles dealing with the state; FEDERAL WRITERS' PROJECT, *South Dakota: A Guide to the State*, rev. ed. (1952), a good descriptive account for the general reader.

(H.S.Sc.)

Southeast American Indians

The Southeastern American culture area is bounded on the east and south by the Atlantic Ocean and the Gulf of Mexico (though some scholars would place the southern portion of aboriginal Florida within the orbit of the circum-Caribbean culture area). To the west, the Southeastern area merges with the southern Plains and the extreme easternmost part of the Southwestern culture area. To the north, the Southeast blends into the northeastern woodlands area with no discernible break in cultural tradition (see NORTH AMERICAN PLAINS INDIANS and EASTERN WOODLANDS INDIANS).

GENERAL CHARACTERISTICS OF THE SOUTHEAST CULTURE AREA

Physiographically, the Southeast is characterized first by a coastal lowland belt broadly encompassing the subtropical zone of southern Florida; the scrub forest, sandy soil, and savannah grassland of the Atlantic and Gulf coastal plains; and the alluvial floodplains of the Mississippi drainage. Second, there is the piedmont of the midland interior, where the landscape changes to rolling hills, crisscrossed by several major river systems and covered predominantly with oak-hickory forest. Third, there is the southern Appalachian Mountains area of eastern Tennessee and the western Carolinas, a land of high peaks and deeply etched valleys, containing hardwood forests and, at high elevations, flora and fauna typical of more northerly regions.

Populations of the Southeast

The Southeast was one of the more densely populated areas of native North America, having an aboriginal population conservatively estimated at 120,000. The bulk of this population resided inland, where advantage could be taken of extensive game resources, wild plant foods, and an abundance of arable land. Only the non-horticultural peoples of south Florida appear to have satisfactorily adjusted to a basically maritime way of life.

Population was distributed among a large number of separate groups—independent villages, autonomous village clusters, and "tribelets." Most of those tribelets disappeared soon after white contact and left only faint traces in recorded history. They perished through the lethal combination of newly introduced diseases, removal into slavery, and direct warfare with white invaders or intertribal conflicts generated by white pressure. The survivors, if any, were assimilated into such larger, more powerful tribes as the Choctaw and Cherokee and various member tribes of the Creek Confederacy. These latter tribes persist to the present as distinctive peoples possessing a rich history and viable cultural heritage. Other intermediate-sized groups, such as the Houma, Catawba, and Chickasaw, survive as marginal enclaves but have lost much of their historic Indian identity. Such groups as the Seminoles (a branch of the Creek that migrated to Florida in the 18th century) and the Lumbee (a large group of Indians in Robeson County, North Carolina, whose precise Indian ancestry is unknown) appear to be entering an active phase of retribalization in which their Indian identification is being reasserted.

Languages of the Southeast

Muskogean-speaking peoples constituted the major linguistic family in the aboriginal Southeast. The Muskogean family included the following five main subdivisions: (1) Choctaw and Chickasaw, two different dialects of a single language, found in Mississippi and western Tennessee; (2) Apalachee, a long extinct language of northwestern Florida; (3) Alabama and Koasati, two closely related languages spoken in the central Southeast; (4) Hitchiti and Mikasuki, two related dialects, formerly spoken in Georgia; and (5) Creek and Seminole, also closely related dialects, spoken in eastern Georgia and later in Florida.

Four Lower Mississippi Valley languages, namely Natchez, Tunica, Chitimacha, and Atakapa, are thought to have distant affinity to Muskogean, but they show sufficient divergence both from the main Muskogean languages and from each other to warrant semi-independent status as linguistic isolates. Timucua, the major language of aboriginal north Florida, was once thought to be related to Muskogean, but its present status is problematic. One linguist believes it may be related to a language spoken in Venezuela, while others feel it may bear ultimate relationship to Siouan.

There are four definite representatives of the Siouan family in the Southeast: Tutelo, Biloxi, Ofo (Mosopelea), and Catawba. These tribes were widely scattered and, with the exception of Biloxi and Ofo, show little relationship to one another and probably represent different prehistoric penetrations of Siouan speakers into the Southeast. Yuchi, the language of a major tribal group once residing in eastern Tennessee and later in Georgia, also demonstrates distant affinities to Siouan but is sufficiently distinctive to be classified as an isolate. Many small piedmont tribes were probably Siouan speaking, but surviving data are insufficient to make definite identifications.

The Cherokee represent the sole member of the Iroquoian family in the presently demarcated Southeast, though the Iroquoian-speaking Tuscarora, Nottaway, and Meherrin, residing on the margin of the area, are included in the Southeast in some culture area maps. The Caddoan speakers on the western boundary of the Southeast belong to a distinctive language family that shows distant relationships to the Siouan and Iroquoian families. The affiliations of many of the smaller coastal and piedmont tribes are unknown. Mention should also be made of Mobilian, an important trade language containing many Choctaw components, which served as a lingua franca in the Mississippi Valley.

TRADITIONAL CULTURE PATTERNS

It is difficult to describe the Southeast in terms of a total cultural pattern or dominant ethos. The environment provided possibilities for such different ecological adaptations as the dependence on maritime resources and the wild zamia root by the Calusa of south Florida and the seasonal buffalo hunts of the horticultural Caddoan tribes. Besides the internal diversity, the external relations of the Southeast are also complex. The lack of geographic barriers to the north and west allowed significant cultural interchange between the Southeast and adjacent areas. There is evidence of overseas' cultural connections with the Antilles; however, the dominant thrust of this diffusion seems to have been from the mainland to the islands. Such individual culture traits as the cane blowgun, double-weave basketry, fibre-tempered pottery, and miscellaneous musical, ritual, and mythological details suggest at least limited contact with northern South America. More definite is prehistoric cultural contact with Meso-America. Not only did such basic cultivated plants as maize, beans, and squash ultimately derive from Meso-America, but numerous symbolic motifs in Southeast religious art have close analogues in ancient Mexico.

The picture of the Southeast that emerges at the time of first European contact is one of intensive cultural change. One senses a period of cultural levelling marked by considerable population movement, warfare, and the formation of confederacies, all of which was accompanied by



Distribution of Southeastern American Indian cultures.
From H. Driver *et al.*, *Indiana University Publications in Anthropology and Linguistics* (1953)

large-scale technological and ideological diffusion. A distinctive Southeastern cultural tradition or style was in the process of being forged, but this fusion was far from complete on the eve of European contact.

The village or town

Social organization. *Settlement patterns.* The basic political unit for most Southeastern tribes was the local village or town. This basic unit varied in size and configuration depending on differences in local ecological potential and cultural preference. Some towns attained populations of over 1,000 individuals, but the more typical Southeastern village numbered less than 500 people. Settlement patterns conformed to two basic types: (1) dispersed hamlets, with households strung out for several miles, usually following valley bottoms or streams, and (2) tightly nucleated settlements often surrounded with protective timber palisades. The heart of the local town was a ceremonial centre consisting of a central council house or temple, which in the interior region might be semisubterranean or located on a pre-existing mound; a central plaza or square, which, among the Muskogean-speakers, was usually surrounded by three or four benches or arbours oriented in the cardinal directions; a ball pole or scalp post sometimes culminating in a carved animal emblem; the residences of the chief and other important local dignitaries; and sometimes storage structures for communal produce.

Considerable variation in house types existed. In much of the area, the Indians constructed circular, conical-roofed winter "hot houses," sealed tight, except for an entryway and smoke hole. Summer dwellings tended to be rectangular, gabled structures made from a framework of upright poles, lashed together and covered with lath, grass, cane matting, or bark and plastered with clay. To the south, housing tended to be more flimsy, with raised floors, palmetto thatched roofs, and, often, open sides. To the west, the Caddoans lived in domed "grass houses." A homestead might contain auxiliary storage buildings and a special cookhouse.

Each town or village was fairly autonomous. Super-ordinate control at the tribal level tended to be weakly developed, although pressure for tribal consolidation and even the formation of intertribal confederacies was greatly increased with the coming of Europeans. A village might be linked to other villages in the same area by ties of kinship, language, and shared cultural traditions; nevertheless, each village claimed sovereignty over its local area and was governed by its own chiefs and war leaders.

Chiefs and other leaders

Stratification. Over most of the Southeast, chieftainship tended to be hereditary within certain lineages. The degree of chiefly power and authority varied, however, from the almost divine kingship of the Great Sun among the theocratic Natchez to the self-effacing status of the peace-making, consensus-seeking *micos* and *ukus* among the more egalitarian Choctaws, Creeks, and Cherokees. War leaders normally achieved their positions on the basis of past accomplishment. War chiefs tended to be active and assertive personalities and younger, by about a generation, than the peace chiefs.

The alternation between peace and war or the occurrence of such competitive activities between alien groups as ball games, communal hunts, and trading expeditions helped to imbue much of Southeastern social structure with a characteristic dualism. The peace chief held sway in the domestic village, whereas the war chief was ascendant in areas external to the village, except when the village was under threat of imminent attack. Young men in the village alternately adjusted to roles appropriate to war and peace, often symbolically represented as red and white activities, and these transformations were usually effected through extensive ritual. This dualistic emphasis was also frequently expressed in the organization of clans, subtribes, and villages into complementary social divisions.

Social stratification was highly developed in some parts of the Southeast, while its significance in other subareas was minimal. Although much has been written about the supposed caste systems among the tribes of the Lower Mississippi, the Chitimacha appear to have been the only society to have possessed true castes in the sense of ranked groups that practiced strict endogamy (marriage within the group).

Social ranking among the Natchez

The elaborate rank system of the Natchez consisted of four groups: three upper classes, composed hierarchically of the suns, the nobles, and the honoured people, and a lower class of commoners (or stinkards, as they are referred to in the early French sources). Upper class indi-

viduals were required to marry into the lower class of commoners, and many commoners also married other commoners. Offspring of males in the upper classes would assume a rank one step below that of their fathers (thus, the child of a sun father and commoner mother would become a member of the noble class). The progeny of upper class females, however, would retain the rank of their mothers rather than descend to a lower station. The system, as described, would be unstable, since the supply of available commoner women would soon be depleted after several generations. Many explanations have been advanced to explain this so-called "Natchez paradox," but the difficulties probably reside in the inaccuracies or incompleteness of the original French sources.

Social stratification also was highly elaborated among the aboriginal inhabitants of Florida. Among the Timucua, for instance, the "king" enjoyed an elevated status considerably above that of his followers and was sometimes carried about in a litter. In many other Southeastern societies there was a trend toward the ranking of towns or clans. Member towns of the Creek Confederacy were sometimes ranked in terms of their tribal affiliations or on the basis of outcomes of inter-town ballgames. The Caddo were said to have ranked their clans on the basis of the reputed strength of the totemic animal ancestor in a kind of natural "pecking order." In other tribes, such as the Cherokee, stratification was only weakly developed, though certain clans might possess special ceremonial prerogatives and recruitment to certain offices might be determined on the basis of clan.

Kinship and marriage. In the Southeast, descent was almost universally reckoned in the female line, though not all societies possessed matrilineal clans. Clans were lacking among the Choctaw, and their existence among the eastern piedmont tribes has been questioned. Clans, where they existed, were apparently not restricted to nor localized within specific villages. The resulting dispersal of clan members throughout a tribe or nation thus served as a kind of social adhesive binding together the larger body politic. Certain ceremonial knowledge and privileges might also be passed down along clan lines, and clans were also important as mechanisms of social control, since vengeance for serious crimes was frequently a clan responsibility.

Marital customs

Marriage was often marked by a symbolic ceremonial exchange whereby the groom presented his bride with game and the bride reciprocated with vegetable food. Because residence after marriage normally found the man moving into the wife's natal household, the husband was expected to contribute to the economic maintenance of his wife's family, as a form of bride service. After a few years the couple might leave to form their own household. Most Southeastern tribes permitted and some even encouraged pre-marital sexuality. After marriage, however, adultery—especially on the part of the wife—was often severely punished. Nevertheless, divorce seems to have been a frequent and, often, almost casual event. Polygamy was permitted in most groups, although a man usually had to gain the assent of his first wife before taking on a second spouse. The levirate, a custom by which a widow was remarried to her deceased husband's brother, was fairly common, particularly as male mortality increased during the wars of the European period.

Socialization. During a woman's late pregnancy, both she and the father were often subjected to various dietary taboos and restrictions on their activities. Children were nursed until they tired of the breast or until the mother again became pregnant. Responsibility for the child's early education was vested in the mother. Later, a young boy received instruction in male skills from his father and his mother's brother; in many systems the mother's brother, as the senior male in the matrilineage, assumed considerable importance as a disciplinarian, tutor, and sponsor for his sister's son. Young girls remained at their mother's side and were trained in various duties associated with the domestic household. Behaviour considered proper was reinforced with praise and encouragement, as when a boy killed his first deer or a young girl manufactured her first basket; behaviour considered improper might be greeted with mild rebuke, ridicule, or shame. Children were rarely subjected to physical punishment. In those few instances in which punishment was deemed necessary, it was generally meted out by someone other than the parents. A popular method of chastisement throughout the Southeast was raking the skin with briars or a special pointed scratching instrument, but, even here, such action was regarded as strengthening or toughening the youth, rather than as delivering direct retribution for misdeeds. Young boys enjoyed considerable permissiveness. They spent much time with their peers in wrestling, playing games imitative of adult activities, and stalking rabbits, squirrels, and birds with blowguns and scaled-down bows and arrows. The freedom and wide behavioral space permitted the young boy contrasted markedly with the restricted sphere, close surveillance, and early responsibility training that characterized the daily routine of his sister.

Rites of passage

Puberty rituals were either absent or only weakly developed in the Southeast. Girls were secluded during their first menstruation, but this event occasioned no public celebration. (Menstrual blood, though, was regarded as a potent and polluting substance, and women either absented themselves from the household during their periods or were subjected to restrictive taboos.) Similarly, no special rituals attended the transition from boyhood to manhood. A boy might receive instructions from tribal elders in esoteric lore or in preparation for special ritual offices, but graduation from such training seldom was marked by a formal commencement. A young man's first participation in a war party and the achievement of military honours were, however, given public recognition. Probably the clearest marker of the passage from adolescence to adulthood was marriage.

Economic systems. As already implied, the primary division of labour was by sex. Women were responsible for cultivating the fields, gathering wild-plant food, cooking and preserving food, rearing the young children, and manufacturing such basic domestic items as cordage, baskets, pottery, and clothing. Men assumed the primary roles of warriors and hunters, occupations that often took them away from the village for extended periods of time. Men also cleared the fields by girdling trees, assisted in the harvest, constructed houses and public buildings, and manufactured ceremonial objects and implements for personal use.

Except for the marginal groups on the western Gulf Coast, villages were semipermanent and located near rich alluvial soil or, in the Lower Mississippi region, near natural levees. Such land was easily tilled, possessed adequate drainage, and enjoyed enduring productivity. Fertility was enhanced by the custom of annually burning off the brush. The length of the growing season in most regions of the Southeast allowed multiple crops. Planting was done in spring, and some produce was available by mid-summer. The major harvest time, however, was late summer and early fall, a time of plenty when most of the major ceremonies were celebrated. Many villages became deserted, except for older people, during the winter months, when families took to the woods in search of game. Men also departed for a shorter hunt in late spring and early summer, after the crops had been planted.

Cultivation of maize

Principal foods and goods. The economic mainstay of the Southeast was maize, the cultivation of which was well established in most areas by the time of first European contact in the mid-16th century. Several varieties of maize were grown, including "little corn" (related to popcorn); flint, or hominy, corn; and flour, or dent, corn. Early corn was baked as roasting ears; later corn was pounded into hominy or cornmeal in wooden mortars made of large, upright, partly hollowed logs. Associated in the maize complex were varieties of beans and squash. Fields were prepared with mattocks and hoes and planted by punching holes in the ground with dibbles (digging sticks), inserting seed, and covering the holes with earth to form a small mound. Cornfields belonged to individual households, but among some tribes communal fields were also cultivated, with the produce going to the chiefs for

support of the civil–religious hierarchy or for redistribution to the needy.

The importance of maize cultivation to the way of life of the southern Indians cannot be overemphasized. Not only did maize provide a high yield of nutritious food with a minimal expenditure of labour, but maize, beans, and squash could be easily dried and stored for later consumption. This reliable food base enabled men to spend much time away from their villages on hunting, trading, and war expeditions. It was not fortuitous that the standard war ration was parched corn.

Other important plants and animals

Secondary cultivated plants included the sunflower (processed for its oil), chenopodium or orache (spinach-like greens), and tobacco. Many additional plants, such as species of wild grapes, plums, and perhaps walnut and pecan trees, can be regarded as being in a condition of incipient domestication, since there is evidence to suggest that Indians exerted some effect on selection. Important wild-plant foods were various types of berries, nuts and acorns, wild potatoes and amaranths, smilax, zamia root, and maple sugar or honey locust sap. The economic botany of the Southeastern Indians can be expanded to encompass the vast array of plants utilized for cordage, clothing, dyes, fish poisons, medicines, building materials, and various tools and utensils. Perhaps mention should be made of the distinctive Southeastern use of two species of holly (*Ilex cassine* and *Ilex vomitoria*) as ingredients in a special decoction, the "black drink," to induce sweating and vomiting in ceremonial and medical contexts.

The only native domesticated animal in the Southeast was the dog, which was used to a minor extent in hunting and was probably more important as a sentinel to warn of the approach of strangers. In accounts of the Hernando De Soto expedition (1539–1543), there are several references to small, fat, barkless dogs that were served to the Spanish visitors by their Indian hosts. Spanish trail hogs, brought by De Soto to feed his troops, became wild and were ancestral to the modern mongrel razorback hog. Horses were introduced later, mostly through the intermediacy of tribes to the west.

The aboriginal Southeast also teemed with wild game: deer, black bears, bison, elks, beavers, squirrels, rabbit, otters, and raccoon—some of which were used for their hides, bone, or fat as well as for food. In Florida, turtles and alligators played an important part in subsistence. Among birds, wild turkeys were the principal quarry, but partridges, quails, and seasonal flights of pigeons, ducks, and geese also contributed to the larder. The feathers of eagles, hawks, swans, and cranes were highly valued for ornamentation, and in some tribes a special status was reserved for an eagle hunter.

Both on the coast and on inland rivers, streams, and lakes, a wide variety of fish were taken in weirs, fish traps, and dip nets and dragnets and by hooks and lines, bows and arrows, and spears. In the interior, poison was administered in ponds and sluggish or dammed streams to gather a rich harvest of stunned fish. Coastal groups gathered oysters, clams, mussels, cockles, and crabs. Interior groups found freshwater mussels and crawfish.

Economic organization and ecology. In the well-endowed Southeastern area, each household group was fairly self-sufficient. The economic specializations and trade networks that did develop tended to centre on subsidiary and luxury items. Salt deposits were unequally distributed and formed one basis for trade. There also was regular trade between the coast and interior, with shells, which were used for beads, pendants, and horns, exchanged for soapstone, flint, furs, and other inland resources. The presence in the Southeast of artifacts made of imported copper and certain types of red clay suggests important trade connections with the western Great Lakes tribes.

Indians are popularly viewed as living in a primeval or virginal territory. Such was not the case in the Southeast. Indians maintained a delicate balance with their environment, and their presence was a vital link in a complex ecological chain. By tillage, controlled use of fire, and hunting, they altered the landscape significantly. Large areas of secondary re-growth favoured certain types of berry bushes and groundnuts. The presence of this secondary-growth flora was, in turn, essential for supporting large populations of browsing deer, squirrels, rabbit, and wild turkeys on which man depended for a large measure of his sustenance. In this process and in combination with hunting, the decline of other animals, such as the wood bison, was probably accelerated. In areas where intensive maize cultivation had already taken hold, such as in the Lower Mississippi, game animals had become scarce in historical time. In the central Southeast, however, the very diversity of plant and animal resources and the highly generalized adaptation of the Indian in exploiting these resources seems to have resulted in maintenance of an equilibrated balance between man and nature.

Belief systems. *World views.* The delicate man–nature relationship is well expressed in what is known of Southeastern religion and world view. The world was perceived as animated by a proliferation of ghosts, witches, and spiritual essences of animals, plants, and natural objects or phenomena. As can be inferred from the frequent elaboration of funerary practices, most groups professed belief in an afterlife. The location of the resting place for deceased souls was either in the direction of the western setting sun or up above in a celestial firmament. It was generally thought that the souls of the recently deceased would hover around the community and try to induce close friends and relatives to join them in their journey to eternity. The elaborate funerary rites and the extensive taboos associated with death were as much a protection for the living as a commemoration of the dead. Death was nowhere considered a natural event but always the result of malevolent animal spirits or witches or the deadly machinations of sorcerers. If death was thought to be caused by human agents, the soul of the deceased would never rest until vengeance had been secured by living relatives.

Death and afterlife

It was also believed that animals possessed souls. Slain animals sought vengeance against man through the agency of their species "chief," a mythological animal with great supernatural power. Hunting thus became a sacred act, much imbued with taboos, ritual, and sacrifice. Most disease was attributed to failure to placate the souls of slain animals.

The plant world was considered friendly to mankind, and the Cherokees thought that every animal-sent disease could be cured by a plant antidote. The economic significance of maize was memorialized by the near universality of the Green Corn, or Busk, ceremonies throughout the Southeast. This major ceremonial was suffused with an ethos of annual renewal in which the sacred fire was rekindled; old debts and grudges were forgotten; and a sense of community was regenerated.

Spiritual power might also reside in physical objects. Medicine men possessed sacred stones, quartz crystals, and other mystically endowed paraphernalia. Other objects were consecrated to symbolize the collective solidarity of the group. The Cherokees made use of a palanquin or litter within which were placed revered objects; the Tukabahchee Creeks possessed sacred embossed copper plates; the temples of several Lower Mississippi tribes contained an assortment of idols and icons. Natural objects could be infused with sacred power in a variety of ways: contact with thunder, as in lightning-struck wood; immersion in a rapidly flowing stream; exposure to the smoke of the sacred fire or of ritually prepared tobacco.

Mythology

Remnants of a formal theology can be reconstructed from early accounts of some of the stratified societies and of tribes who survived the immediate ravages of European contact. Most groups possessed origin myths, often involving a primal deluge into which prototypical beings from on high plunged to secure a portion of mud that magically expanded to create the Earth island. The subsequent course of mythological history is frequently related in terms of a cosmic struggle between good and bad culture heroes, one of whom bestows boons on mankind, the other serving as the source of the fatality and misfortune inherent in the human condition. Southeastern myths and folktales are populated by an incredible host of nature spirits, monsters, tricksters, giants, and little people.

Among many tribes, evidence survives that suggests belief in a supreme being, or master of breath. This ultimate divinity was frequently associated with the sun and its earthly aspect, fire. In addition, the world was viewed as divided into four quarters defined by the cardinal directions; each section had a presiding deity and appropriate colour symbolism. Concern with the remote supreme being seems to have rested more with the priesthood than with the everyday activities of the average man. The life of the latter was more intimately tied up with more proximal spiritual beings who were felt to intervene more directly into human affairs.

Priests and diviners. In some of the wealthier stratified societies, priests were given specialized training and became full-time religious practitioners responsible for the spiritual health of the community and assuming responsibilities for conducting the major collective religious rituals that punctuated the annual cycle.

Complementary to the priesthood were various individual magico-medical practitioners, such as sorcerers, conjurors, diviners, and medicine men, who were generally only part-time specialists and catered to individual needs and crises, especially the treatment of illness. Medical therapy was intricately enmeshed in the magical view of the world but might include such practical procedures as isolation, sweating, bathing, bloodletting, sucking, vomiting, and the internal and external application of herbal medicines.

EVOLUTION OF THE CULTURES TODAY

The complex history of the Southeastern Indians after contact with Europeans can be outlined here only briefly. The 16th century witnessed European exploration of the Southeast, though without permanent settlement. During the 17th century, white settlement was established on the coastal fringes, and the deerskin trade made the Indians increasingly dependent on whites for firearms, metal tools, and luxury items. By the 18th century the Southeastern Indians became counters in the power struggle between France, Spain, England, and the nascent United States for control of North America. With the ascendance of the United States, the military threat of the Indians became gradually neutralized, and Indians ceded large tracts of their land in an attempt to placate the insatiable appetite of land-hungry frontiersmen and the ensuing waves of white settlers.

Removal to Indian Territory

A short-lived Indian renaissance occurred during the first third of the 19th century, when the major surviving Southeastern groups became known as the Five Civilized Tribes. Unabated pressure for Indian land continued, however, and federal policy eventually culminated in the 1830s in the removal of the Cherokees, Choctaws, Creeks, Chickasaws, and Seminoles to the Indian Territory west of the Mississippi. In the 1890s the tribal governments of the Five Civilized Tribes were dissolved and their lands were allotted, thus paving the way for Oklahoma statehood.

Most of the surviving Southeastern Indian population currently resides in Oklahoma. Although this population reflects a full spectrum of assimilation—from oil company executives to culturally conservative "full bloods"—the increasingly numerous segment of people occupying the culturally conservative end of the continuum live in poverty. Small contingents of Cherokee, Seminole, and Choctaw who managed to escape the general removal of the 1830s presently live on small reservations in their traditional homelands. In addition, several remnant groups, such as the Catawba, Lumbee, and Houma, remain in the Southeast, though much of their cultural distinctiveness has disappeared.

BIBLIOGRAPHY. JOHN R. SWANTON, *The Indians of the Southeastern United States* (1946), is the standard starting point for information on the traditional cultures of the Southeast. Other indispensable works by Swanton are *Indian Tribes of the Lower Mississippi Valley and Adjacent Coast of the Gulf of Mexico* (1911), *Early History of the Creek Indians and Their Neighbors* (1922, reprinted 1970), *Social Organization and Social Usages of the Indians of the Creek Confederacy* (1928), *Religious Beliefs and Medical Practices of the Creek Indians* (1928), *Social and Religious Beliefs and Usages of the Chickasaw Indians* (1928), *Myths and Tales of the Southeastern Indians* (1929), *Source Material for the Social and Ceremonial Life of the Choctaw Indians* (1931), and *Source Material on the History and Ethnology of the Caddo Indians* (1942). Also valuable are various works by FRANK G. SPECK, including, *The Creek Indians of Taskigi Town* (1907), *Ethnology of the Yuchi Indians* (1909), and, with LEONARD BROOM, *Cherokee Dance and Drama* (1951). JAMES MOONEY, *Myths of the Cherokee* (1900), remains the basic source on the Cherokee and may be usefully supplemented by WILLIAM H. GILBERT, *The Eastern Cherokees* (1943). For the Seminoles, CLAY MACCAULEY, *The Seminole Indians of Florida* (1887); ALEXANDER SPOEHR, *Camp, Clan and Kin Among the Cow Creek Seminole of Florida* (1941), and *The Florida Seminole Camp* (1944); and various articles by WILLIAM C. STURTEVANT, referred to in his essay "Creek into Seminole" in E.B. LEACOCK and N.O. LURIE (eds.), *North American Indians in Historical Perspective* (1971), provide basic background. For the piedmont tribes, see JAMES MOONEY, *The Siouan Tribes of the East* (1894); and recent works by DOUGLAS S. BROWN, *The Catawba Indians* (1966); and CHARLES M. HUDSON, *The Catawba Nation* (1970).

On more specialized topics, MARY R. HAAS, "Southeastern Indian Linguistics," in C.M. HUDSON (ed.), *Red, White, and Black* (1971), is the most reliable and up-to-date published review of linguistic relations; WILLIAM C. STURTEVANT, *The Significance of Ethnological Similarities Between Southeastern North America and the Antilles* (1960), clears up some difficult issues; EMMA L. FUNDABURK and MARY D.F. FOREMAN, *Sun Circles and Human Hands* (1957), provides fine illustrations and commentary on prehistoric and historic Southeastern art; JOHN WITTHOFT, *Green Corn Ceremonialism in the Eastern Woodlands* (1949), is an important contribution to Southeastern religion; as is JAMES H. HOWARD, *The Southeastern Ceremonial Complex and Its Interpretation* (1968). Many detailed scholarly works exist on the post-contact history of the Southeastern Indians, among which might be mentioned GRANT FOREMAN, *The Five Civilized Tribes* (1934, reprinted 1971), and *Indian Removal* (1932); ANGIE DEBO, *And Still the Waters Run* (1940, reprinted 1966), and *The Rise and Fall of the Choctaw Republic* (1934); ROBERT S. COTTERILL, *The Southern Indians* (1954); VERNER CRANE, *The Southern Frontier, 1670–1732* (1929); DAVID H. CORKRAN, *The Cherokee Frontier* (1962) and *The Creek Frontier* (1967); and THURMAN WILKINS, *Cherokee Tragedy* (1970).

(R.D.Fo.)

Southeast Asian Cultures

Southeast Asia comprises those continental margins and offshore archipelagoes of Asia lying south of China, north of Australia, and east of India. Its continental portion embraces the autonomous nations of Burma, Thailand, Laos, Cambodia, and the two Vietnams. Its insular portion consists of the island nation of Singapore, at the tip of the Malay Peninsula, and the two sprawling archipelago nations of Indonesia and the Philippines. It also includes Malaysia, a political amalgam of two other northern Borneo states and the Malay states of the peninsula. In addition, there are the Portuguese-ruled colony of eastern Timor and the British dependency of Brunei in northern Borneo. The region extends over 2,000 miles (3,200 kilometres) from north to south and 3,500 miles (5,600 kilometres) from east to west. In mid-1971 it had over 300,000,000 inhabitants, or a little over 8 percent of the world's population.

Most of Southeast Asia has a stable, homogeneous tropical climate. Its large, shallow archipelagic seas and extensive fringing bays help stabilize the region's average monthly temperatures at around 81° F (27° C).

GENERAL CHARACTERISTICS OF SOUTHEAST ASIA

Geography and languages. Four vast, southward-flowing river systems shape continental Southeast Asia's topography and major settlement patterns. They are the Irrawaddy and Salween (Burma), Chao Phraya (Thailand), and the Mekong (marking the Thailand–Laos frontier and traversing both Cambodia and South Vietnam). The shorter, eastward-flowing Red River (Song Hong) reaches the Gulf of Tonkin farther north, near the Chinese border. All except the Salween flow through broad alluvial plains and fertile deltas, where intensive rice agriculture sustains dense population centres and large cities. No comparably large river systems exist in the islands. Closest in length are the huge meandering


INDIA
CHINA
TAIWAN
BURMA
Shan
Palaung
Lahu
Meo
Yao
Vietnamese
Lao
Mon
Thai
Khmer
Cham
Jarai
Rhade
Rejang
Malay
Minangkabau
Batak
Achinese
Gajo
Dayak
Iban
Bajau
Dusun
Samal
Tagbanua
Tagalog
Bicolano
Pangasinan
Pampangan
Mangyan
Cebuano
Ilonggo
Bukidnon
Subanon
Maranao
Yakan
Tausug
Sangir
Toradja
Buginese
Makasarese
Sundanese
Javanese
Madurese
Balinese
Sasak
Lampung
Ifugao
Ilocano
Kalinga
Apayao
Tinggian
Bontoc Igorot
Kankanay
Sagada Igorot
Ibaloi
Tboli
Bagobo
Magindanao
Tiruray
Tasaday
Maguindanao
Maranao
Kalagan
Tausug
Samal
Yakan
Sama
Badjao
Kadazan
Murut
Kelabit
Punan
Kayan
Kenyah
Bisaya
Land Dayak
Sea Dayak
Ngaju
Ot Danum
Kadazan
Iban
Kayan
Kenyah
Punan
Kelabit
Murut
Bisaya
Land Dayak
Sea Dayak
Ngaju
Ot Danum
Bajau
Samal
Dusun
Tausug
Sama
Badjao
Kadazan
Murut
Kelabit
Punan
Kayan
Kenyah

skills, social patterns and religions of two major civilizations. From about AD 100 to 1500, Indian, rather than Chinese, civilization became imprinted most explicitly in the region. Chinese influence spread overtly only among the people of Vietnam. Throughout "farther India" (Burma, southern Thailand, Cambodia, and southern Vietnam), on coastal west Malaya, in eastern Sumatra, and on Java's north coast, pioneer Indian settlements grew up along the trade routes. As they waxed commercially, they expanded their religious and political influence. Indian Sanskrit scholars, monks, and priests settled and helped consolidate these communities into regional, theocratic, and socially stratified "Indianized states," economically based upon irrigated rice agriculture. These Indianized states were variously and at different times either Buddhist or Hindu-Brahmin in their dominant religious orientation. Until the mid-14th century they struggled among themselves for political supremacy and the control of regional trade. In doing so they implanted Indian religions, legal codes, social patterns, cosmologies, syllabic writing systems, classical literature, graphic arts, drama, and music throughout much of Southeast Asia.

A second major historical influence has been that of Islām. Between the 13th and the 15th centuries Arab and Indian Muslim traders began to dominate the maritime trade with the coastal peoples of Sumatra, Malacca, the Malay Peninsula, and many of the islands eastward to Mindanao. Often settling and intermarrying among the local ruling families of these scattered principalities, they imparted a pervasively Islāmic way of life. Islām embodied a monotheistic philosophy of religious brotherhood, revealed and concretely chartered by the Qur'ān. It contrasted with the already widespread polytheistic Hinduized religion that emphasized caste-segmented ritual and a variety of beliefs. By the 17th century, missionaries and converted chieftains (sultans) had widely extended Islām along the trading routes to Indonesia's Spice Islands (Moluccas), into the southern Philippines, and, on the mainland, even among the Cham population of southern Vietnam. While also helping to undermine Java's Indianized state of Majapahit, Islāmic culture had begun by this time to penetrate deeply—mainly in insular Southeast Asia—into the belief systems, languages, literatures, and domestic law of many indigenous peoples.

Indigenous culture traits

Despite the spread of Indian, Islāmic, and Chinese civilizations into the region, Southeast Asia's interior, less accessible peoples have retained certain widespread, indigenous culture traits from the pre-Christian era. These include slash-and-burn (shifting field) agriculture, the terrace irrigation of wet rice, and the domestic use of chickens, pigs, and water buffalo. Other such traits are a highly diversified use of bamboo (to construct houses, tools, weapons, musical instruments, containers, etc.), separate premarriage dormitories for men and women, megalithic monuments, pile-supported dwellings, and multifamily longhouses. Associated patterns have included trial marriage, the blood feud, headhunting, strong local patterns of etiquette and custom, weak institutions of political leadership, and the practice of making decisions through rituals of ordeal and divination. Most of the technologically simpler peoples of even the remote interior have acquired the use of iron and the blacksmith's skills. Yet, such has been their generally ingenious, careful, and effective use of all available plant and animal resources that the French geographer Pierre Gourou was able to describe Tonkinese peasant life in northern Vietnam before World War II as a "civilization of the vegetable kingdom."

CULTURES OF THE MAINLAND

The Burmese, Thai, Cambodian, and Laotian peoples. The lowlands of Burma, Thailand, Cambodia, and Laos are each occupied by one politically predominant ethnic group—respectively, the Burmese, Thai, Khmer, and Lao peoples. In 1970 they probably totalled about half of mainland Southeast Asia's population. The Burmese inhabit the valleys of the Irrawaddy, Sittang, and Chindwin rivers, the Arakan Coast, and the Tenasserim panhandle; they extend northward to about 25° N. The Thais live on the alluvial plains of the Mae Nam Chao Phraya (Chao Phraya River) and its tributaries in central and south Thailand. The Khmers, or Cambodians, inhabit Cambodia's interior plain drained by the Mekong and Tonle Sap rivers. The Lao occupy both banks of the Middle Mekong River and extend into Thailand. Although separated by language, ethnicity, and nationality, these peoples are alike in their agrarian economy, social organization, and adherence to Theravāda Buddhism.

Settlement patterns

These mainly rural peoples live in settlements that vary from compact house clusters to the more common linear communities extending along rivers, irrigation canals, roads, or paths. The size of a settlement may range from a few families to over 3,000 people. Houses are rectangular, gable-roofed, and usually built on piles two to six feet above the ground; they range in width from 10 to 25 feet and in length from 20 to 40 feet. Most have plaited bamboo walls and floors, with thatched roofs; houses built of planks with roofs of tile or sheet iron usually proclaim wealth and high status. Better houses may contain several rooms and have shuttered windows; poorer houses often consist of single compartments opened only by doors. Kitchens, often essentially wood-framed earthen hearths with stones to support cooking pots, may be separately partitioned, sometimes occupying rear verandas. Roofed front verandas, reached from the ground by short ladders, serve among the Burmese and Thai for other household chores and for sociability. Furnishings are usually simple but may include low eating tables, bed mats, storage bins, baskets, water jars, shelves, Buddha images, household altars, and spirit houses. Almost every settlement includes a Buddhist temple compound and, additionally, among the Burmese, a brick pagoda containing a religious relic. Temple compounds in the larger settlements of all these peoples may also contain monks' quarters, meeting halls, and tomb monuments.

Rice, the major crop and staple food, is grown in many different varieties, usually selected for their productivity and endurance under local conditions. It is generally grown in permanent, diked fields watered by rainfall or —as in Burma's dry zone—by irrigation. Both annual cropping and multicropping are widely utilized. Draught buffalo (water buffalo) and bullocks are commonly used to plow and harrow the fields for rice planting; animal dung, ashes, and paddy stalks are frequently, if erratically, used as fertilizer. Common planting techniques include broadcasting, dibbling, and transplanting from seedling nurseries, depending upon local soil and water conditions. Hand sickles are generally used for rice harvesting. Threshing and winnowing techniques are quite varied and include trampling by buffalo.

Supplementary vegetables, tree crops, and other crops are also widely grown, both in houseplot gardens and in rotation on multicropped ricefields. Vegetables include beans, eggplants, tomatoes, onions, garlic, chillies, cucumbers, yams, sweet potatoes, white potatoes, gourds, cassava, peppers, ginger, mint, turmeric, and other exclusively local plants. Tree crops include bananas, coconuts, mangoes, durians, betel nuts, citrus, jackfruit, papaya, guavas, tamarind, and sapodilla. Other crops, sometimes raised commercially, include rubber, cotton, maize, millet, sesame, sugarcane, and various condiments. No single farming household, however, will normally raise more than a few of these items. Burma's Tenasserim area specializes in tree crops, while in Cambodia the sugar palm (*Borassus flabellifer*) has become a national symbol because of its local importance as a source of thatching material, fruit, sugar, and juice.

A major animal food is fish—fresh, dried, or salted. Although particularly important in the Burmese diet, it is found everywhere in local trade and, especially in its dried and salted form, readily reaches even communities distant from fishing areas.

Social organization typically is based upon the nuclear-family household—*i.e.*, essentially parents and their unmarried children; there is no tradition of extended-family groups or clans. Traditional society is largely stratified in hereditary ranks according to degree of royal ancestry.

aristocratic background, and even religious or military achievement. The variety of such ranking has allowed individuals to move from one rank to another and occasionally even from the bottom of the traditional social ladder—*i.e.*, from among formerly indentured people and slaves.

Religious practices among mainland Buddhists

Religion is a blend of Theravāda Buddhism and local spirit worship. Buddhist theology posits that man undergoes a cycle of separate existences and that all animate life is endowed with individual souls. The moral and behavioral guidelines of his faith enable a devout Buddhist to progress toward his goal of Nirvāṇa by acquiring merit during his lifetime. There are many merit-earning acts. They include becoming a monk, financing construction of a temple, having one's son ordained as a monk, visiting Buddhist shrines, contributing to the repair of a temple, giving daily food to monks, entering the monkhood as a novice, attending the temple on all Buddhist holy days, and generally obeying Buddhist precepts. Conversely, by flouting Buddhist precepts, one can diminish one's stockpile of merit and even risk being reincarnated as a less desirable form of life in the next existence. Devout Buddhists are enjoined from taking life and usually even refrain from killing mosquitoes. Yet, this does not prevent most from eating meat or fish. If another person kills an animal, no merit is lost by eating its meat. The fisherman loses no merit because his catch is thought to die from leaving the water and not from the fisherman's action.

The worship and propitiation of local spirits provides the complementary, non-Buddhist religious dimension. Known as *nat*s in Burma and *phi* in Thailand and among the Lao, these spirits are ubiquitous, are thought to exercise great influence over daily human fortune, and are susceptible to approach and manipulation. The Burmese formally recognize 37 major *nat*s and a horde of lesser ones; the Thai, Khmer, and Lao recognize comparably large numbers. There are spirits of one's house, village, and ricefield; there are many others of lakes, rivers, forests and of objects and periods. Equally various are the rituals necessary to keep them placated and amicable. Accidents, sickness, epidemic, drought, floods, and domestic problems are generally blamed upon the malevolence of spirits. Individual households and villages maintain shrines for placating their respective guardian spirits, usually through offerings of food and flowers. Male and female mediums can, when in trance, communicate with certain spirits and ascertain from them the shape of future events.

Because entering the monkhood is a major merit-generating act, most men spend part of their lives as monks. A young man will be initiated at about 14 years of age as a novice monk and remain in the monkhood temporarily, for several months or half a year. His initiation, a ritual re-enactment of Lord Buddha's own renunciation of material wealth and assumption of monastic discipline, is a major festive occasion; the initiate's head is shaved, and he receives a new name, recites the monastic vows, and dons the monk's orange robe. During his monkhood the novice obeys the same rules governing diet, celibacy, and material possessions that discipline the senior monks. He lives in the temple monastery, begs his morning food in the community, and recites the sacred texts. He usually leaves the monastery without taking the examinations for advancement in the monkhood. He may return later in life either for another short period or to enter the monkhood permanently.

The full-time Thai Buddhist monk usually assumes an active pastoral role in the community of his home temple. The role commits him to officiate at the weekly temple services and at the monthly and seasonal rituals. Normally, it also requires that he conduct individual family ceremonies, as invited, for weddings, house blessings, cremations, and ordination of novices. He is a locally influential religious counsellor on important matters such as determining auspicious dates for personal activities. He may even study *sajjaad*, or magical healing, and thus extend his leadership into matters of health.

Since illness is often attributed to hostile spirits, its cure is not an established responsibility of the professional Buddhist monk. There are many sorts of locally recognized healers. Among the Thai, they include the spirit doctor (*maw phii*), who, through therapy such as text chanting, body massage, and the use of herbal medicines, is able to coax out, or extract, the hostile intrusive spirit. A *maw khan* specializes in recalling a patient's wandering personal soul and thereby restoring his health. Mediums can often diagnose illness by singling out the responsible *phi* and even, when in trance, contacting it directly for therapeutic advice. Such healers and diagnosticians may employ explicitly Buddhist symbols and ritual acts, while the pastoral monk, on the other hand, may utilize "magical healing" among his congregation; yet, healers and monks seem seldom to conflict. Using different techniques, they both provide some "mental-health" therapy; in addition, the healers are frequently astute both as physical therapists and as physicians.

The Vietnamese. The Vietnamese are heavily concentrated in the northern Red River Plain of Tongking and in the southern Mekong Delta area. They are settled only thinly along the coast between. In the north, where population density is very high, rice cultivation is supplemented by root crops such as manioc, taro, and sweet potatoes; in the south, more exclusively rice-growing areas predominate. Villages tend to specialize in pursuits such as metalworking, silk production, and coastal fishing; trade between such specialized communities has traditionally been important.

Chinese cultural influence in Vietnam

Chinese influence has profoundly affected Vietnamese society, setting it apart from the mainland societies farther west. Vietnamese use paternal surnames and, for purposes of ritually honouring their paternal ancestors, congregate into patrilineal clan groups. The oldest male of the clan's senior branch usually conducts the ritual and maintains the ancestral clan shrine, its altars, and its commemorative tablets. Since this vital responsibility descends only among males, the importance of having male children is stressed more than in Thailand or Burma. Vietnamese Buddhism, formally of the Mahāyāna school, is also of Chinese origin. Far fewer men enter the monkhood in Vietnam than in the westerly Theravāda mainland countries. Those who do usually become full-time professionals; they follow primarily scholarly pursuits, do not beg for daily food in their communities, and play little or no pastoral role. Some become divination specialists, making use of Taoist ritual techniques, another Chinese cultural influence.

Vietnamese society has also incorporated traditional Confucian social doctrine binding subject to ruler, son to father, wife to husband, younger to older brother, and friend to friend in ideally respectful, permanent relationships. These ties have closely supported Vietnam's traditional hierarchical systems of civil and religious authority extending from the supreme emperor down to the village-level "council of notables." The intervening bureaucracy, mainly educated Mandarin landholders, has traditionally represented the emperor and is split into a number of ranks, each with its special privileges, powers, and status trappings. One could attain or better one's rank in the Mandarin bureaucracy through education and by passing examinations in Chinese characters and Classics. While this fostered some mobility both in rank and in residential location among the bureaucracy, the peasantry remained closely rooted to its villages. This "rootedness" is religious in nature, exemplified in the villagers' attachment to local clan shrines and also in the Vietnamese *dinh*. The *dinh* is a walled compound containing a sanctuary for the village's patron deity—often a legendary or historical personage—together with sanctuaries for lesser local spirits. While the *dinh* does not contain monk's quarters, it often provides a meeting place for the discussion of village affairs. Above all, it is a focal point for the discharge of the villager's crucial ritual obligations to the locally powerful deities and to the numerous spirits that, in traditional belief, control his world.

Montagnard cultures

The hill peoples. In Vietnam, as throughout Southeast Asia, the lowland peoples tend to be culturally distinct from their upland neighbours, even when they belong to

the same ethnolinguistic group. Vietnam's own interior hill peoples, the many groups collectively known by the French word Montagnard (Mountaineer), have traditionally been despised by the surrounding Vietnamese, Cambodians, and Thai as "savages." Like the other scattered hill and mountain peoples of northern Thailand, Laos, and Burma, most are economically dependent upon shifting field cultivation and upon supplementary trade in the products of forest and of local craftsmanship. Mostly they live in scattered, impermanent settlements often located on steep slopes at high altitudes. They are predominantly and elaborately animistic in their religious beliefs and ritual practices, and their community leaders are usually endowed with some shamanistic ability as curers, diviners, or trance mediums.

Rice is the most desired food and usually also the staple crop among the mainland hill peoples. Some groups acquire it mainly by trading other crops for it, such as tea and opium. In Laos the hill peoples are generally self-sufficient in rice but also produce cane, wax, Spanish peppers, shellac, cardamom, and many other trade items; these often pass through the ethnically Laotian, river-mouth trading villages along the Mekong tributaries to their ultimate lowland buyers. In return, the mountain farmer receives such essential items as cloth, steel tools, and salt. Today, he also receives an increasing flow of Western consumer goods and even cash.

INDONESIA

Varieties of Indonesian societies

Major Indonesian groups. *Rice farmers of Java and Bali.* The peoples of Indonesia may be divided into several categories. The first includes the irrigation rice farmers of the strongly Hinduized inland areas of Java and Bali. These two peoples comprise about half of Indonesia's total population, and their densely settled homelands reflect Indonesia's severest population pressures. But the Balinese have escaped the Islāmization experienced by the Javanese; their expressively rich religion, Ugama Bali, shows the widespread influence of classical Indian cosmology, ritual organization, and litany. Such influence may also be seen in the ancestral shrines, village temples, and courtyard structures adorned with Hindu art motifs carved into wood and volcanic stone. While both Balinese and Javanese share a common, pervasively Hinduized cultural heritage, the Balinese retain certain old Hindu caste distinctions and stress status differences more formally than the Javanese, especially those between aristocrat and commoner. By his simultaneous membership in a number of groups based on kinship and occupation, the typical Balinese villager also participates in a formally structured local community.

Islāmic peoples of Indonesia. The second category of societies includes the trade-oriented Islāmic peoples scattered along the coastal and riverine portions of the archipelago. Economically, these peoples range from full-time merchants to petty traders who supplement their livelihood by occasional gardening and cottage industry. In religious orientation, they are Muslims and share a fundamental respect for Qur'ānic tradition and law. While linguistically facile in Malay, which serves as their commercial lingua franca, they also speak a variety of native languages. The three largest constituent societies are those of the Malays of lowland eastern Sumatra and western Borneo and the mutually similar Makasarese and Buginese societies composed of peoples who stem from southwestern Celebes but are today scattered widely along the coast of the archipelago. Other small societies include hamlets, villages, and port-town enclaves of mixed Javanese, Arab, Portuguese, Indian, Chinese, Dutch, and other ancestries.

Marginal peoples. A third category of Indonesian societies comprises mainly isolated peoples unaffected by Indian or Islāmic influence. Their economy has consisted mainly of shifting-field rice cultivation or, as in extreme eastern Indonesia, the gardening of sago and various root crops. By the late 19th century, several of these peoples had experienced Christian-missionary influence. Until the 20th century, many retained their traditional customs of sporadic intervillage warfare and headhunting. Their traditional religious beliefs were richly animistic, with conspicuous emphasis upon ancestor-revering ritual. Included in this category are the Toradja of central Celebes, the scattered Dayak peoples of Borneo, a number of interior peoples of Nusa Tenggara (principally the islands of Lombok, Sumbawa, Flores, Timor, and Sumba), several smaller Sumatran groups (Gajo, Rejang, Lampung), and certain peoples of Ceram and the Moluccas in extreme eastern Indonesia.

Major Sumatran peoples. Three other major peoples inhabit Sumatra: the Achinese of the island's northern end, the interior Batak somewhat farther south, and the upland Minangkabau of west central Sumatra. None has experienced the extensive Indian cultural influence felt in Java and Bali. Yet, historically, all have adopted foreign religions: the Achinese and Minangkabau have long been Islāmized, while the Batak have been significantly and increasingly influenced by Christian missionaries. All three peoples depend heavily upon rice agriculture (both wet and dry) and share western Indonesian patterns both of language and of material culture.

The matrilineal Minangkabau

The Minangkabau have traditionally occupied handsome, multifamily wooden longhouses (*rumah gadang*). Together with the historically cognate Minangkabau Malays of the Malay Peninsula (Negeri Sembilan state), they probably constitute the world's largest matrilineal society. Each Minangkabau man and woman, even after marriage, maintains permanent membership in his or her maternal, natal longhouse. Hence, Minangkabau fathers often split their residence between their wife's and their own natal longhouses. Their children, growing up in the mother's longhouse, tend to experience their fathers as "visitors," whereas their frequently co-resident uncles (mother's brothers) of the same longhouse often become equally or more familiar. In his wife's longhouse a man holds no property and, at best, only exercises supervisory control over his wife's property affairs. In his own natal longhouse he inherits property through his mother and exercises influence, control, and even formal leadership both in the property and the ritual affairs of his maternal "corporation." He tends to assume a partial "father role" toward his sister's children, while, with his own children, he partially relinquishes it to his wife's brothers. By Islāmic law he is entitled to four wives simultaneously, but polygyny among the Minangkabau has been relatively rare. So also has been the observance of Islāmic property law, which favours males as heirs over females, for it conflicts sharply with Minangkabau customary laws of inheritance.

Although matrilineal organization is unique among the Minangkabau in Indonesia, it may once have been more widespread. It also occurs in Negeri Sembilan and even among certain Montagnard groups in Vietnam, such as the Mnong. The traditional Minangkabau clans have steadily subdivided into more numerous, locally autonomous sub-clans and have lost some of their traditional ways. In Negeri Sembilan the introduction of smallholder rubber cultivation in the early 1920s began to divert labour away from the traditional wet-rice cultivation, which had been a major source of economic strength for the matrilineal clan.

The Javanese. Java comprises about 8 percent of Indonesia's total land area (excluding Irian Barat), but with Madura, a small island to the east, it supported over 76,000,000 people in 1971, or 64 percent of Indonesia's population of 119,000,000. Three culturally similar ethnolinguistic groups predominate on Java and Madura: the Javanese, the Sundanese, and the Madurese. The Javanese, occupying central and eastern Java, are most numerous, numbering in 1970 about 54,000,000.

Sedentary farm communities in Java

Most Javanese are lowland sedentary farmers, tilling small irrigated ricefields and household gardens. A large Javanese population also inhabits the mountainous interior, growing cassava as a staple crop. Upland villages consist mostly of dispersed hamlets; in the lowlands relatively nucleated communities predominate, although usually strung along roads or rivers. Rural houses are small, rectangular, and usually built close together on the ground. House frames are of wood, walls of plaited bam-

boo, roofs of thatch, and floors of tamped earth. Small gardens, rice granaries, and livestock yards (for chickens, goats, and buffalo) frequently augment the rural family homestead. The countryside is interlaced by many hundreds of densely populated, contiguous rural settlements. The average "landed" farmer owns perhaps 1.24 acres (0.5 hectare) of cultivable land (rice and garden); many others own far less, and millions subsist entirely by renting or sharecropping land. Most rural Javanese live at a bare subsistence level.

Javanese society divides into two major strata: a small upper stratum and a large lower stratum of rural villagers and manually employed townspeople. Javanese also recognize a vertical division into two religious groups: the *santri* and the *abangan*. *Santri* Javanese seriously observe Islāmic principles in their daily life; they conduct the five daily prayers, fast during the *pasa* month (Ramaḍān), refuse to eat pork, attempt to make the Mecca pilgrimage, pay their religious tithes, and obey other Islāmic obligations. *Abangan* Javanese proclaim no such devotion and ignore or even flout, most of these observances. Most Javanese are *abangan*, but *santri* are found widely throughout Javanese society. In the ancient, princely court areas of central Java and in many towns, there are conspicuous localized *santri* communities of merchants, craftsmen, and entrepreneurs. Scattered *santri* families are widely influential also among the nobility, in urban intellectual circles, and in rural villages.

The Javanese life cycle is strongly influenced by non-Islāmic religious beliefs about cosmic and social order. While one's social position and ultimate fate are foreordained, correct family-focussed ritual helps prevent accidents, illness, and misfortune. The central ritual, or *selamatan*, is basically one of distributing sanctified food; this explicitly generates spiritual awareness and emotional equilibrium within the family, particularly during crises. Major family *selamatan*s are performed during pregnancy and at childbirth, circumcision, betrothal, marriage, and death. An elaborate series of protective *selamatan*s also follows childbirth until, on the 245th day, the postnatal crisis period ends and the baby ritually "descends to Earth" during a major family *selamatan*. Another *selamatan* series also follows a death; it usually symbolizes the deceased's gradual metamorphosis into an idealized, sacred forebear who will continue, it is believed, supernaturally to assist his living descendants. Participants in such family *selamatan*s include one's kinsmen, neighbours, and—to recite the appropriate Qur'ānic verses and prayers—local religious leaders.

Child rearing and marriage ceremonies

Social order and emotional self-control are also dominant values in Javanese child rearing. Whatever his social stratum, a Javanese child begins early to learn the complex language etiquette and other behaviour denoting respect toward his parents, other senior kinsmen, and—more widely—all senior people. Discipline begins at about two years of age and variously assumes the form of scolding, bodily punishment, supernatural punishment (by ghosts, monsters, evil spirits), or invidious comparisons with older siblings or unrelated age peers. The shy and withdrawn personalities of many adult Javanese seem to be partly the result of this conditioning. Among preadolescents, boys tend to expand their outside social contacts more widely than girls; the latter, even when in school, maintain responsibilities at home in cooking, paddy threshing, and care of younger siblings. For boys, circumcision between the ages of 10 and 15 marks the transition to adolescence and, among *abangan*, to the ideally proper "Muslim" status; for girls the transition is observed by a *selamatan* at first menstruation. Many Javanese parents avoid imparting sexual knowledge to their children. Chaperonage of unmarried women is fairly common, especially in upper Javanese society, and premarital pregnancy brings great shame to a girl's family. Abortion is considered sinful.

Marriage celebrations involve the most intricate and important rituals of the Javanese life cycle. Weddings are occasions of great display, rich in Hindu–Buddhist symbolism. While most marriages are probably no longer parentally arranged, parental approval remains an important factor. Young men often send intermediaries to gain the approval of a girl's guardian, who, following Islāmic law, is usually her father or brother. Getting married is a protracted process of betrothal (negotiation, gift exchanges, formal announcement), setting an auspicious wedding date, holding the ceremony, paying the bride price, celebrating the subsequent wedding *selamatan*, and possibly carrying out postwedding familial exchanges. The usual marriage age, especially in rural areas, is 15–18 for girls and 17–20 years for young men, and a newly married couple is usually not expected to be economically or residentially independent. The couple frequently lives for several years with the bride's parents.

During and after middle age, many Javanese become increasingly preoccupied with the relationships between natural and supernatural phenomena. Comprehending them requires sensitivity to *kebatinan*, problems of inner experience, as well as observation and knowledge. These are sought through individual and group meditation (*sudjud*), with the aid of systematic exercises, recitations, prayers, devotional offerings, abstinence, and visits to ancestral graves. The primary devotional symbol tends to be "one god," who may also be identified with a particular angel, ancestor, or spirit. Javanese meditation groups differ in their concepts of life, death, the afterlife, and the spirit world, but they usually claim that their philosophy represents *ilmu kedjawen*, or "Javanese science." While much of the imagery is indigenous, it also conspicuously includes both Hindu–Buddhist and Islāmic symbolism.

The Western tendency to make a sharp dichotomy between the natural and the supernatural is generally alien to most Javanese. Their religious philosophy contemplates, instead, the fundamental continuity of man with nature within the known and sensed universe. It stresses the inexorable dependence of man upon superior forces and teaches the value of constant awareness of—and mystical surrender to—these forces. As such, it evokes the expression of the pervasive Javanese values of harmony, balance, and aesthetic sensitivity.

THE PHILIPPINES

Some 70 ethnolinguistic groups are found in the Philippine archipelago. The dominant, aggregated society, however, consists of the mainly lowland-dwelling Christians on the few large islands, and the label Filipino usually refers to these people. They include a number of large ethnolinguistic groups, which, while strikingly similar in overall culture, differ in languages, food habits, art forms, dress, housing, and ritual practices. The

Linguistic groups in the Philippines

7,000,000 speakers of Cebuano live mainly in the Visayan Islands (central Philippines) and northwestern Mindanao. Speakers of Tagalog, the second ethnolinguistic group, number about 6,000,000; they have a high literacy rate and predominate in Manila, around Manila Bay, and in central southern Luzon. Ilocano speakers number between 3,000,000 and 4,000,000, settled mainly along Luzon's northwest coastal plain and in the Cagayan River Valley. Speakers of Ilongo (Visayan Islands), Bikol (Bicolano; southeast Luzon), Waray-waray (Samar-Leyte), Pampangan, and Pangasinan (central Luzon) constitute other large groups. In 1970 these groups probably included over 30,000,000 of the nation's total population of 37,000,000. Roman Catholics officially comprise about 84 percent of the national population, Philippine Independent (Aglipayan) Church members another 5 percent, and Protestants 3 percent. Practically all Christians belong to these ethnolinguistic groups.

The non-Christian minority peoples are either Muslim or pagan. Those professing Islām, derogatorily labelled Moros, constitute about 5 percent of the national population. About 90 percent of them belong to one of four relatively large ethnolinguistic groups: the Magindanao and Maranao peoples of west central Mindanao and the Tau Sug and Samal of the Sulu Archipelago to the southwest. Smaller Islāmic groups include the Bajau, Yakan, Melebuganon, and Sangir. Pagan peoples occupy the mountainous interiors of many of the large islands. In Mindanao they include the Subanon, Bagobo, Mandaya, Bukidnon, and Tiruray; in the Visayan Islands, the Tag-

banua (Palawan) and Mangyan (Mindoro); and in northern Luzon, some 10 ethnolinguistic groups in the former Mountain Province area (Kalinga, Apayao, Ifugao, Tinggian, and both the Sagada and Bontoc Igorots).

Family relationships among Filipino Christians

Filipino Christians. Christian Filipino society centres upon the nuclear-family group. The father officially heads the family, but women occupy a high social position and exercise much authority, as suggested by their control of 80 percent of Filipino business firms and their fairly dominant position in the scientific–academic community. Within the nuclear family, respect for elders is inculcated early in life; honorific pronouns and articles are obligatory in speech to parents and elders. Collective responsibility within the family is highly valued; older siblings must care for and, if necessary, make considerable sacrifices for their younger siblings, especially in educational opportunity. Stinginess toward close relatives is a serious breach of fundamental Filipino social values.

Social relationships outside the nuclear family are distinctively patterned. The individual recognizes relatives on both the maternal and paternal sides. By carefully bestowing gifts or services upon individual members of his kindred group, a Tagalog develops enduring ties of reciprocal obligation called *utang na loob*, literally "debt of the inside." Two persons become linked through a long period of alternating indebtedness that leads to a complementary relationship wherein obligations often cease to be calculated. While seldom explicitly discussed, such partnerships can involve deep feelings of mutual gratitude and friendship. Normally, an individual's social network includes many such dyadic ties in different stages of development. Even a child's obligation to care for his parents is thought ultimately to rest upon basic indebtedness to the parents for their gift of life. To renege upon one's *utang na loob* obligations is a grave breach of personal trust; it causes shame (*hiya*) and can even lead to violent revenge by the aggrieved.

Filipino society is stratified according to wealth, occupation, and education. Wealthier upper class people usually live in small towns or regional capitals, in non-Malay-style houses; they tend to have professional occupations. People of different status levels frequently establish ritual kinship ties with one another through the widespread Spanish Catholic custom of ritual co-parenthood (*compadre*). This can facilitate an even wider extension both of *utang na loob* relationships and of entire patron–client networks, the latter being an essential ingredient of political power. Socially important family alliances may also be consolidated by marriage. While national law and Ilocano custom both prescribe equal property inheritance, Ilocano parents, like other Filipinos, will often bestow all their property on only one heir. While concentrating family resources effectively, this custom also perpetuates parental control over "key" marriages and family linkages among many groups.

Rural settlements among the Christians are made up of Malay-style houses (bamboo, pile-supported, thatch-roofed) with their associated ricefields, gardens, and groves of coconut and bamboo. Domesticated animals are mainly chickens, pigs, and water buffalo. In areas of shifting agriculture, the farmer will often move away seasonally from his home to small guardhouses near the crops in distant fields. Where plantation-scale sugar or copra is grown, as in the Visayan Islands, Filipino worker settlements may be conspicuously large and compact.

Islāmic peoples

Other Philippine societies. The Islāmic peoples of the Philippines fall into several categories. The interior Maranao of Mindanao's Lake Lanao area, while primarily wet-rice farmers, also maintain vigorous cottage industries in mat making, weaving, woodwork, bronzework, and brasswork. Less conspicuous as craftsmen but depending both on fishing and rice farming are the Magindanao, the Philippines' largest group of Islāmic people, who inhabit the Cagayan River Valley inland from Cotabato. The Tau Sug, mainly occupying the northern Sulu Archipelago, depend upon extensive maritime trade, pearl diving, handicrafts, dry- (and some wet-) rice agriculture, and a varied horticulture and arboriculture. The Samal of the southern Sulu islands maintain homes, mosques, and agricultural plots ashore but live much of the time on boats, where they engage in fishing and maritime trade. Only their coastal dwellings and Islāmic community ties separate many of these Samal from the pagan Bajau, impoverished fishermen of the same area who live exclusively aboard small family houseboats. The Yakan, of Mindanao's Zamboanga Peninsula and Basilan Island, are fishermen, craftsmen, and gardeners but only occasionally Islāmic.

Social organization among these groups remains significantly traditional. The larger societies are composed of locally powerful social pyramids usually headed by hereditary aristocratic chiefs, or *datu*s. A pyramid includes the *datu* and his close kinsmen (including in-laws) of highest social rank, the usually related nobility of lesser rank, and a large subordinate group of dependent local retainers and commoners linked by ties of kinship, fealty, or debt bondage. A *datu*'s title will often proclaim his allegedly direct descent from the Prophet and symbolize his considerable local authority in both religious and civil matters. In practice he usually shares this power with an Islāmic religious adviser (*kadi*) and a council of ranking noblemen. Both locally and in their external ties to the greater society, such socio-political pyramids are sustained by intricate combinations of leadership charisma, kinship solidarity, respect for hereditary rank, family size, alliances by marriage, and control over resources, such as riceland and slave labour. Polygyny, cousin marriage, and hypergamy (female marriage upward) are three enduring customs by which, particularly among the nobility, kinship ties are used to bind groups together and become major channels of alliance between families of equal rank within different local pyramids.

Most small sultanates and principalities among the Maranao and Magindanao have this pyramidal organization. The Maranao are more politically fragmented and more recently Islāmized than the Magindanao. The Tau Sug, ethnically the core population of the traditional sultanate of Sulu, have been Islāmized since the 15th century. Historically vigorous in resisting all foreign political and religious domination (Spanish, U.S., and Christian Filipino), their small *datu*-led principalities maintain a regionally marked tradition of political coordination. Today, the Tau Sug are still widely regarded as conspicuously proud, aggressive, and devotedly Islāmic. For their close neighbours, the less Islāmized Samal, the Tau Sug have constituted an ethnically distinct superior social class. While a Tau Sug man may marry a Samal, rarely if ever does a Tau Sug woman marry "down" with a Samal. The Samal maintains a similar class position toward the Bajau, who, as they have begun to settle ashore and adopt Islām, are increasingly allowing their women to marry "up" among Samal men.

In their social organization, economy, technology, dress, art, and proud religious identity, the Islāmic peoples of the southern Philippines are most immediately similar in culture to the coastal peoples of southern Celebes and of western and northern Borneo, such as the Banjarese, Makasarese, Brunei Malays, and others. The similarities stem largely from their common experience of Islām, their traditional dynastic interrelationships, and their collective antipathy toward Christian colonizers. Today, the Philippines' Islāmic groups constitute an increasingly self-conscious political minority with distinct problems of national assimilation.

The many pagan minority peoples of the Philippines range from the small hunting bands of pygmy Negritos (on Luzon and Palawan) to the physically "Malay" interior rice farmers of northern Luzon, the Visayan Islands, and Mindanao. They also include the Tasaday, a tiny band of rain-forest hunters discovered in southern Mindanao only in 1970–71.

PATTERNS OF CHANGE

Traditional customs have changed greatly. Except for Thailand, the nations of the region have all experienced the colonial imposition of western legal and administrative codes. Burma and Malaysia have known the Pax Britannica, and Indonesia has experienced prolonged

Dutch rule. The Philippines have been subjected to successive Spanish and U.S. administrations. Laos, Cambodia, and the Vietnams have undergone French administration.

The persistence of old ways

Localized patterns of customary law remain. Many pagan and Islāmized peoples of the southern Philippines and northern Borneo have intricate conceptions of proper interpersonal behaviour and impose sanctions, especially fines, for violations of standards. These often involve questions of sexual and marital propriety and give rise to constant village-level litigation before local headmen or "native-court" adjudicative bodies. Below the level of litigation, sensitivity to minor local custom, or folkways, remains exceedingly widespread, especially throughout insular Southeast Asia. Few topics are more common in the folklore of the Malay languages area than themes of "other villages having other ways."

Divining is still common. Sarawak peoples such as the Iban (Sea Dayak) predict future events with the aid of a rich lore about the flying formations and habits of particular birds. They and other Borneo peoples also often read the entrails of recently killed animals to determine future events. Trial by ordeal is probably still common among many mainland hill peoples and in some interior regions of the archipelagoes; in the 1960s ordeal by water immersion was in occasional use before certain "native courts" of eastern Sabah (Malaysia). The reading of horoscopes is widespread throughout the region, even among educated urban people. The Burmese and Javanese have rich traditions of horoscopy, drawing judgments of the future from an enormous array of temporal, chromatic, auditory, and physiological cues.

The major religions have been dealt with above. In addition to the creeds introduced from outside, indigenous beliefs in natural and ancestral spirits are exceedingly common. Often they are fused with Buddhist, Islāmic, and Christian elements, but they are found in plain form among the smaller, more isolated, and still essentially pagan peoples. The widespread practice of revering ancestral spirits expresses the generally sacred significance of family continuity. The Vietnamese clan shrine is but one major example. The practice occurs in various forms among the Balinese, Batak, Burmese, Thai, Makasarese, and other peoples.

Ethnic images and regional patterns

Ethnic styles and stereotypes have a subtle but important influence on personal judgments and social relationships in Southeast Asia. Among Filipinos, the Ilocano is stereotyped as industrious, tough, frugal, and without a sense of humour; the Bicolano is supposedly even-tempered and religious. Among Islāmic Indonesians, the Makasarese is thought to be excessively proud, quick to take offense, and implacably vengeful. Such stereotypes are likely to be even cruder among culturally less similar peoples. Among the Burmans (the major ethnic group of Burma), for example, the Shan peoples of northern Burma tend to be regarded as uniformly "untrustworthy." Throughout much of Southeast Asia, immigrant Chinese businessmen are considered unfailingly industrious, penurious, and shrewd to the point of dishonesty. The Chinese often reciprocate by stereotyping others, as in Malaysia, as lazy, untrustworthy, and superstitious.

With all their diversity, Southeast Asia's people have several cultural patterns in common. Women command substantial social and economic authority; only among recently immigrant Arabs and Indians are they distinctly excluded from any social sphere, although on formal occasions there is often a physical separation of the sexes. The prevalent type of family organization is that of the nuclear, monogamous family linked to both maternal and paternal kinsmen; unilineally organized large family groups are found only among the patrilineal Vietnamese, the unilineal peoples of Sumatra and the Malaysian state of Negeri Sembilan, and among several hill groups of the mainland and of eastern Indonesia. Widespread sensitivity to, and preoccupation with, social status is another conspicuous pattern. This is visible in the still-traditional court etiquette of Thailand and central Java, but it is also apparent in the widespread sensitivity to language "levels" and to the implications of making a "proper" marriage. Such sensitivity goes with an equally widespread concern for maintenance of self-esteem, or "face"; there is an elaborate personal etiquette for maintaining outwardly harmonious social relations and thereby avoiding the infliction of shame or indignity upon another person.

Population movements

Recent tendencies. The growth of population and the expansion of cities are among the most important of recent trends in Southeast Asia. Manila, Saigon, Kuala Lumpur, and Bangkok are among the world's fastest growing cities. This urban growth opens up wide opportunities for the intermingling of different groups and cultures.

Some population shifts have resulted from government programs. South Vietnam's numerous relocation camps and resettlement schemes have forced vast numbers of uprooted Vietnamese civilians and Montagnards to settle among new neighbours and cope with new economic circumstances, often under very adverse conditions. The Indonesian government has resettled substantial numbers of Javanese in new communities in central Sumatra. Similarly, the Philippines government has resettled Christians from the crowded Visayan Islands to the northern and western sectors of Mindanao. Rural-development projects in Malaya have relocated substantial numbers of land-poor Malays to pioneer areas cleared for future rubber plantations.

In the 1960s newly opened rubber and palm-oil plantations in eastern Sabah (Malaysia) drew substantial numbers of Indonesian immigrant labourers from the Makasarese area of southern Celebes and from Indonesian Timor. Timber-cutting operations attracted many Christian Filipino labourers south to interior Mindanao. Among the Lawa hill peoples of northern Thailand, the increased costs of maintaining their trained timbering elephants has forced many to sell their animals and seek wage-labour jobs in the lowlands. Families and whole villages of Meo, Lahu, and Yao hill peoples have moved into northern Thailand from upper Burma and Laos in search of new, high-altitude farming areas for the cultivation of opium poppies. Thousands of Vietnamese families, dislocated since 1965 by the massive destruction of their agricultural resources through saturation bombing and herbicide spraying, have had to leave their home communities in search of new livelihoods.

Rise of national languages

Cultural homogenization. The cities and suburbs of Southeast Asia have become commercial centres for the sale of foreign manufactures; they symbolize a transnational culture of salaried economic prosperity, literacy, white-collar employment, and recreation. As centres of government for new, multi-ethnic nations, one of their most urgent tasks is to develop that basic vehicle of national culture, the national language. Bangkok is energetically fostering early school education in the national language (Thai) and literacy in its Indian-derived writing system. The Indonesian language, developed from the literary Malay language of Sumatra, is universally taught in Indonesian schools and has come to be widely spoken and read (in a Romanized alphabet) by younger Indonesians. Malaysia's new national language, another dialect of literary Malay written in a Romanized alphabet, is being vigorously promoted as a common medium to link the large resident Chinese population with the Malays and other indigenous peoples of Sarawak and Sabah. In Manila a version of Tagalog written in a Roman alphabet has been declared a national language and is gaining acceptance. Burmese and Cambodian, both written in Indianized scripts, are routine school subjects in Burma and Cambodia. Vietnamese, mostly written in the Romanized Quoc-ngu script, is increasingly being taught among South Vietnam's hill peoples.

New nonlinguistic skills are also spreading. Road building in Thailand, Malaysia, and the Philippines has taught local workers the use of motorized equipment ranging from bulldozers to passenger vehicles. In northern Thailand, Mindanao, and northern Borneo, the chain saw has enormously accelerated local timber-cutting operations. In Borneo this tool, together with growing access to inexpensive outboard motors, is revolutionizing the economy of many communities along the rivers. In Vietnam and

Thailand the U.S. military forces have trained thousands of Vietnamese and Thai in such skills as helicopter operation, ordnance logistics, aircraft maintenance, and the use of electronic computing and tracking systems.

Other tendencies to cultural "homogenization" include an increase in inter-ethnic marriages. In the cities of Java they even occur on a significant scale between ethnically different Christians and Muslims. Another aspect is the displacement of traditional codes of behaviour among urban immigrants: in Sabah (Malaysia), for example, rural Dusun women coming as nurse trainees to the capital city discover that their "native law and custom" does not apply to social or sexual behaviour in the city.

Movements reacting against cultural uniformity have also developed, usually as efforts to reassert traditional cultural values. Since Indonesia obtained its independence, the Javanese have experienced a great proliferation of meditative sects, or *aliran kebatinan*, each usually led by a local guru, or teacher. Many *aliran* meet regularly to examine the detailed ramifications of "Javanese science" and to repudiate materialistic secular values such as efficiency, individual success, and Western scientific attitudes. Nativistic local leaders have also been conspicuous in Borneo, Sumatra, Burma, the Philippines, and peninsular Malaysia; to rally and hold followers, they often invoke the symbols, episodes, and folk heroes of regional history. In some instances these groups have developed into armed revolutionary movements.

BIBLIOGRAPHY

General and mainland: Much of the ethnological and historical material is summarized in RALPH LINTON, *The Tree of Culture* (1955); and A.L. KROEBER, *Anthropology* (1948). The seminal work on the history of Indian influence in southeast Asia is GEORGES COEDES, *Les États hindouisés d'Indochine et d'Indonésie*, new ed. (1964; Eng. trans., *The Indianized States of Southeast Asia*, 1968). FRANK M. LEBAR, GERALD C. HICKEY, and JOHN MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia* (1964), covers recent work on mainland ethnology. More specialized works are THOMAS A. SEBEOK (ed.), *Current Trends in Modern Linguistics*, vol. 2 (1967); and MANNING NASH (ed.), *Anthropological Studies in Theravada Buddhism* (1966). Detailed studies of lowland and upland communities in Thailand, Laos, and Burma are H.K. KAUFMAN, *Bangkhuad* (1960); K.G. IZIKOWITZ, *Lamet: Hill Peasants in French Indochina* (1951); F.K. LEHMAN, *The Structure of Chin Society* (1965); MANNING NASH, *The Golden Road to Modernity* (1965); and J.E. DE YOUNG, *Village Life in Modern Thailand* (1955). Other studies of the mainland include MELFORD E. SPIRO, *Burmese Supernaturalism* (1967); H.P. PHILLIPS, *Thai Peasant Personality* (1965); and G.C. HICKEY, *Village in Vietnam* (1964), an ethnographic account of a delta village south of Saigon. Major studies of peninsula Malays are RAYMOND W. FIRTH, *Malay Fishermen*, 2nd ed. (1966); ROSEMARY FIRTH, *Housekeeping Among Malay Peasants*, 2nd ed. (1966); J.M. GULLICK, *Indigenous Political Systems of Western Malaya* (1958); M.G. SWIFT, *Malay Peasant Society in Jelebu* (1965); and JUDITH DJAMOUR, *Malay Kinship and Marriage in Singapore* (1959).

Indonesia—Java: ALICE G. DEWEY, *Peasant Marketing in Java* (1962), is a detailed study of the intricate transaction strategies of the marketplace. CLIFFORD GEERTZ, *The Religion of Java* (1960), deals with three Javanese religious variants. HILDRED GEERTZ, *The Javanese Family* (1961), analyzes Javanese socialization and personality. ROBERT R. JAY, *Javanese Villagers* (1969), details the economic and social aspects of Javanese peasant life. KOENTJARANINGRAT (ed.), *Villages in Indonesia* (1967), is a collection of 13 community studies from throughout Indonesia. LESLIE H. PALMIER, *Social Status and Power in Java* (1959), analyzes social structure in a Javanese town. SELOSOEMARDJAN, *Social Changes in Jogjakarta* (1962), offers a broad picture of central Javanese social history from the late colonial Dutch period through 1958. G.W. SKINNER (ed.), *Local, Ethnic, and National Loyalties in Village Indonesia* (1959), contains papers of five specific areas.

Indonesia—Outer: RUTH T. MCVEY (ed.), *Indonesia* (1963), is a comprehensive survey. JANE BELO, *Trance in Bali* (1960), deals with Balinese ritual. R.E. DOWNS, *The Religion of the Bare'e-Speaking Toradja of Central Celebes* (1956), summarizes and interprets earlier reports by Dutch scholars. P.R. GOETHALS, *Aspects of Local Government in a Sumbawan Village (Eastern Indonesia)* (1961), analyzes social control in a village of shifting cultivators. Other works on Indonesian peoples include CORA A. DUBOIS, *The People of Alor* (1944); CLARK E. CUNNINGHAM, *The Postwar Migration of the Toba-Bataks to East Sumatra* (1958); J.C. VERGOUWEN, *Het rechtsleven der Toba-Bataks* (1933; Eng. trans., *The Social Organization and Customary Law of the Toba-Batak of Northern Sumatra*, 1964); C. SNOUCK HURGRONJE, *The Achehnese*, 2 vol. (1906); and P.E. DE JOSSELIN DE JONG, *Minangkabau and Negri Sembilan: Socio-Political Structure in Indonesia* (1952).

Malaysian Borneo: Studies of village life in Borneo include W.R. GEDDES, *Nine Dayak Nights* (1957) and *The Land Dayaks of Sarawak* (1954); J.D. FREEMAN, *Iban Agriculture* (1955); H.S. MORRIS, *Report on Melanau Sago Producing Community in Sarawak* (1953); HEDDA MORRISON, *Sarawak* (1957), a photographic portrait; I.H.N. EVANS, *The Religion of the Tempasuk Dusuns of North Borneo* (1953); and H.L. ROTH, *The Natives of Sarawak and British North Borneo*, 2 vol. (1896, reprinted 1968).

Philippines: Studies of Christian communities include CHARLES R. KAUT, "Utang na loob: A System of Contractual Obligation Among Tagalogs," *SWest. J Anthrop.*, 17:256–272 (1961); JAMES N. ANDERSON, "Buy-and-Sell and Economic Personalism: Foundations for Philippine Entrepreneurship," *Asian Survey*, 9:641–668 (1969); FRANK LYNCH, *Social Class in a Bikol Town* (1959); H.T. LEWIS, *Ilocano Rice Farmers* (1971); W.F. and C. NYDEGGER, *Tarong: An Ilocos Barrio in the Philippines* (1966); and MARY R. HOLLNSTEINER, *The Dynamics of Power in a Philippine Municipality* (1963). RICHARD W. LIEBAN, *Cebuano Sorcery* (1967), analyzes the social and medical contexts of sorcery beliefs in the central Philippines. H.C. CONKLIN, *Hanunóo Agriculture* (1957), is a study of shifting cultivation in a pagan Mindoro community. R.F. BARTON, *Ifugao Law* (1919, reprinted 1969), *Religion of the Ifugaos* (1946) and *The Kalingas* (1941, reprinted 1969), are studies of northern Luzon pagan peoples. EDWARD P. DOZIER, *Mountain Arbiters* (1966), has provided a more recent view of Kalinga social control patterns. FELIX M. and MARIE KEESING in *The Ethnohistory of Northern Luzon* (1962), reconstruct the peopling of the northern Philippines; and in *Taming Philippine Headhunters* (1934), provide an early account of changes wrought by American administration of the Mountain Province area. A.L. KROEBER, *Peoples of the Philippines*, 2nd ed. (1928), presents an overview of the area's traditional ethnological diversity.

(P.R.Go.)

Southeast Asian Peoples, Arts of

First conceived in 1943 as a theatre of war and later generally accepted as a distinct geographical and cultural region, the term Southeast Asia refers to the huge peninsula of Indochina and the extensive archipelago of what used to be called the East Indies. The region can be subdivided into mainland Southeast Asia and insular Southeast Asia. The political units contained in this region are Burma, Thailand, Laos, Cambodia (now the Khmer Republic), North Vietnam, South Vietnam, Malaysia, Singapore, Indonesia, and the Philippines. The Philippines originally was not included because Philippine history has not followed the general historical pattern of Southeast Asia, but, because of its geographic position and the close affinities of its primitive cultures with the primitive cultures of Southeast Asia, it is now usually regarded as the eastern fringe of Southeast Asia. A common geographical and climatic pattern prevails all over Southeast Asia and has resulted in a particular pattern of settlement and cultural development. Mountain people generally have a cultural level lower than that of the valley dwellers. As a consequence, Southeast Asia is culturally fragmented.

This article is divided into the following sections:

I. The cultural setting of Southeast Asian arts
II. Literature
  General considerations
    Regional distinctions
    Prestige of the writer
  Pre-European colonial period
    Burma
    Thailand
    Laos, Cambodia, and Vietnam
    Malaysia and Indonesia
  European colonial and modern periods
    Burma
    Thailand
    Laos, Cambodia, and Vietnam
    Malaysia and Indonesia
    The Philippines

III. Music
General characteristics
Society and music
The relation of music to dance and theatre
Musical traditions and practice
Historical developments
Origins
Burma
Thailand, Laos, and Cambodia
Vietnam
Indonesia and Malaysia
The Philippines
IV. The performing arts
Diverse traditions in the performing arts
The folk tradition
The court tradition
The popular and Western traditions
Characteristics of dance
Dramatic and nondramatic forms
Styles and conventions of movement and costuming
Characteristics of drama
Thematic origins and materials
Characters
Dramatic materials
Costumes, makeup, and settings
Origins and development of the performing arts
Prehistory and links to the present
Spreading of styles
Popular theatre and Western rule
Diverse national forms and traditions
Cambodia
Thailand
Laos
Burma
Indonesia
Malaysia
Vietnam
The Philippines
Evaluation and study
V. Visual arts
General considerations
Religious–aesthetic traditions
Artistic styles
General development of Southeast Asian art
Burma
6th to 11th century
11th century to the present
Thailand and Laos
Dvaravati Mon kingdom: 6th to 11th century
Khmer conquest and Tai immigration: 11th to 13th century
The Thai kingdom: 13th to 17th century
18th century to the present
Laos
Cambodia and Vietnam
Cambodian kingdoms of Funan and Chenla: 1st to 9th century
Kingdom of Khmer: 9th to 13th century
13th century to the present
Vietnam kingdom of Champa: *c.* 2nd to 15th century
Vietnam: 2nd–19th century
19th and 20th centuries
Indonesia
Central Javanese period: 7th to 13th century
East Javanese period: 927 to 16th century
16th century to the present
Bali
Java: 20th century
The Philippines
Folk arts
The ornamental-fantastic style
Textiles

I. The cultural setting of Southeast Asian arts

Racial composition

Southeast Asia has been the crossroads of many races who have been contending against each other for centuries. At present, all the peoples of Southeast Asia are Mongoloid in racial origin. The first to come were the Austronesians (Malayo-Polynesians), sometimes described as Proto-Malays and Deutero-Malays. At one time they occupied the eastern half of mainland Southeast Asia, but later they were pushed toward the south and the islands by the Austro-Asiatics. At present, peoples of Austronesian origin occupy Malaysia, the Republic of Indonesia, and the Republic of the Philippines. There were three main Austro-Asiatic races, the Mons, the Khmers, and the Viet-Muong. The Mons were at one time dominant, but they lost their racial identity in the 18th century and became absorbed by the Burmese and the Tais; only a few thousand Mons are now found living near the Burma–Thailand border. The Khmers from the 9th century to the 15th built a great empire, but much of its territory was lost to its neighbours so that only the small kingdom of Cambodia remains today. The Viet-Muongs now occupy South and North Vietnam. A Tibeto-Burmese tribe, the Pyu, founded an empire of city-kingdoms in the Irrawaddy Valley in the early centuries of the Christian Era, but the Pyus disappeared, and the Burmese, taking the leadership, founded their kingdom of Pagan and have occupied Burma up to the present day. In the 13th century the Tai-Shans lost their kingdom of Nanchao in Yunnan, China, and entered the Mae Nam Chao Phraya Valley to found kingdoms that gradually evolved into the kingdoms of Siam (Thailand) and Laos.

The Indian influence

In Southeast Asia winds of change often came as storms. Indian commerce expanded into Southeast Asia in the early centuries of the Christian Era and, in spite of its peaceful nature, caused revolutionary changes in the life and culture of the peoples of the region. The Indians would sojourn in the region in small numbers for two or three monsoons only. The success of their commercial venture and the safety of their persons depended entirely on the goodwill of the inhabitants. At that time, the Indians were at a higher level of culture, and they brought new ideas and new art traditions. Since these ideas had some affinity with indigenous ideas and art forms, the natives, who had already reached a high level of culture themselves, accepted them but were not overwhelmed by an influx of new traditions. The Hindu–Buddhist culture of the Indians made a tremendous impact and came to form the second layer of culture in Southeast Asia, but the first layer of native ideas and traditions has remained strong to the present day.

European influence

Changes often came to Southeast Asia, usually because it possessed a commodity that was in great demand by the rest of the world. The Indians came because they were looking for fresh sources of gold after the Roman imperial source had run dry. In the 15th, 16th, and the 17th centuries, insular Southeast Asia attracted Islāmic merchants from India and farther west and later the Portuguese and the Dutch as a rich source of spices. As with the Hindu–Buddhist merchants of the past, the Islāmic traders came not as missionaries, though they did spread their religion in the region. The Portuguese came as conquerors and as militant missionaries of their Roman Catholic form of Christianity, and, for those reasons, their cultural traditions were found unacceptable to the natives. In the 17th century the Dutch came as conquerors and colonists for whom the attraction was first spices and then coffee, rubber, and petroleum. Since mainland Southeast Asia produced no spices for export, it was less vulnerable to the navies of Portugal and the Netherlands, so the region was not greatly affected by the Muslims, Portuguese, and Dutch. In the 19th century, Britain and France became interested in mainland Southeast Asia as the back door to China and sought to possess it as a colony. By the end of the 19th century, Burma had fallen to Britain, Siam was allowed to retain its independence only with the tacit permission of the two powers, and the rest had fallen to France. When, in the middle of the 20th century, the whole of Southeast Asia became free again, it was found that the arts of the peoples had remained merely static during the years of colonial rule; in some regions they had been driven underground by long neglect, but European culture and European art forms had made only a small impact.

The native heritage

The peoples of Southeast Asia were once thought to have shared a lack of inventiveness since prehistoric times and to have been "receptive" rather than "creative" in their contacts with foreign civilizations. Later excavations and discoveries in Burma and Thailand, however, inspired some scholars to argue against the accepted theory that civilization came to Southeast Asia from China in prehistoric times; rather, these scholars contend, the

peoples of mainland Southeast Asia were cultivating plants, making pottery, and working in bronze about the same time as the peoples of the ancient Near East, and therefore civilization spread from mainland Southeast Asia to China and India. Southeast Asians have never produced any theory of art or literary or dramatic criticism, for they are always more concerned with doing the actual work of producing beautiful things. Because the Southeast Asians, especially in the western half of the mainland, worked on nondurable materials, it is not possible to trace the development and evolution of art forms stage by stage. The region has always been thickly forested, so it was natural that the first material to be used for artistic purposes should have been wood. The wood-carving tradition, begun in primitive times, was retained even when they learned to work with metals and with stone and continued to flourish long after the great age of stone sculpture and stone architecture, which ended in the 13th century. Proto-Neolithic paintings discovered in a cave near the Salween River in the western Shan state of Burma have very close affinity with the later carvings on posts of houses among the Nāgas on the western hills of Burma. Similarly, cave paintings of a pair of human hands with open palms, one holding the sun and the other holding a human skull, are reflected in the later aesthetic tradition of Southeast Asia: the sun symbol is found as an art motif all over the region, and a suggestion of awe, triumph, and joy at acquiring a human head is found in carvings under the eaves of the Nāga houses. The cave painting testifies to the continuity of the magico-religious tradition connected with all the arts of the area.

The art of casting the bronze drums found at Dong Son, near Hanoi, which are similar to the bronze drums used by mountain tribes throughout Southeast Asia, was thought to have come from China, but recent excavations in Thailand proved that the drums and the so-called Dong Son culture itself are native to mainland Southeast Asia. In any case, the continuity of the aesthetic tradition of Southeast Asia can be seen in the bronze drums that were cast by the Karens of Burma for centuries until the early years of the 20th century. The mountains of mainland Southeast Asia provided gold, silver, and other metals, and the art of metalworking must have developed quite early. Silver buttons, belts, and ornaments now made and worn by the hill peoples in Southeast Asia have behind them a very ancient tradition of workmanship. The same artistic tradition is found in textile designs.

Music, dance, and song were originally associated with tribal rituals. From the beginning, the main characteristic of Southeast Asian music and dance has been a swift rhythm. The slow and stately dances of the Siamese court were of Indian origin; when they were introduced into Burma in the 16th century, the Burmese quickened the tempo, but, even with that modification, the dances were still called Siamese dances to distinguish them from the native ones. In their oral literature—namely, in folk songs and folktales—the emphasis is on gaiety and humour. Typically, Southeast Asians do not like an unhappy ending.

The role of kings and gods

In all the regions of Southeast Asia, the arts flourished under the patronage of the kings. About the time of the birth of Christ, tribal groups gradually organized themselves, after some years of settled life as rice cultivators, into city-kingdoms, or conglomerations of villages. A king was thus little more than a paramount tribal chieftain. Since the tribes had been accustomed to worshipping local spirits, the kings sought a new spirit that would be worshipped by the whole community. One reason why the gods of Hinduism and Buddhism were found so readily acceptable to Southeast Asia was this need for new national gods. The propagation of the new religions was the task of the kings, and consequently the period from the 1st to the 13th century was a great age of temple building all over Southeast Asia. Architecture, sculpture, and painting on the temple walls were the arts that flourished. In the ancient empires of eastern Indochina and the islands, scholars of Sanskrit, the language of the sacred works of Hinduism, became part of the king's court, producing a local Sanskrit literature of their own. This literary activity was confined to the hereditary nobility and never reached the people, except in stories from the great Hindu epics *Mahābhārata* and *Rāmāyaṇa*. Because the Hindu religious writings in Sanskrit were beyond the reach of the common people, Hinduism had to be explained to them by Hindu stories of gods and demons and mighty men. On the other side of the peninsula, in the Pyu-Burmese empire of Prome, which flourished before the 8th century, there was no such development—first, because Hinduism was never widely accepted in Burma and, second, because the more open Burmese society developed neither the institution of a god-king nor that of a hereditary nobility. Although Pāli scholars surrounded the king in later Pagan, Pāli studies were pursued not at the court but at monasteries throughout the kingdom so that even the humblest villager had some faint contact with Pāli teachings. While the courts of the kings in Cambodia and Java remained merely local centres of Sanskrit scholarship, Pagan became a centre of Pāli learning for Buddhist monks and scholars even from other lands. As in the case of stories from the Indian epics, stories of the *Jātaka*s (birth stories of the Buddha) were used to explain Buddhism to the common people, who could not read the scriptures written in Pāli. Just as scenes from the great epics in carving or in fresco adorned the temples in Cambodia and Java, scenes from the *Jātaka*s adorned the Pagan temples.

Sanskrit and Pāli literature

Musicians of the Pyu kingdom played before the Emperor of China in 801, and the various musical instruments at the performance have their counterparts at the present day, not only in Burma but throughout Southeast Asia. At Pagan the people were so fond of music that even the collection of taxes became an occasion to dance and sing, and a royal official, endowing a temple, inscribed a prayer asking that in all his future existences until he reached Nirvāṇa "might he be woken up every morning to the strains of music sweetly played on flute and violin." In spite of this love for music and dance, no dramatic art seems to have developed in Burma, perhaps because Sanskrit, in which there was a dramatic tradition, was not studied. In contrast, at the courts of Cambodia and Java, the Sanskrit drama, Hindu dances, and native dance traditions combined and produced the court opera ballets. These dramatic elements later reached the common people by way of the shadow play.

Pervasive love of music

The patronage of the king and the religious enthusiasm of the common people could not have produced the great temples without the enormous wealth that suddenly became available in the region following the commercial expansion. With the Khmer and Javanese empires, the wealth was produced by a feudalistic society, and so the temples were built by the riches of the king and his nobles, combined with the compulsory labour of their peasants and slaves, who probably derived some aesthetic pleasure from their work because of their religious fervour. Nonetheless, their monuments, such as Borobuḍur, in Java, and Angkor Wat, in Cambodia, had an atmosphere of massive, all-conquering power. At Pagan, where wealth was shared by the king, the royal officials, and the common people, the temples and the monasteries were built by all who had enough not only to pay the artisans their wages but also to guarantee their good health, comfort, and safety during the actual construction. The temples were dedicated for use by all monks and lay people as places of worship, meditation, and study, and the kings of Pagan did not build a single tomb for themselves. The Khmer temple of Angkor Wat and the Indonesian temple of Borobuḍur were tombs in that the ashes of the builders would be enshrined therein; the kings left stone statues representing them as gods for posterity to worship, whereas at Pagan there was only one statue of a king, and it depicted him on his knees with his hands raised in supplication to the Buddha. Consequently, the atmosphere that pervaded the temples of Pagan was one of joy and tranquillity. This golden age of wealth and splendour in Southeast Asia ended in the 13th century with a sudden violence, when Kublai Khan's Tatar Chinese armies destroyed both the Burmese and the Khmer empires and his navy attacked Vietnam and Java.

Significance of the temples

The tiny kingdoms that subsequently sprang up all over Southeast Asia continually fought among themselves; their kings were neither powerful nor rich, and the royal courts became centres of military planning and political intrigue. During the 13th and 14th centuries, in the new Javanese kingdom of Majapahit and the new Burmese kingdom of Ava, vernacular literatures came into being. Again, differences in social structure had aesthetic repercussions. In Majapahit the king was powerful and gave his patronage to the newly arisen literature, confining it to the court. At Ava the vernacular literature bloomed throughout the kingdom, and the king, lacking power and prestige, prevailed upon some established writers to join the court circles and give them glamour.

After Majapahit, a new cultural force—namely, Islām—reached insular Southeast Asia, and over the two layers of primitive and Hindu–Buddhist cultures was added the third layer of Islām. In mainland Southeast Asia, a new Burmese empire arose over the ruins of the old and continued its task of spreading Buddhism. Hindu tradition reached the Burmese court secondhand in the 18th century as the result of the Burmese conquest of Siam and was one of the factors that contributed to the rise of a Burmese drama. On the other side of the peninsula, Vietnam, reconquered by China, fell more and more under the influence of Chinese culture. After a short period of Islāmic bloom, native culture in insular Southeast Asia was subjected to alien rule. In Burma and Siam alone among the states of Southeast Asia, native arts continued to flourish because, after centuries of warfare, they finally emerged as strong kingdoms.

Predominant artistic themes

The predominant themes of Southeast Asian arts have been religion and national history. In religion the main interest was not so much in actual doctrine but in the life and personality of the Buddha and the personalities and lives of the Hindu gods. In national history the interest was in the legendary heroes of the past, and this theme appeared only after the great empires had fallen and the memories of their glory and power remained. The Buddha image, which went through various stages of development, remained the favourite motif of sculpture and painting. The depiction of scenes from his previous lives in fresco and relief sculpture also had the purpose of teaching the Buddhist ethics to the people, as the *Jātaka*s emphasized certain moral virtues of the Buddha in his previous lives; it also gave an opportunity to the artist to introduce local colour by using, as background, scenes from his own contemporary time. The depiction of scenes from the Hindu epics also had the same purpose and gave the same opportunity to the artist. Many figures from the Buddhist and Hindu scriptures, such as gods and goddesses, heroes and princesses, hermits and magicians, demons and dragons, flying horses and winged maidens, became fused with similar native figures, and, gradually, folklore plots became merged in the general religious themes. The *nāga*, a superhuman spirit, was taken from Buddhist and Hindu texts and merged with native counterparts, with the result that different images of the *nāga* appeared in various regions. The Burmese *nāga* was a snake with a crested head. The Mon *nāga* was a crocodile, and the Khmer and Indonesian *nāga* was conceived as a nine-headed snake. The demons of various kinds from all over Southeast Asia became merged under one name of Pāli-Sanskrit origin, *yakkha* or *yakṣa*, but they retained their separate identities in sculpture and paintings of their own different countries. The lion, which was unknown to the monsoon forest but was a figure of Hindu and Buddhist mythology, evolved into a native symbol and art motif. The primitive worship of the snake-dragon as a god of fertility was retained in the Khmer Empire; the nine-headed *nāga* became a symbol of security and of royalty, and stone *nāga*s guarded the palaces and temples. Buddhism frowned upon *nāga* worship. In Burmese and Mon sculpture the *nāga* was always shown as a servant of the Buddha, putting his body in coils to make a seat for his master and raising his great hood as an umbrella over his master's head. According to tradition, the guardian figure of a Mon temple was a two-bodied lion with a man's head, and the guardian figure of a Burmese temple was the crested lion. The Tais made themselves heirs to both the Khmer and the Mon art traditions relating to the *nāga*, but the guardian figure of their temples was the benevolent demon. Primitive symbols and animal imagery merged with Indian animals and entered the arts. The Pyus embossed the primitive symbol of the sun on their coins as insignia of their power, and the Burmese transformed it into their favourite bird, the peacock, on the excuse that Buddhist mythology associated the peacock with the sun; the Mons adopted the red sheldrake as their symbol, and in Indonesia the mythical bird called Garuḍa, the vehicle of Viṣṇu (Vishnu), became merged with the local eagle. The figures of these birds also became decorative motifs. Animals of the Southeast Asian forests whose figures had adorned primitive dwellings of wood and thatch were stylized and came to adorn palaces and monasteries. Primitive geometrical patterns mixed with new spirals and curves from India, and Indian floral designs merged with those of trees and fruits and flowers copied from the monsoon forests.

The *nāga*s

Relationship to other cultures

The arts of Southeast Asia have no affinity with the arts of other areas, except India. Burma was always an important route to China, but Burmese arts showed very little Chinese influence. The Tais, coming late into Southeast Asia, brought with them some Chinese artistic traditions, but they soon shed them in favour of the Khmer and Mon traditions, and the only indications of their earlier contact with Chinese arts were in the style of their temples, especially the tapering roof, and in their lacquer ware. Vietnam was a province of China for 1,000 years, and its arts were Chinese. The Hindu archaeological remains in South Vietnam belong to the ancient kingdom of Champa, which Vietnam conquered in the 15th century. The Buddhist statues in North Vietnam were Chinese Buddhist in style. The essential differences in aesthetic aim and style between the arts of East Asia and those of Southeast Asia could be seen in the contrast between the emperors' tombs of Vietnam and the temple-tombs of Cambodia and Indonesia or the opulent and dignified Buddha images of Vietnam and the ascetic and graceful Buddha images of Cambodia and Burma. Islāmic art, with its rejection of animal and human figures and its striving to express the reality behind the false beauty of the mundane world, also has no affinity with Southeast Asian arts. Both Hinduism and Buddhism taught that the sensual world was false and transitory, but this message found no place in the arts of Southeast Asia. The world depicted in Southeast Asian arts was a mixture of realism and fantasy, and the all-pervading atmosphere was a joyous acceptance of life. It has been pointed out that Khmer and Indonesian classical arts were concerned with depicting the life of the gods, but to the Southeast Asian mind the life of the gods was the life of the peoples themselves—joyous, earthy, yet divine. The European theory of "art for art's sake" found no echo in Southeast Asian arts, nor did the European division into secular and religious arts. The figures tattooed on a Burmese man's thigh were the same figures that adorned a great temple and decorated a lacquer tray. Unlike the European artist, the Southeast Asian did not need models, for he did not strive to be realistic and correct in every anatomical detail. This intrusion of fantasy and this insistence on the joyousness of human life have made Southeast Asian arts unique.

II. Literature

GENERAL CONSIDERATIONS

Regional distinctions. From the point of view of its "classical" literatures, Southeast Asia can be divided into three major regions: (1) the Sanskrit region of Cambodia and Indonesia; (2) the region of Burma where Pāli, a dialect related to Sanskrit, was used as a literary and religious language; and (3) the Chinese region of Vietnam.

There are no examples of Chinese literature written in Vietnam while it was under Chinese rule (111 BC–AD 939); there are only scattered examples of Sanskrit inscriptions written in Cambodia and Indonesia; yet most of the literary works produced at the court of Pagan in

Burma (flourished *c.* 1049–1300) have survived because the texts were copied and recopied by monks and students. But in the 14th–15th centuries, vernacular literatures suddenly emerged in Burma and Java, and a "national" literature appeared in Vietnam. The reasons behind the development of each were the same: a feeling of nationalistic pride at the final defeat of Kublai Khan's invasions, the desire of the people to find solace in literature amidst change and struggles for power, and the lack of wealth and patronage to channel artistic expression into building temples and tombs. In Vietnam and Java literary activity centred on the courts; but in Burma the first writers were the monks and, later, the laymen educated in their monasteries. In the new Burmese kingdom of Ava (flourished after 1364), the Shan kings were proud of their Burmese Buddhist culture, and they appointed the new writers into royal service, with the result that courtiers became writers also. The Tai kings of Laos and Siam led their courts in learning Pāli from the Mons, whom they had conquered, and Sanskrit from the Khmers, whom they harassed; nevertheless, seized with national pride and influenced by the Burmese example, they developed their own vernacular literature. But Cambodia itself declined. Although the monks in the Theravāda Buddhist (*i.e.*, the Southeast Asian form of Buddhism) monasteries produced a few works in Pāli, no vernacular literature emerged until finally Khmer-speaking people (those living in the area comprised approximately of modern Cambodia) were borrowing many words from the Tais.

For its vernacular literatures, Southeast Asia can be divided into (1) Burma; (2) Thailand, Laos, and Cambodia; (3) Vietnam; (4) Malaysia and Indonesia; and (5) the Philippines (which produced a vernacular literature only in the 20th century, after the imposed Spanish and English languages and literatures had made their impact).

The era of royal patronage

Prestige of the writer. During the time of the kings, a Southeast Asian writer enjoyed patronage and a prestigious position in society. He could not, however, make a living by writing as a profession. Manuscripts had to be written by hand, and only in the case of famous works might one or two duplicates be made, again by hand. There was no question of selling the manuscript. A writer could only hope to attract the notice of his king and obtain a monetary reward or a royal office. By the time that printing presses were introduced, in the colonial period in the 19th century, the kings were gone and with them their writers. Colonial rule overwhelmed and destroyed vernacular literary traditions, leaving intact only oral literatures in the forms of folktales and folk songs. Literary criticism, as understood in Western cultures, had never been known, either in the ancient or modern literatures of Southeast Asia. Apart from a few stray writings on versification, therefore, no works of literary criticism or literary history existed until the colonial period. Even then, the interest of European scholars was chiefly confined to archaeology, and only a few made the attempt to study some special type or period of a vernacular literature (for example, vernacular versions of the *Rāmāyaṇa*, the great Sanskrit epic of India, or of 14th-century Javanese verse). There is a work in French dealing with Thai literature and a work in Burmese dealing with Burmese literature; but apart from these no study of any Southeast Asian literature as a whole has yet been made. For this neglect, native scholars are much to blame. Works on literary criticism and the history of literature would help give perspective to indigenous writers in the 20th century, both to those who want to cling to their native traditions (as is the case with writers in Burma and Thailand) and those who want to make a complete break with the past (as is generally the case in Vietnam, Indonesia, and Malaysia).

PRE-EUROPEAN COLONIAL PERIOD

Burma. The Burmese borrowed many words from Pāli but not to the extent that the Indonesians, Khmers, and the Thais borrowed Sanskrit words. The Burmese language was monosyllabic and tonal, and since there was no accent or stress, the feature that distinguished verse from prose was the regular occurrence of rhyme. They modelled their literature not on classic examples from Pāli or Sanskrit but on their own traditional folk songs.

The 15th century. In the 15th century, four types of verse existed: (1) *pyo* (religious verse), which retold stories of Buddha's birth and teaching and were taken from the *Jātaka*s (a collection of folktales adapted to Buddhist purposes and incorporated into the Pāli canon), to which were added imaginative details and a Burmese background; (2) *linkar* (shorter religious verse), or a devotional poem, characterized by a metaphysical flavour comparable in many ways to that which informs the work of the early 17th-century English poets George Herbert and Robert Herrick; (3) *mawgoon* (historical verse), half ode, half epic, written in praise of a king or prince and developing out of military marching songs; (4) *ayegyin* (lullaby), an informative poem usually addressed to a young prince or princess and written in praise of his royal ancestors.

Literature in the 15th century is dominated by three monks: Shin Maha Rahta Thara, who wrote for the court of Ava, and Shin Maha Thila Wuntha and Shin Uttamagyaw, both of whom were of village stock and did not go to court but remained on in their village monasteries. Shin Maha Thila Wuntha, in the closing years of his life, turned to prose and wrote a chronicle history of Buddhism. In this period several courtiers, both men and women, also began to achieve some literary success, and the genre called *myittaza* (epistle) first evolved, which is a long prose letter written by a monk and addressed to the king to advise him of his duties.

Poets at court

The 16th century. In the 16th century, the Burmese conquered Siam, and their subsequent knowledge of Thai romantic poems gave rise to a new verse form called the *yadu* (the seasons). They borrowed only the theme, however, and not the form, and they developed it as an emotional poem, passionate, yet with something of the cool intellectual strength of the poems of the English metaphysical poets John Donne and Andrew Marvell. The most famous writers of the *yadu* were two court poets, Phyu and Nyo; a general of the army called Nawaday; and Natshinnaung, king of Toungoo. The wide popularity of the poems eventually gave rise to a mock-heroic form called *yagan* ("Kick the *yadu*").

Golden age of literature. In the early years of the 18th century, U Kala compiled a history of Burma, written in precise and clear prose; the closing years, which coincided with the establishment of the third Burmese Empire, saw a great period of literature. The Thai court, brought as captives to the Burmese capital, introduced to the Burmese poetic romances and their *Rāma* play (based on the *Rāmāyaṇa*). Contact with the Thais stimulated the growth of a Burmese court drama and led to the appearance of Burmese court romances in poetic prose. The king's treasurer, however, made fun of the Thai importations and wrote the *Rama Yagan*, in which the high romance and courtly elegance of the 4th-century-BC *Rāmāyaṇa* ("The Life of Rāma") were given a rustic setting, with hilarious results. From the quiet of their monasteries, the monk Awbatha wrote novel-like rendering of the *Ten Long Jātakas* and the monk Kyeegan Shingyi wrote homely, pithy, and sometimes even humorous myittaza ("epistles") from villagers to their relations in the cities.

The defeat suffered by the Burmese in the Anglo-Burmese War of 1824–26—their first defeat since the time of Kublai Khan in the 13th century—introduced a note of melancholy to Burmese literature. During this first half of the 19th century, many of the new melancholy lyrics were set to music. Two great writers were a product of this period: the dramatist U Kyin U and the courtier U Pon Nya, the greatest writer of the time, whose plays, epistles, and songs are full of humour and zest for life.

Thailand. Until 1824 Thai literature was entirely the province of the king and his court: the king maintained a corps of writers, and it was the custom to attribute authorship of any literary work to the king himself. Thai

vernacular literature began with verse, based on Sanskrit models but relying on an elaborate rhyme scheme because the Siamese language was tonal. The two earliest known poems were *Yoon Pai* ("The Defeat of the Yoons"), an epic-ode having similarities to the Burmese *mawgoon* genre, and *Mahajati* ("The Great *Jātaka*"), a poem stressing ethical ideas, similar in form to the Burmese *pyo*. Both poems, written during the period 1475–85, give ample proof that Thai writers, using Sanskrit, Khmer, and Burmese models, could nonetheless produce a truly Thai work.

Elaborate rhyme scheme of Thai poetry

First golden age: King Narai (1657–88). All literary activity ceased in the 16th century because of the unsettled conditions that prevailed before and after the annexation of the country by the Burmese. Independence was regained toward the close of the century, and under King Narai (1657–88), at his court in Ayutthaya, Siamese literature achieved its first golden age. Narai was himself a great poet, and during his reign new verse forms were evolved. He wrote poetic romances, based on stories from the "Fifty Jātakas," which were in fact folktales belonging to the region retold in Pali and disguised as *Jātakas* by an unknown Tai monk. Narai also wrote the final version of the poem of tragic romance, *Pra Lo* ("Lord Lo"), which had first been composed by an anonymous author in a much earlier reign. Among courtier poets of this time, the most famous were Maharajaguru; Si Prat, a wild young gallant who wrote the romantic poem *Aniruddha* (the name of the hero of the poem) and some passionate love songs; Khun Devakavi, author of cradle-songs using many Sanskrit and Khmer words but modelled on the Burmese *ayegyin;* and Si Mahosot, the author of an ode-epic in praise of King Narai. A new genre, the travel poem, also became popular; and the first versions of the plays *Rāma* and *Inao* (based on Hindu–Khmer–Javanese models) were composed by the King and his corps of writers. Perhaps the only prose work of the period was the *History of Ayutthaya* by Luang Prasroeth, which was lost and came to light only in the 20th century. It showed some signs of being influenced by U Kala's *History* (of Burma).

Second golden age: King Rama II (1809–24). Siam was conquered by the Burmese in 1767, and a new dynasty was established in a new capital, Bangkok. Some effort was made to revive the country's culture, largely destroyed following the sack of the old capital of Ayutthaya; and under the poet-king Rama II a second golden age of Thai literature occurred, during which women achieved prominence as poets for the first time. The King, with his writers, composed the final versions of *Rāma* and *Inao* and also a popular romance, "Khun Chang and Khun Pen," based on an incident in Thai history. The most famous poets were Prince Paramanuchit, whose ode-epic *Taleng Phai* ("The Defeat of the Mons") testified to his greatness, and Sunthon Phu, the King's private secretary, who was born of humble parents but made his way in the court by the excellence of his poetry. A strongly religious king, Rama III disbanded the corps of writers and discouraged the performance of plays at his court. Sunthon Phu lost his position but wrote his most famous poem, *Phra Aphaimani*, away from the court. A long fantasy-romance, this work can be regarded as the end of court domination in literature. Further, a royal official composed a Thai translation in prose (*Sam Kok*) of the Chinese classic *Romance of the Three Kingdoms.* The author, Pra Klang, was admittedly a royal official; nevertheless, the work was meant for the people rather than the court. It was followed by a spate of imitations and finally resulted in the development of the historical novel.

Final version of *Rāma*

Laos, Cambodia, and Vietnam. Laotian literature was in many respects a dialect branch of Tai literature, and, as in Thailand, it was the creation of the royal court. A number of popular romantic poems and prose lives of famous monks were composed, but their authors were unknown: all works, in fact, were by custom written anonymously.

The kings of Cambodia, fallen from high estate and often mere vassals of Thailand, could not inspire the rise of a vernacular literature. Only in the monasteries was there any literary activity, and this was written in the Pāli language.

In Vietnam, the emperors of the Tran dynasty (13th–14th century) were themselves poets and patronized a new literature—which, nevertheless, was still written in Chinese and was therefore national rather than vernacular. The writings themselves, however, were by no means a mere branch of Chinese literature. The country was afterward conquered once more by China and it was not until it regained independence that, under the patronage of the Le dynasty emperors (15th–16th century), a new age of literature began. Although the Chinese language was still used, some writers were beginning to use the vernacular (employing Chu-nom script, consisting of modified Chinese characters). Nguyen Trai, Emperor Le Thanh Tong, and Nguyen Binh Khiem were the great poets of this period. In 1651 Father Alexandre de Rhodes, a Roman Catholic missionary priest, invented a new romanized script (Quoc-ngu) that became the national script. Literature then began to reach the common people.

The national script

Literary works written before the end of the 18th century have not survived; the best known are those written in the 19th century, before the country became a French colony in 1862. Ho Xuan Huong, Nguyen Cong Tru, Chu Manh Trinh, and Tran Ke Xuong were famous court poets. Nguyen Du (1765–1820) wrote moral tales in verse that appealed not only to the court but to the common people. His most famous work was *Kim Van Kieu*, a poem of 3,253 lines, showing a strong Chinese influence (the plot was taken from a Chinese historical novel, and its ethical basis was both Confucian and Chinese Buddhist). The plays of the period, although written in Vietnamese, followed Chinese dramatic traditions because the Vietnamese theatre was still Chinese in style and practice.

Malaysia and Indonesia. Malaysia and Indonesia together have about 300 different languages and dialects, but they have a single common linguistic ancestor. Before the coming of Islām to the region in the 14th century, Javanese had been the language of culture; afterward, during the Islāmic period, Malay became the most important language—and still more so under later Dutch colonial rule so that, logically, it was recognized in 1949 as the official Indonesian language by the newly independent Republic of Indonesia.

During the period of Indian cultural influence, Sanskrit flourished in the great empires that included both the Malay Peninsula and the islands of present-day Indonesia. In the 11th century, at the court of Emperor Airlangga, a national literature (as distinct from a vernacular literature) emerged. It was written in courtly Javanese mixed with Sanskrit words, and it used Sanskrit metres and poetic style. In the 14th century in Majapahit (the new Javanese Empire that had been established after the final defeat of Kublai Khan's forces) a vernacular literature based on the speech of the common people came into being. The most important work of this new literature was *Nāgarakertāgama* (1365), a long poem in praise of the king (though it was not a product of the court) that also contained descriptions of the life of the Javanese people at the time. Although it employed a number of Sanskrit words, the style and metre were Javanese, not Sanskrit.

The Indian Hindu epics had already been popularized in the Malay Peninsula and in the islands of Indonesia (by way of the shadow-puppet play), and in this period fresh versions began to be written in the new Javanese. Romances, called *hikayat*, both in verse and in prose, also appeared—having as their source native myth and legend. Soon Malay, Balinese, Sundanese, and Madurese vernacular literatures emerged, all dealing with the same themes.

Romance themes from myth and legend

The coming of Islām coincided with the rise of Malacca and the decay of Majapahit; but the popular fantasy-romances were able to survive by adopting a Muslim, instead of a Hindu, guise. New romances, telling the stories of heroes known to Islām, such as Alexander the Great, Amīr Hamzah, and Muḥammad ibn al-Ḥanafiah, were

added to their number, and translations of Persian Muslim stories and of works on Muslim law, ethics, and mysticism further enriched Malay literature.

The finest work of all in the Malay language was the *Malay Annals*, written in about the 15th century. It gave a romanticized account of the history of the kingdom of Malacca and a vivid picture of life in the kingdom. Although a court record that begins with ancestral myths, it goes on to describe latter-day events of the kingdom with realism and humour.

In the Malay Peninsula, the coming of colonial rule did not at once overwhelm the existing native literature. As at the courts of the sultans of the British federated Malay states, the old traditions continued for some time. In Indonesia, however, a complete break was made with the cultural tradition, and by the end of the 19th century young Indonesians were dazzled by Dutch literature. A 20-year-old girl, Raden Adjeng Kartini, wrote in Dutch a remarkable series of letters that were later collected and published. By about 1920, however, the Dutch government itself had decided against encouraging further a national literature in Dutch, which the nationalist leaders were also anxious to break away from. This joint desire for a vernacular literature bore fruit in about 1933, when novels and poems in Malay (Indonesian) became popular. The new tradition developed after independence, and its outstanding writers were, in poetry, Chairil Anwar and Sitor Situmorang. Important novelists include Ananta Toer and Takdir Alisjahbana.

EUROPEAN COLONIAL AND MODERN PERIODS

Impact of colonial rule

The entire region of Southeast Asia, with the single exception of Thailand, fell under colonial rule, and Thailand itself survived more as a buffer state than as a truly independent kingdom. At the courts of the kings of Laos, Cambodia, and Vietnam, which fell under French suzerainty, and in the palaces of the sultans of the British Malay states, vernacular literatures managed to survive for a time; but since these literatures had long ago ceased to develop—as a result of harassment by the Thais in the case of Laos and Cambodia, by the Portuguese in the case of Malaya, and by the French in the case of Vietnam—they soon became moribund. In all of Southeast Asia, except Burma and Thailand, the vernacular languages themselves lost their status, as the languages of the colonial rulers became the languages of administration and of a new elite. A revival of interest in the native languages and literatures occurred only toward the close of the colonial period, as a consequence of national movements for freedom.

Burma. In Burma, unlike India and other parts of the British Empire, English did not fully replace Burmese as the language of administration. In the almost classless Burmese society the language of the court and of literature was also the language of the people, which prompted the British government to retain Burmese as a second official language and to make both languages compulsory for study in schools and colleges. As a result, no English-speaking elite emerged, English literature did not dazzle native scholars, and, although its growth was retarded, Burmese literature did not disappear. With the intensification of the movement for freedom, around 1920, political tracts, novels, short stories, and poems reflected a political bias against colonial rule. In 1930, at the University of Rangoon, a group of young writers developed a new style of Burmese prose and poetry, a style little influenced by Western literature. In the post-independence period, novels and poems became centred on biographical and historical writings.

Thailand. Administrative and educational reforms introduced by King Mongkut (1851-68) as an answer to the threat of colonial conquest created a liberal atmosphere and a new reading public, and soon many of the old courtly writings were popularized in the form of romantic prose fiction. About 1914, King Rama IV, a graduate of Oxford University, attempted to win back for the palace the leadership in literature; although he produced some fine adaptations of Shakespeare's plays, they made no impact on the people, with whom romantic fiction remained popular. Because of increased contact with the West, after World War II novels and short stories based on western models began to rival the earlier prose romances.

Laos, Cambodia, and Vietnam. Because of France's restrictive colonial educational policy, French language and literature never reached the common people. Moreover, the French-speaking elite, engrossed in French literature, neglected the native literature. With the growing vehemence of the freedom movement in the 1930s, however, there developed in Vietnam a new school of vernacular poetry that was less traditional and more nationalistic. But in the turbulent years that followed, the poets, including Ho Chi Minh himself, became occupied more with war than with literature.

Malaysia and Indonesia. The first Malaysian newspaper in the vernacular language, which appeared in 1876, introduced a new style of prose, less literary and nearer to the spoken Malay. Becoming immensely popular, the new style was further developed by other newspapers. (Although the early innovators were influenced by the English language, their followers were influenced by Arabic.) In about 1920 this new "Malaysian Malay" finally replaced the old literary Malay. The Translation Bureau, established by the British government in 1926, translated a great number of English books into the new Malay. In Indonesia, also, the old cultural language, literary Javanese, ceased to be used; by the end of the 19th century young Indonesians, overwhelmed by Dutch literature, started to write in Dutch. For example, a young girl, Raden Adjeng Kartini, wrote in Dutch a remarkable series of letters, containing criticism of Indonesian society, that were later collected and published; and a group of young men wrote poems in Dutch, although with an Indonesian background. By about 1920, however, the Dutch government itself had decided for political reasons to discourage further development of a national literature in Dutch, and the nationalist leaders had become eager for a new literature in the native language. This common aim bore fruit in 1933, when a literary journal under the editorship of Takdir Alisjahbana appeared, containing poems and essays written by various authors in the new Malay, which they now called Indonesian. The editor himself later wrote in Indonesian a number of popular novels containing social criticism, which were imitated by other writers. During the Japanese occupation of Indonesia and Malaya, this new Indonesian literature became popular also in Malaya. The adoption of Bahasa Malay (Indonesian) as the official language of Indonesia in 1949 gave further impetus to the development of the vernacular literature in both countries. A group of young Indonesian writers led by Ghairil Anwar produced a number of poems in which the theme of nationalism was subordinated to a new spirit of tolerance and universalism.

Malaysian Malay

The Philippines. Philippine literature had its beginnings in great epics that were handed down orally from generation to generation and sung on festive occasions. When the Philippines became part of the Spanish Empire in the 16th century, printing was introduced, and all the early published works in the vernacular (Tagalog) were of Christian religious subjects. Eventually, some individual romantic legends taken from the epics were published, but they had acquired a European flavour. An outstanding work in the early years of the 19th century was an epic romance called *Florante at Laura* by the first native writer to achieve prominence—Francisco Balagtas—who wrote in Tagalog. In the latter half of the 19th century, an intellectual renaissance coincided with the beginnings of a national movement toward freedom; writers began using Spanish, for their work was part of the nationalist propaganda. The most famous author was José Rizal, who wrote a series of brilliant social novels, beginning with *Noli me tangere* ("Touch Me Not"). Other prominent writers, all essayists, were Mariano Ponce and Rafael Palma. There were poets also—for example, José Palma, whose poem "Filipinas" was later adopted as the national anthem. After the United States had taken over the Philippines, Spanish was gradually replaced by En-

19th-century intellectual renaissance

glish, and new writers began to use that language as their medium. But before a new national literature could evolve, World War II took a heavy toll of writers, and those who survived became caught up in the political changes that followed. Many still write in English—the Spanish tradition, too, remains strong—but more and more writers are turning to Tagalog for literary expression. (M.H.Au.)

III. Music

GENERAL CHARACTERISTICS

Society and music. *Rural and urban music.* A general musical division exists between the urban and rural areas of Southeast Asia. Urban centres comprise the islands of Java and Bali and places in Thailand, Laos, Cambodia, and Burma, where big ensembles of gong families play for court and state ceremonies. Rural areas include other islands and remote places, where smaller ensembles and solo instruments play a simpler music for village feasts, curing ceremonies, and daily activities. In cities and towns influenced by Hindu epics such as the *Rāmāyaṇa* and *Mahābhārata*, shadow and masked plays and dances utilizing music play important communal roles, while in less urbanized areas, in lieu of musical plays, chants and songs in spirit worship and rituals are sung in exclusive surroundings—a ritual procession on the headwaters of Borneo, a drinking ceremony in the jungles of Palawan, a feast in the uplands of Luzon.

In both regions the physical setting is usually the open air—in temple yards and courtyards, under the shade of big trees, in house and public yards, fields and clearings. Many musical instruments are made of natural products of a tropical environment, and their sounds are products of this milieu. The music of buzzers, zithers, and harps is thus akin to sounds heard in the tropical vegetation of Southeast Asia. In Bali, for example, special ways of chanting and sounds of the jew's harp (also jaw's harp) ensemble (*genggong*) imitate the croaking of frogs and noise of animals.

Relation to social institutions. Music in Southeast Asia is frequently related to ceremonies connected with religion, the state, community festivals, and family affairs. In Java, important Islāmic feasts, such as the birthday of Muḥammad or the fast of Ramaḍān, as well as animistic ceremonies marking the harvest and cycles of human life, are celebrated with shadow plays (*wayang* [*wajang*]). In Bali, the *gamelan gong* orchestra opens ceremonies and provides most of the music for temple feasts. The *gamelan selunding*, an ensemble with iron-keyed metallophones (like xylophones but with metal keys), plays ritual music, and the *gamelan angklung*, so called because it formerly included tube rattles, or *angklung*, is used to accompany long processions to symbolic baths near the river.

In Malaya the court orchestra, or *nobat*, was held almost as sacred as the powers of the sultan himself. Among the Land Dayak and Iban in Borneo, ceremonial chants are sung in feasts related to rice planting, harvesting, and honouring the omen bird *kenyalang* and other spirits.

The relation of music to dance and theatre. In the Thai masked play, or *khon*, dancers, chorus, soloists, and orchestra are all coordinated. The musicians know the movements of classical dance and coordinate musical phrases with dance patterns, turns, and movements. In the shadow play, or *nang sbek*, the dancer, who manipulates a leather puppet, must keep his foot movements in time with vocal recitations. During pauses in which the gong ensemble plays an interlude, the dancer must change steps accordingly. In general, when there is solo singing, the instrumental ensemble remains silent or plays only a few instruments in contrast to interludes of acrobatic shows or scenes of fighting, when the full orchestra clangs on all the instruments. In Balinese dancing, body movements, paces, and directions are dependent on drum strokes and signals from a wood block (*keprak*) and cymbals (*tjengtjeng*). The dancers generally rehearse with the musicians to know exactly when choreographic changes take place.

Musicians' knowledge of the dance

As theatre, the stories of *Rāmāyaṇa* and *Mahābhārata* have different musical supports, depending on the country. In Bali, *Mahābhārata* shadow plays are presented to the accompaniment of a quartet of metallophones known as *gender wayang*. In Cambodia, where the preference is for stories of the *Rāmāyaṇa* (which is called *Ramker* in Cambodia), the music is a full gong ensemble similar to the Thai *pi phat* ensemble, while in Burma, a percussion orchestra of drums and gongs in circular frames accompanies singing, dancing, and dialogues in all types of plays.

Musical traditions and practice. *Vocal music.* The role of the voice in music making differs from that of European music in both concept and execution. Men's and women's voices are each not divided into high and low ranges but are used for their colour qualities. In the Javanese shadow play, for example, the narrator (*dalang*) assumes many singing and speaking qualities to depict different characters and scenes. Arjuna, the chief *wayang* hero, is represented with a clear voice, speaking in a single tone. Puppets with bigger bodies are given lower, resonant voices. In Thai masked plays there is no desire to produce full open tones, as in Italian bel canto. A vocal tension accounts for shades of "nasal" singing that can be discerned in commercial recordings of Thai, Javanese, Cambodian, and Vietnamese music. In the Javanese orchestra (gamelan) the voice tries to imitate the nasality of the two-stringed fiddle (*rebab*). In Bali, a particular use of men's voices is in the *ketjak*, a ritual in which groups seated in concentric circles combine markedly pronounced syllables into pulsing rhythmic phrases. In village settings among the Kalinga of Luzon, in the Philippines, singing, speaking, or whispering of vowels is so subtle as to blur the border line between speech and song. On the Indonesian island of Flores, leader–chorus singing, with the chorus divided into two or more parts, is accompanied by a prolonged note (drone) or by a repeated melodic, rhythmic fragment (ostinato). In Borneo, or Mindanao and Luzon in the Philippines, a man or woman may sing an epic or a love song in a natural voice with little or no attempt to nasalize it. Epic singing, with long or short melodic lines, goes on for several nights, and some of the sounds are mumbled to give words and their meanings a particular shading. Further, a sensuousness in the quality of Islāmic singing is achieved through the use of shades of vowel sounds, vocal openings, and a bell-like clarity of tones.

Voice quality and dramatic meaning

Instrumental music. Although gong orchestras consisting of gongs, metallophones, and xylophones bind Southeast Asia into one musical cultural group, the types of ensembles and sounds they form may be classified into four areas. Java and Bali make up one unit because of their predominant use of bronze instruments in orchestras that make one homogeneous sound. Thailand, Laos, and Cambodia form another subdivision, with families of musical instruments producing heterogeneous sounds: the bronze group makes slowly decaying sounds, wooden xylophones play short sounds, and a reed blows a penetrating melody accompanied by a fourth group of cymbals, drums, and another gong. The Burmese orchestras differ from the Indonesian and Thai groups by the unique use of a row of tuned drums (sometimes called a drum circle), with sounds consisting of sharp attacks and quick-vanishing waves. The fourth area, Indonesia, Malaysia, and the Philippines, uses several types of suspended and horizontally laid gongs. These gongs produce various combinations of sounds. In Nias, an island west of Sumatra, one group of three heavy suspended gongs plays three rhythms of homogeneous sounds. Suspended gongs with a wide rim and a high knob are played alone, with another gong or with a drum on the Philippine islands of Mindanao and Palawan and the Indonesian island of Kalimantan (Borneo). Gongs laid in a row, called *kulintang*, are melody instruments accompanied by a percussion group. The most developed melodies are found in Mindanao, and the area of distribution extends to Borneo, Sumatra, and Celebes, in Indonesia. The sets of tuned gongs found throughout Southeast Asia are also called gong chimes, gong kettles, and gongs in a row.

Homogeneous and heterogeneous sound

Tonal systems. In contrast to the Western diatonic-scale system (based on seven-note scales comprised of whole and half steps) and its association with relatively "fixed" pitches, there prevails a gapped system in Southeast Asia (*i.e.*, scales containing intervals larger than a whole step) with elastic intonation. Examples include the five-tone *slendro* and the seven-tone *pelog* of Java and the seven-tone scale of Thailand. In each of these systems the distances between corresponding tones in two different sets of octaves are not exactly the same. For example, one Javanese *slendro* octave has the following intervals expressed in cents (a unit of pitch measurement; 1,200 cents make 12 semitones or one octave): 246, 241, 219, 254, 246; another has 245, 237, 234, 245, 267. In contrast, two tunings of the Western chromatic scale theoretically always have 12 semitones of 100 cents apiece.

Modal systems

Related to tonal systems are modes, which in Southeast Asia use tones of a particular scale system to form melodies. Associated with a given mode are a hierarchy of pitches, the principal and auxiliary tones, endings of melodic phrases (cadential formulas), ornaments, and the vocal line. Modes express emotions and are applied to different times of the day and night and to particular situations in stage plays. They are clearly present, with local variations, in Java, Vietnam, and Burma but are less distinct in Bali, Thailand, Laos, and Cambodia.

In rural areas a multitude of scales with mixed diatonic and gapped systems and no modes are used.

Musical time and improvisation. Musical time is generally divisible in units of two or four in urban music, but it occurs more freely and without a metric pulse in rural areas, especially in singing. Musical improvisation or the use of variations based on a melodic theme is not universal. It is essential to the playing of the *rebab* and singing in the Javanese gamelan, the tappings on the Burmese circle of drums, and the percussive playing on the *kulintang*. But, in fast playing in the Balinese gamelan, exact repetitions of patterns are necessary, for there is no time for the performer to think of alternative formulas. Similarly, the separate rhythmic patterns of five instrumental parts do not change in the gong (*gangsa*) music of the Ibaloi of Luzon. Repetition is the essence of the music.

HISTORICAL DEVELOPMENTS

Origins. *Early bamboo instruments.* The widespread use of bamboo musical instruments in practically all parts of Southeast Asia points to the antiquity of these instruments and, probably, that of the music they play. A historical citation of mouth organs and jew's harps in the Chinese *Shih Ching* ("Classic of Poetry") shows that these instruments were known in the 8th century BC. Previous to this time, other bamboo musical instruments were probably in use, just as bamboo tools were used in pre-Neolithic times.

The music of pre-Neolithic types of bamboo musical instruments, such as are played in the 20th century, may be just as old as these instruments. One general feature that points to this antiquity is the widespread and frequent use of a very simple musical element: a sustained tone (drone) or repetition of one or several tones (ostinato). Sustained tones appear in the mouth organ, where one or two continuous sounds are held by one or two pipes while a melody is formed by the other pipes. Prolonged tones may also be heard in rows of flutes played by one person in Flores. One flute acts as ostinato and the rest make a melody. In group singing, an underlying held tone is common. Repetition of tones occurs in bamboo instruments (jew's harps, percussion tubes and half percussion tubes, zithers, clappers, slit drums) as well as in nonbamboo instruments. In the *kudjapi*, a two-stringed lute, one string is used for the ostinato and the other to pluck the melody. In the log drum, two players play fast rhythms of continuous sounds while another player taps improvised rhythms.

Drones and ostinato

Bronze instruments in gong families of Indonesia, Thailand, and Burma employ repeated sounds acting as ostinati. A widespread and preponderant use of drone-like or repeated sounds in Southeast Asia shows that they are probably an ancient fundamental musical element.

Early bronze instruments. The earliest bronze musical instruments are kettle gongs (deep-rimmed gongs), which date back to about 300 BC and are found in Vietnam, Bali, Sumatra, Borneo, Thailand, and Burma. In Burmese gongs the use of a heavy beater for the centre and a lighter stick to strike the side denotes an opposition of a full and a tiny sound applied today also to the *babandil* and other gong ensembles in Palawan and Borneo.

Gongs that predominate in Southeast Asia are those with a boss, or central beating knob. The many varieties differ according to their shapes, chemical properties, playing position, number in a series, manner of playing, musical function, and sound. Flat gongs without a central boss are not as widely used. They are found in the hills of Thailand, Laos, Cambodia, Vietnam, in some parts of Indonesia, and in the northern Philippines and may have come to Southeast Asia either through China in the 6th century or from the Middle East.

Musical traditions. The influence of the great traditions of Asia—Indian, Chinese, Islāmic, and Khmer (Cambodian)—on native Southeast Asian music varies in different countries. From India come principally two ancient Sanskrit epics—the *Mahābhārata* and the *Rāmāyaṇa*. Deep attachment to themes from the *Rāmāyaṇa* pervades the whole Southeast Asian region, except the Philippines, where Indian influence was weakest. Musical instruments attributed to India and appearing in 9th-century reliefs at the Buddhist temple of Borobuḍur and Hindu temple of Prambanan, in Java, are bronze bells, bar zithers, cymbals, conical drums, flutes, shawms, and lutes. They may still be found in several islands of Indonesia. Khmer gong circles, stringed instruments, mouth organs, drums, and oboes still in use in rural Cambodia and Vietnam are depicted in the 12th-century ruins at Angkor Wat in Cambodia. Prehistoric lithophones, or stone chimes, excavated in Vietnam in 1949, may have been the ancestors of kettle gongs. Chinese-type musical instruments (two- and three-stringed fiddles, bells, and drums), the use of the Chinese pentatonic (five-tone) scale, and duple and quadruple time (typical Chinese metres) are used in Vietnam, Burma, Thailand, Laos, and Cambodia, Islāmic musical instruments—drums, two-stringed fiddles (*rebab*), and three-stringed lutes—may be heard in Java, while melismatic singing (many notes to one syllable), especially in Islāmic rituals, is usual among the Malay groups on Borneo.

Mingling of influences

There are also musical instruments and elements that have developed locally. The mouth organs of Borneo, Laos, and Cambodia are probable ancestors of the Chinese *sheng* and the Japanese *sho* (mouth organs). Jew's harps, tube zithers, ring flutes, buzzers, xylophones, two-stringed lutes, and various types of gongs with boss (knobbed centre) are some of the most typical instruments of Southeast Asia. A probably ancient manner of measuring flute stops in Mindanao—dividing flute segments into proportional lengths to produce the octave, fifth, and other intervals—recalls a very old Chinese account of cutting bamboo tubes into lengths that would sound these same intervals.

In general, music in Southeast Asia is a tradition taught to each succeeding generation without the use of written notation. From exclusive families of musicians in courts, gamelan music was transmitted to the people. Epic and ritual songs are learned by rote and handed down from older to younger generations. Hence, skill in instrumental music is developed by imitation and practice.

Burma. Just as today all types of Burmese plays are accompanied by the traditional Burmese orchestra, the beginnings of Burmese theatre contained a music that, like the theatre, was probably based on primitive religious rituals. Before Indian and Chinese musical influences, the inspirational source of Burmese music and dance was the miracle plays (*nibhatkhin*), which, in turn, were based on singing, dancing, and entertainment in local folk feasts that date back to antiquity. The worship of spirits (*nats*) at Chinese festivals was accompanied by women who, through song and dance, communicated with and were possessed by these spirits.

Following this practice, professional entertainers taking the place of women danced, sang, and played instruments during the first *nibhatkhin*. These practices led to the dancing and singing associated with the *pwe*, a popular play for public and courtly entertainment.

Foreign musical influences came from India, China, and Thailand. Indian elements appear in musical terms, theories about scales, and in some musical instruments—oboe, double-headed drums, cymbals, and the arched harp. Chinese influence appears to be older and is apparent in the use of the pentatonic scale and such musical instruments as table zithers (related to the Chinese *ch'in*), a dragon-head lute resembling a Chinese *p'i-p'a*, and two- and three-stringed fiddles. From Thailand and the Khmer civilization of Cambodia probably came both the use of gongs in a circular frame and the dramatization of episodes from the *Rāmāyaṇa*. In the traditional orchestra for state ceremonies, for the theatre, and, formerly, for royalty, three simultaneous variations of the same theme are performed by two sets of melodic percussion—a circle of about 21 tuned drums (*saing-waing*) and a circle of about 21 tuned gongs (*kyi waing*)—and at least one oboe (*hne*) or a flute (*pulwe*). To this is added a playing of a percussion group comprising a double-headed drum (*patma*), a pair of cymbals (*la gwin*), and clappers playing a duple or a quadruple metre. In three rhythmic patterns applied by these percussion groups to specific song types, the strong beats are always marked by the clappers.

Melodic traits

Melodies played on traditional instruments (*saing-waing*, harp, *pattala* or xylophone) are frequently broken by rests and consist of segments of two, three, or four notes that form phrases, usually of eight or 16 beats. Several phrases make up a number of verses to complete a musical rendition. Melodies, based on modes, are constructed according to the previously discussed elements usually found in the modal music of Southeast Asia. Song types exist in Burmese music and are assigned to specific modes.

The Burmese arched harp (*saung gauk*) has features that may be traced back to pre-Hittite times and the Egyptian 4th dynasty (*c.* 2613–*c.* 2494 BC). Scarcely existent outside of Burma, this instrument has undergone a renascence in the 20th century. A more popular solo instrument is a wooden xylophone *pattala*.

The following instruments may be found among ethnic groups in rural Burma: idiophones, or resonant solids—bamboo jew's harps, clappers, cymbals, wooden slit drums, bronze kettle gongs, drums; membranophones, or vibrating-membrane instruments—goblet drums; chordophones, or stringed instruments—crocodile zithers, monochords with calabash resonators, three- and four-stringed fiddles; aerophones, or wind instruments—lip-valley flutes, ring flutes, panpipes, double-reed winds, buffalo horns, and mouth organs.

Thailand, Laos, and Cambodia. Although their individual political histories differ, the music practiced in Thailand, Laos, and Cambodia is almost identical. The musical instruments and forms of this region spring from the same sources: India, the indigenous Mon-Khmer civilizations, China, and Indonesia. In Thailand, three types of orchestras, called *pi phat, kruang sai,* and *mahori,*

Thai ensembles

exist. The *pi phat,* which plays for court ceremonies and theatrical presentations, uses melodic percussion (gongs in a circle, xylophones, metallophones) and a blown reed. The *kruang sai* performs in popular village affairs and combines strings (monochords, lutes, and fiddles with two and three strings) and wind instruments (oboes and flutes); while the *mahori,* as accompaniment of solo and choral singing, mixes strings (floor zithers, three-stringed fiddles, and lutes) and melodic percussion (gongs and xylophones) with the winds (flutes and oboes). All three ensembles are provided with a rhythmic group of drums, cymbals, and a gong to punctuate the melody parts. Some of the above musical instruments and their functions may best be illustrated in the *pi phat* ensemble below.

A slow-moving theme is played by gongs arranged in a circle (*khong wong yai*) with variations in smaller gongs (*khong wong lek*), two wooden xylophones (*ranat ek, ranat thum*), and two box-shaped metallophones (*ranat thong ek, ranat thong thum*). The last three pairs of instruments vary the theme by playing twice as fast or by repeating, anticipating, and revolving around it. A double-reed oboe (*pi nai*) hovers above the melodic percussion, providing the only blown sound in the ensemble. Together with the punctuating gongs and drums, the whole orchestra displays a polyphonic (many-voiced) stratification of instrumental parts, using unisons and octaves mainly in the strong beats.

A melody may be broken down into phrase units consisting of two or four measures that may be joined by four other phrase units to make a phrase block, and a given number of blocks constitutes one musical composition. Three speeds of rendition—slow, medium, fast—in either duple or quadruple time are marked by two alternating strokes in a pair of cymbals; a dampened clap marks a strong beat, and a ringing vibration denotes a weak beat.

The tuning system is made up of seven tempered (approximately equidistant) tones to an octave. But the melodies constructed out of this system use only five tones out of seven—which sound close to a Chinese pentatonic scale. This scale may be constructed in any of seven levels or tones of the Thai tuning system. Further, through a process called *metabole,* melodies may move from one level to another.

Shadow-play music and dance

In the Cambodian shadow play (*nang sbek*) two narrators alternate in chanted recitative to explain the role of the leather puppets. Dancers parading these figures across the screen and simulating their actions are accompanied by an orchestra. A limited number of tunes is played to eight dance positions (walk, flight or military march, combat, meditation, sorrow or pain, promenade, reunion, and metamorphosis). In the play these poses are assumed by princes, princesses, monkeys, demons, peasants, or ascetics.

Among different ethnic groups, such as the Khmer Saoch, Pwo Karen, Bu Nor, Kae Lisu, Kui (Kuoy), and Samre, a rural music related to that of the ancient Khmer peoples is played by aerophones (buffalo horns, mouth organs, vertical flutes), idiophones (flat gongs, gongs with boss, cymbals, jew's harps), chordophone (bamboo zithers), and membranophone (circle of drums). Other important instruments for solo performance or as accompaniment to songs are the three-stringed crocodile zither (*chakhe*), a four-stringed lute (*grajappi*), a plucked monochord with a gourd resonator (*phin nam tao*), and a bamboo whistle flute (*khlui*).

Vietnam. Although Vietnamese music belongs to the great Chinese musical tradition, which includes the music of Korea, Mongolia, and Japan, some of its musical elements are indigenous or come from other parts of Southeast Asia, and some derive from Champa, an ancient Hinduized kingdom of Vietnam. Archaeological finds in the village of Dong Son revealed that the ancient Vietnamese used kettle gongs, mouth organs, wooden clappers, and the conch trumpet. From the 10th to the 15th century a joint Indian and Chinese element left its musical imprint. The Chinese seven-stringed zither (*ch'in*) and a double-headed drum were played together, or a Champa melody was accompanied by a drum. It was at this time that two traditional Chinese ensembles—Great Music and Little Music—and an elementary Chinese theatrical art were introduced. From the 15th to the 18th century the Chinese influence reached its height. Court music (*nha nhac*) was played by two orchestras. One, located in the Upper Hall of the court, consisted of a chime of 12 stones, a series of 12 bells, a zither of 25 strings (Chinese *se*), a zither with seven strings (Chinese *ch'in*), flutes, panpipes, a scraper in the shape of a tiger, a double-headed drum, a mouth organ, and a globular whistle. The second orchestra in the Lower Hall used 16 iron chimes, a harp with 20 strings, a lute with four strings (Chinese *p'i-p'a*), a double flute, a double-headed drum, and a mouth organ. Ceremonial music, almost nonexistent in the 20th century, was patterned after court music.

Traditional court music

In Buddhist ceremonies, prayers were recited in three

ways: as recitation in a low voice, as a cantillation (sung, inflected recitation) following the six tones of the Vietnamese language, and as chant accompanied by an orchestra of two drums, bell, gong, cymbals, and fiddles.

Music as entertainment is mostly a vocal art played without ritual outside the court and still enjoyed by many people. The *hat a dao* found in the north is the oldest form. It is a woman's art song with different instrumental accompaniments, dances, a varied repertoire, and a long history of evolution.

From the 19th century to World War II, Vietnamese music reaffirmed its character. Although the playing of court music was restricted, popular music was encouraged, leading to northern and southern styles that were patronized by both the aristocracy and commoners. Western musical influence in this period was manifest in the use of the mandolin, the Spanish guitar, and the violin, as well as by the introduction of European classical music and composition following Occidental forms. In the late 20th century traditional Vietnamese music began to disappear, but attempts were being made to revive it in the early 1970s.

Vietnamese rural folk music is built on the same musical principles as court music. The main difference lies in its application to village activities—work, games, courting, marriage, cure for the sick, entertainment, feasts.

Elements of Vietnamese music

Common elements characterize and unify all Vietnamese music. It is based on an oral tradition, with written notation serving only as a reading guide. Melodies are generally built out of a pentatonic system (for example, C, D, F, G, A) to which two auxiliary tones (E, B) may be added to make other pentatonic melodies. A song, usually preceded by a prelude, may be sung in slow, moderate, or fast tempo divisible by two or four, with a simple contrapuntal (countermelody) accompaniment using unisons and octaves at beginning points of phrases. Outside of the first beats, intervals of fifths, fourths, thirds, and even seconds are allowed. An important aspect of melodies is the idea of mode (*dieu*), the elements of which do not essentially differ from those of Javanese and Burmese music.

Indonesia and Malaysia. *Java.* A Javanese philosophical concept based on mysticism, refinement (*halus*), and the inner life as related to Hindu, Islāmic, and Indonesian thought may best be represented in music by the Javanese gamelan, an orchestra made up mostly of bronze instruments producing homogeneous blended sounds.

Instrument groups of the gamelan

The instruments in the ensemble may be divided into three groups of musical function. The first group comprises thick bronze slabs (*saron demung, saron barung, saron panerus*) on trough resonators playing the theme usually in regular note values without ornamentation. The second group consists of elaborating or panerusan instruments, which add ornaments to the main theme. In this group gongs in double rows (*bonang panembang, bonang barung, bonang panerus*) play variations with the same ratio of speed as the saron group. In softer sounding music for indoor performance, other panerusan instruments with very mellow sounds come in. These are three sizes of thin bronze slabs with bamboo resonators—gender panemburng or slentem, gender barung, and gender panerus. Other elaborating instruments are the wooden xylophone (*gambang*), the zither (*tjelempung*) with 26 strings tuned in pairs, an end-blown flute (*suling*), and a two-stringed lute (called a *rebab* by the Javanese), which leads the orchestra. In loud-sounding music, the soft-sounding instruments are not played, and the drum (*kendang*) leads the orchestra. The third group provides "colotomic," or punctuating beats in four rhythmic patterns played separately by four types of heavy, suspended, or horizontally laid gongs.

Scales and modes

Two tuning systems prevail. The *slendro* tends to have five equidistant but flexible (or varying) pitches in an octave, while the *pelog*, with seven equally flexible tones, has a more varied structure. One tuning with intervals expressed in cents (140, 143, 275, 127, 116, 204, 222) may roughly be represented by the following notes in a descending scale: C↑, A♯, G♯, G↓, F↑, D♯↓, C♯↑, and C. (Arrows up are tones slightly higher than Western tempered tuning [in which a semitone is equivalent to 100 cents] and vice versa for arrows down.) Melodies from these tunings are governed by a modal structure (*patet*) the elements of which are similar to those of Vietnamese and Burmese music.

In West Java the most popular ensembles use a vocal part, a two-stringed fiddle (*rebab*) or a bamboo flute (*suling*), and a zither (*kachapi*). In the gamelan, submodes (*surupan*) are formed by the use of vocal tones—sung or played on the *suling* or *rebab*—which amplify the number of scales in both the *pelog* and *slendro* systems.

Bali. In contrast to the introspection of Javanese music, the Balinese gamelan exudes a music of brilliant sounds with syncopations (displaced accents) and sudden changes, as well as gradual increase and decrease in volume and speed and feats of fast, precise playing. The tuning system, musical instruments, and polyphonic stratification are similar to those of the Javanese gamelan, although in Bali the seven-tone *pelog* is not popular. Most gamelan are tuned to a five- or four-tone system, and the concept of modes is not as clearly developed as in Java. A variety of gamelan exists, each with a special function, instrumentation, repertoire, and tuning system. The *gamelan gong* orchestra is among the most extensive in its number of instruments. A modern version, *gong kebjar*, omits the *trompong* (gongs in a row) and *saron* (bronze slabs over a trough resonator) and replaces them with *gangsa gantung* (metallophone with bamboo resonators) and *rejong* of four gongs to produce exuberant outbursts of sound. The *gamelan gambuh*, now rare, comprises four end-blown flutes, one *rebab*, and a group of percussion. The *gamelan semar pegulingan*, played formerly in royal courts but now almost disappeared, emphasizes the *trompong* as a solo instrument. The *gamelan pelegongan* is a virtuoso orchestra that accompanies *legong* dances, while the *gamelan pedjogedan* is an orchestra of xylophones for dance (*djoged*) and entertainment in the marketplace. The *gender wayang* is a quartet of *slendro* tuned metallophones specially employed for shadow plays. The *gamelan angklung*, a village orchestra assembled during ceremonies, anniversaries, and cremations, originally consisted of rattling tubes that are now replaced by metallophones. The *gamelan ardja* is characterized by a soft timbre (tone colour) and the use of a one-stringed bamboo zither, the *guntang*, to accompany musical comedy and popular plays.

Varieties of Balinese gamelan

Other parts of Indonesia. In the islands of Flores, Nias, New Guinea, Celebes, and Borneo idiophones make up perhaps the most varied collection of musical instruments—gongs of various profiles, slit drums, jew's harps pulled with a string, clappers, bells, xylophones, percussion sticks, bull-roarers, and stamping tubes. Particularly interesting are idiophones made of bones, shells, skulls, fruits, seeds, planks, pellets, crab claws, clogs, coconut, and shark bones. Membranophones are represented by drums shaped like a cylinder, goblet, vase, round frame, hourglass, cone, cup, barrel, or a tube. Aerophones present an array of vertical and transverse (horizontally played) flutes, panpipes, ring flutes, shawms, clarinets, gourd trumpets, conch shells, ocarinas, and flutes with different mouthpieces. Chordophones include bamboo zithers, spike fiddles (in which the neck skewers the body), one- and two-stringed lutes, musical bows, monochords, guitars, *rebabs*, bar zithers, and sago zithers. In Flores, part singing with a sustained drone is frequent. Songs in Nias use diatonic (whole and half steps), chromatic (half steps), and gapped melodies largely less than an octave in range. In Sarawak descending melodies make up a tetrachord (four adjacent tones forming the interval of a fourth). In New Guinea departures from songs with gapped scales include fanfare, stair descent, and tiled melodies (the last consisting of short phrases repeated at different pitch levels).

Malaysia. At least three principal cultural influences —Indonesian, Hindu, and Islāmic—left their musical marks in Malaysia. The Indonesian influence is seen principally in musical forms, participants, and paraphernalia of the Malaysian shadow play (*wayang kulit*). It is said

that the Indian epics and, especially, the *Pandji* tales of Java came to Malaysia via Indonesia, but there are songs in certain plays and musical instruments (*e.g.*, the double-headed drum and oboe) that could have reached Malaysia from India through other routes. Islāmic traces are evident in melismatic songs among the Malay groups in songs connected with religious rituals and in choral singing in the *ma'yong* plays. Chinese music, a more recent development, is largely practiced among the Chinese communities, principally in Singapore.

Before Malaysian independence, the *nobat,* an old royal instrumental ensemble dating back to about the 16th century, played exclusively for important court ceremonies in the palaces of the sultans of Perak, Kedah, Selangor, and Trengganu. Today, in Kedah, the ensemble consists of five instruments: one big goblet drum (*negara*), two double-headed drums (*gendang*), one long oboe (*nafiri*), one small oboe (*nafiri*), and one gong. The music, which consists of ten surviving pieces, is broadcast today and performed live.

Three shadow plays exist, principally in the state of Kelantan. The *wayang gedek* is the Thai form; *wayang Djawa,* a Malay form, is almost extinct; and the *wayang Siam,* which is a combination of Thai and Malay influences, is the most popular form of puppet shadow play. The operator of the performance is the narrator (*dalang*), who manipulates the leather figures, introduces important characters, and describes different scenes with the accompaniment of the orchestra. The music is led by a two-stringed lute (*rehab*) in the *Rāmāyaṇa,* or an oboe (*serunai*) in *Mahābhārata* and *Pandji* cycles. The melodic instruments are supported by a percussion group consisting of pairs of goblet-shaped drums (*gedombak*), cylindrical drums (*gendang*), barrel drums (*geduk*), gongs lying on a support (*chanang*), suspended gongs (*gong*) or, sometimes, a row of gongs played by two or three men, and one pair of cymbals (*kesi*). The music usually begins with a prelude followed by a list of pieces the sequences of which are dictated by the narrator.

The *ma'yong,* a dance drama that probably dates back to more than 1,000 years, was introduced in Kelantan under the patronage of the royal courts. In the 20th century it exists as a folk theatre with an all-female cast. The music that accompanies 12 surviving stories is played by an orchestra of one bowed lute (*rebab*), two suspended gongs, and a pair of double-headed drums (*gendang*). A heterophony (simultaneous variation of the same melody) between a solo voice, a chorus, and the *rebab* creates a music with a Middle Eastern flavour.

Variety in instruments

A rich musical heritage in the rural sections of Malaya is shown in musical instruments used by Malay, Thai, Semang, and Sakai groups. Idiophones include shell and coconut rattles, the jew's harp (mostly pulled by a string, rather than plucked), bull-roarers, bamboo clappers, and the bamboo slit drum. Aerophones include the buffalo horn, wooden and clay whistles, nose flutes, end-blown flutes, and the oboe. Chordophones are two- and three-stringed fiddles with coconut resonators, monochords, and tube zithers. One membranophone is a double-headed cylindrical drum.

In Borneo among the Malay, Kadazan, and Iban groups, the principal instruments are gongs in a row (*gulintangan*) played with suspended gongs of different types (*chanang, gong, tawak-tawak*). Among the Murut, Kenyah, and Iban the mouth organ with a calabash resonator (*sompoton*) plays a melody with a drone accompaniment. The jew's harp (*ruding*), bamboo zither (*tongkungon*), nose flute (*tuali*), hourglass drum (*ketubong*), and vertical flute (*suling*) may be heard among different ethnic groups. Iban ceremonial songs are sung in connection with rice festivals and rituals to prevent sickness, while mourning songs make up a rich repertoire of solo and leader–chorus singing. The Kenyah are particularly adept at blending low voices of men singing a melody supported by a drone.

The Philippines. Two musical cultures—Western and Southeast Asian—prevail in the Philippines. Western music is practiced by about 90 percent of the population, while Southeast Asian examples are heard only in mountain and inland regions, among about 10 percent of the people.

The Western tradition dates back to the 17th century, when the first Spanish friars taught plainchant and musical theory and introduced such European musical instruments as the flute, oboe, guitar, and harp. There subsequently arose a new music related to Christian practices but not connected with the liturgy. Processional songs, hymns in honour of the Blessed Virgin, Easter songs, and songs for May (Mary's month) are still sung in different sections of the country. A secular music tradition also developed. Guitars, string ensembles (*rondalla*), flute, drum, harps, and brass bands flourished in the provinces among the principal linguistic groups and still appear during town fiestas and important gatherings. Competing bands played overtures of Italian operas, marches, and light music. Young men, like their counterparts throughout the Hispanic world, sang love songs (*kundiman*) in nightly serenades beneath the windows of their ladies. It was not uncommon in family gatherings for someone to be asked to sing an aria, play the harp, or declaim a poem. Orchestral music accompanied operas and operettas (*zarzuelas*), while solo recitals and concerts were organized in clubs or music associations. With the advent of formal music instruction in schools, performance and composition rose to professional levels. In the 20th century several symphony orchestras, choral groups, ballet companies, and instrumental ensembles performed with varying regularity.

Gong ensembles

A Southeast Asian musical tradition exists completely apart from the Western tradition. In the north, flat gongs are played in different instrumental combinations (six gongs; two gongs, two drums and a pair of sticks; three gongs). In the ensemble with six gongs, four are treated as "melody" instruments, one as ostinato, and another as a freer layer of improvisation. The melody consists of scattered tones produced by strokes, slaps, and slides of the hands against the flat side of the gong. Other musical instruments in the northern Philippines are bamboo. These are the nose flute (*kalleleng*), lip-valley or notched flute (*paldong*), whistle flute (*olimong*), panpipes (*diwdiwas*), buzzer (*balingbing*), half-tube percussion (*palangug*), stamping tube (*tongatong*), tube zither (*kolitong*), and jew's harp (*giwong*). Leader–chorus singing among the Ibaloi is smooth and sung freely without a metric beat, while the same form among the Bontoc is emphatic, loud, and metric. Scales in songs and musical instruments use from two to several tones within and beyond an octave and are arranged as gapped, diatonic, and pentatonic varieties.

In the southern Philippines (the islands of Mindanao and Sulu), the more developed ensemble is the *kulintang,* which consists of eight gongs in a row as melody instruments accompanied by three other gong types (a wide-rimmed pair; two narrow-rimmed pairs; one with turned-in rim) and a cylindrical drum. The *kulintang* scale is made up of flexible tones with combinations of wide and narrow gaps sometimes approaching a Chinese pentatonic variety and oftentimes not. Its melody is built on nuclear tones consisting of two, three, or more tones to form a phrase. Several phrases may be built, repeated, and elongated to complete one rendition lasting two to three minutes. Pieces of music are played continuously for a long period during the night.

In the central west Philippines on the island of Mindoro, love songs are sung that are based on reciting tones with interludes played by a miniature copy of the Western guitar or a small violin with three strings played like a cello.

(Jé.Ma.)

IV. The performing arts

In variety of dance and theatrical forms and in the number of performing groups, no area in the world except India and Pakistan compares to Southeast Asia. Some form of the performing arts is a normal part of life throughout the several nations. Sophisticated performing groups cluster in and around the present and former court cities—Jogjakarta and Surakarta in Java, Ubud and Gianjar in Bali, Bangkok in Thailand, Mandalay in Bur-

ma, Siem Reap near Angkor and Phnom Penh in Cambodia, Hue in Vietnam—where drama, puppetry, dance, and music have been cultivated for ten centuries or more. Hundreds of commercial theatrical and dance groups perform in such newer centres as Rangoon, Saigon, and Djakarta and in scores of provincial cities and towns. Wandering troupes of actors, puppeteers, singers, and dancers travel from village to village in areas adjacent to these population centres. There are few communities in which some form of folk dance is not peformed by local people.

Richness of form and experience

In the West, music, dance, and drama are usually separate arts, whereas in all areas of Southeast Asia, drama, dance, mime, music, song, and narrative are integrated into composite forms, often with masks or in the form of puppetry. The spectator's senses, emotions, and intellect are bombarded simultaneously with colour, movement, and sound. The result is a richness and a vividness in the theatre that is absent in most Western drama, so much of which rests on a literary basis.

More than 100 distinct forms or genres of performing arts can be distinguished in Southeast Asia. These can be grouped, according to which of the various stage arts is emphasized, into (1) masked dance and masked dance-mime, (2) unmasked dance and dance-drama, (3) drama with music and dance, (4) opera, (5) shadow-puppet plays, and (6) doll- or stick-puppet plays.

DIVERSE TRADITIONS IN THE PERFORMING ARTS

Four relatively distinct traditions exist in the performing arts: folk, court, popular, and Western.

The folk tradition. Dances in the folk tradition are exceptionally numerous and widespread. Some are performed as religious ritual, others, particularly on the Indonesian island of Bali, by highly trained and respected artists, and still another kind as entertainment in which the community participates. Folk theatre is more complex than folk dance and thus less widespread, but it has deep connections with religious ritual. Although the origins of most folk performing arts lie in remote times, later court forms exerted important influence on many of the folk forms. Conversely, folk forms have been a source of inspiration to court artists.

The court tradition. The shadow play and masked and unmasked dance are court arts reflecting centuries of subtle refinement under the patronage of kings and princes.

Shadow plays

In Southeast Asia the shadow theatre is a major classic art. Leather puppets of mythological figures, the bodies intricately incised to allow light to pass through, are attached to sticks for manipulation. A lacy shadow is created by a flaming lamp as the puppet is pressed against the back of a vertical screen of white cloth. The flickering and insubstantial shadow seen from the other side creates for the understanding viewer a mystic world with deep symbolic meaning. In Java, Bali, Malaysia, Cambodia, and Thailand shadow plays and their techniques have been emulated by human actors and dancers and have been the models for marionette and doll-puppet theatre.

Dance troupes have been a part of court life at least since recorded history began. In the mainland courts of Cambodia, Thailand, Laos, and Burma, concubines of the ruler's harem who performed female dances were segregated from male performers, giving rise to separate forms of female unmasked dance and male masked dance-mime. Although certain dances traditionally are performed only by men or only by women in Indonesia and Vietnam, mixed casts have a long history, especially in dramatic pieces. Court dance on the mainland and in Indonesia has been influenced by Indian dance style, and Vietnamese dance by the dance styles of Chinese opera, but they have acquired a distinctly Southeast Asian character. Court dance reached its greatest development when applied to mythological and legendary themes, often taken from the shadow theatre. The resulting dance-dramas and masked dance-mimes of Thailand, Cambodia, and Java are world famous for their magnificent scale and elegance of execution. Some of these court arts are no longer performed, and others face increasing difficulty securing financial support, yet they remain important.

The popular and Western traditions. In the popular traditions are those 400 to 500 professional troupes who perform, except in the Philippines, in commercial theatre buildings of major cities for an urban ticket-buying audience. Some forms of popular theatre are directly modelled on court dance-drama, but most are spoken drama in which court-derived music, song, and dance movements have been inserted. Local legend and history provide the subject matter for many of these plays. As in much of Asia, the performer in the popular tradition is seldom accorded status and may be despised as a vagabond.

The spoken drama, the ballet, and the modern dances are known only superficially in Southeast Asia. The sole exception is the Philippines, where amateur performances of Western plays constitute the country's main theatrical tradition. Southeast Asian audiences generally find Western plays based mainly on dialogue to be uninteresting and deficient in artistic qualities. European and American films and television programs, however, are widely shown and appreciated, and popular Western dances are found in major urban areas. Undoubtedly the impact of these forms on local audiences will continue to increase, possibly to the detriment of the indigenous traditions.

CHARACTERISTICS OF DANCE

Dramatic and nondramatic forms. In the parts of Southeast Asia influenced by Indian forms—everywhere except for Vietnam and the Philippines—nondramatic and dramatic dance are both known. Nondramatic, or "pure," dances that do not express emotional states of characters are numerous in both folk and court traditions. Among court dances, the Javanese *bedaja* is typical. Nine dancers move in unison, without emotional expression, in precisely fixed choreographic patterns designed to demonstrate sheer grace of movement. The *maebot*, composed as a Thai "alphabet of dance," is used to train pupils in the basic movements of court dance. Other dances that include character impersonation yet are not explicitly storytelling dances lie between nondramatic and dramatic dance. In the Thai *praleng*, two performers wearing god masks and holding peacock feathers in both hands perform an offertory dance to the god before the main dance-play begins. The Balinese *legong*, danced by a pair of preadolescent girls, may have only the most tenuous dramatic content. Its interest lies in the girls' unison rapid foot movements and fluttering movements of eyes and hands. Dramatic dance is seen at its best in full dance-dramas and in the excerpts from them that are sometimes danced in concert form.

Use of eyes and hands

Styles and conventions of movement and costuming. General characteristics of both dramatic and nondramatic dance are (1) slowness of tempo except in battle scenes, (2) controlled and reserved movements rather than expansive ones, (3) little of the leaping typical of Western ballet but, instead, a feeling of closeness to the ground, and (4) extensive use of arm and hand gestures. From Indian dance has come an open and flexed position of the legs, a side-to-side sliding movement of the head and neck, and a rigidly codified vocabulary of hand and finger gestures known as *mudrās* or *hastās* in India. In most cases the Indian elements have been altered greatly over their 1,000-year period of assimilation. In Thai, Cambodian, and Lao dance, the 24 to 32 Indian *mudrās* have been reduced to nine; in Javanese dance seven can be recognized, and in Bali only one or two. They have also been altered in their shape, and the many specific meanings attached to each in India have become fewer, while in some cases a gesture has no specific meaning. Such hand gestures as shading the eyes and tying the sash, which appear in Javanese dances, are unknown in India. Foot movements in India typically follow the rhythm of a drum, often with vigorous stamping sounds that are emphasized by bells on the ankles, but such movements are virtually absent in Southeast Asia. The exaggerated eye, eyebrow, cheek, mouth, and chin movements through which the Indian dancer expresses a broad gamut of emotions are nowhere to be seen. Balinese dancers use darting eye movements, but the court

dancer's face is composed into an almost unchanging expression of aloof gentility.

Close contact between neighbouring countries has led to the development of two regional Indian-influenced dance styles, one for Thailand, Cambodia, Laos, and Burma and one for Indonesia and Malaysia. Characteristics of the former style include the soft *pi phat* music of bamboo xylophones, drums, gongs, and oboe as accompaniment, bent-back finger positions not seen elsewhere in Asia, similar and often identical movements for male and female roles, courtship dances in which lovers touch each other and move in unison, and, in dance-drama, lengthy pure-dance pieces inserted solely for their beauty. In the latter style, the performance is accompanied by music of the gongs and metal bars of the gamelan orchestra. Scarves draped from the waist or neck are flicked for effect and manipulated to indicate strength or flying, and male and female dance are clearly distinguished by the powerful masculine lunges of the men and the tiny steps of the women, who also dexterously manipulate the train of the skirt with their feet. Visually, the mainland dance sparkles. Costumes of brilliant silk are covered with sequins and even jewels, and golden crowns and sparkling body ornaments glitter with reflected light. The male dancer in Indonesia wears a soft batik skirt of brown and white, the female a black velvet bodice. Arms and shoulders are bare and powdered golden brown, creating a subdued and warm effect.

Courtship dances

The main style in Vietnam, apart from folk dance, is dramatic and highly pantomimic, like the movements of Chinese opera. In classical opera, the flowing white sleeves and the pheasant feathers bobbing from the general's headdress are twirled and flicked by the actor in many conventionalized movements derived from Chinese forms. Battle scenes are choreographed into precise dance patterns, but the acrobatic movements common in Chinese opera are seldom seen.

CHARACTERISTICS OF DRAMA

Thematic origins and materials. Most traditional plays and dramatic dances are derived from mythological and legendary sources. The tribal epics that relate the origin of the Ifugao and the Bicolano peoples in the Philippines and a number of animistic stories in Indonesian shadow theatre are indigenous myths of great age, while the widely used, romantic *Pandji* cycle from Java and the Thai *King Abhai Mani* and *Khun Chang Khun Phan* are more recent local legends. The most important dramatic sources, however, are borrowed from the Indian *Rāmāyaṇa* and *Mahābhārata* epics, from the *Jātaka* Buddhist birth stories, from Chinese novels (such as *The Romance of the Three Kingdoms*) and Chinese operas, and from a host of Islāmic stories, including the *Thousand and One Nights* and the Amīr Ḥamzah tales. These foreign stories are turned into local legends. For example, the Indian Prince Rāma becomes a Thai, a Balinese, or a Javanese prince, embodying the heroic traits admired in each of these countries.

Plays are invariably extensive and have many scenes. It is not unusual for a play to present action over several generations, an indication of the value placed on cultural continuity. A recurring theme concerns restoration of harmony on earth by a ruler acting in accord with divine law. A kingdom is restored, a prince unjustly exiled returns to assume his throne, a usurper is punished, or the prosperity of the land is assured by consummating a particularly desirable marriage. As in Western drama, the hero gains his ends through struggle. Because he acts as the human representative on earth of the known cosmic will, however, his actions exhibit a natural sweetness and serenity, even in the midst of violence, that is foreign to Western drama. Meditation is often the means whereby the hero gains the power to achieve his goal. In more recent plays based on local history and on contemporary events, the assumption of cosmic harmony has been muted, and emphasis has shifted to depicting human conflicts —nationalist versus Western colonialist, modern daughter versus conservative parents, for example—that may or may not resolve happily.

Theme of the restoration of earthly harmony

Characters. Gods, demigods, kings descended from the gods, and princes and princesses are the heroes and heroines of traditional drama and dance. Powerful religious seers advise them, allies and ministers serve them, crude foreign ogres oppose them, and grotesque, slapstick clown-servants are their attendants. The clowns have been the subject of much speculation. Like the *vidūṣaka* clown of Indian Sanskrit drama, they are gluttons, practical and even cynical, and confidants to their masters' passions and weaknesses. Scholars have theorized that the chief Javanese clown figure, Semar, is derived from an ancient Javanese god who was deposed from his supreme position by the introduction into the drama of the later Hindu gods. In the midst of mythological plays, the clowns comment irreverently on political or social issues of the day, seemingly as spokesmen for the common man in an otherwise aristocratic world. Comic and serious scenes alternate.

Dramatic materials. A written script may be used as the starting point for performance, but usually actors, dancers, musicians, and stage crew improvise from a brief scenario. Specific musical selections are matched to certain kinds of scenes, characters, or actions, and standard movements for entrances and exits are known. Standard descriptive phrases of the kind common in all oral literature are used to introduce the hero and his kingdom, and more than a dozen types of recurring scenes are identifiable. A major interest in playgoing lies in perceiving the skill with which performers rearrange and subtly vary these familiar elements from play to play. Narrative commentary accompanying the dances often interprets a specific action in its broad context, thus helping to universalize the theatrical experience.

Improvisation by variation

Costumes, makeup, and settings. Costume and makeup have great importance in plays and dances. By means of elaborate systems of changing the cut, colour, and ornamentation of costume, the shape of the hairdress, the configuration of the crown, or the facial delineation and colour of masks, at least 300 different dance and dramatic characters can be identified. Doll- and shadow-puppet figures are carved according to similarly elaborate means of identification. Persons familiar with a dance or theatrical form can identify most characters by name or by type. Costumes, masks, and puppets may be works of art highly prized in themselves. Court and folk performances once used no scenery at all. Canvas scenery depicting stock scenes is now used by most popular troupes, but unfortunately it is often as inartistic as it is inexpensive. Only the Thai National Theatre, major troupes performing the popular *cai luong* drama in Saigon, and troupes performing in the Western tradition throughout Southeast Asia attempt to design three-dimensional scenery for each play.

ORIGINS AND DEVELOPMENT OF THE PERFORMING ARTS

Prehistory and links to the present. Knowledge of prehistoric performing arts is necessarily slight. That the performing arts were known and apparently widely practiced by the prehistoric peoples who had settled the mainland and the island archipelagoes is suggested by large bronze drums cast before the time of Christ, numerous pre-Hindu tribal myths in remote areas of the Philippines and elsewhere, masked dances of many types still performed by isolated tribes in Kalimantan (Borneo) and in New Guinea, and descriptions of music and dance by Chinese visitors beginning as early as the 1st century AD. Simple dances were almost certainly accompanied by rhythmic percussion sounds and probably by the tuned metal bars or gongs thought to be indigenous to Southeast Asia. Some scholars suggest that tribal ancestors, animistic spirits, and animals were represented, perhaps in shadow form. Whatever their nature, these were folk performances, in part religious rites connected with seasonal festivals and in part joyful entertainment.

A number of existing dances and dramatic forms show prehistoric links. In the *trott*, a Cambodian deer-hunting dance, masked dancers representing hunter, demon, bull, girls, and deer enact the ritual of a deer hunt to ensure its success in real life. The Dayak of Kalimantan perform a

The *trott*, *barong*, and *nat*

dance to exorcise sickness. The *barong* dance-drama of Bali is staged by a village in which malicious spiritual forces are believed to have gained dominance over protective ones. By enacting the stand-off battle between the protective Barong lion figure and the destructive Rangda witch figure, the village ritually restores an equilibrium between the contending forces. A local *nat*, or animistic spirit, of which there are 37 in Burma, can be invoked by the dance of a professional "spirit wife," or *natkadaw*, through whom the *nat* communicates with the living. A disputed theory holds that the shadow play began as a ritual in which the spirits of magically powerful tribal ancestors were called to earth, in their natural form as shadows or shades, for advice.

Spreading of styles. Between *c.* AD 100 and 1000, dance and drama in Southeast Asia were profoundly affected by the introduction of dance style and the vast Hindu historical epics of India. First in Cambodia, then in turn in Thailand, Laos, and Burma, the epic *Rāmāyaṇa* became the source of dance and shadow plays. In Java the *Mahābhārata* dominated, whereas in Bali and Malaysia both epics were popular. Indian influence, however, can be exaggerated. There is no evidence that Sanskrit play texts or written dramatic treatises such as the *Nāṭya-Śāstra* became known. Strong local performing traditions made it possible to assimilate elements of Indian dance and Hindu stories, and, in subsequent development, Southeast Asian dance and theatre grew ever farther away from Indian styles.

Copper inscriptions from Java identify clowns, actors, musicians, and possibly puppeteers in the 9th century, and epic literature of succeeding centuries contains numerous descriptions of shadow plays that were popular and emotionally gripping. By at least the 4th century, epic recitations were a part of the Brahmanic worship of ancient Cambodia. Carvings of the beautiful *apsaras*, or heavenly dancing girls, adorning the temples of Angkor attest to the importance of court dance in Cambodia between the 10th and 13th centuries.

Josephine Powell, Rome



Apsaras, heavenly dancing girls, bas-relief from Angkor Wat, Angkor, Cambodia, early 12th century.

Accidents of history often carried the performing arts across national boundaries. It is believed King Jayavarman II brought dancers and musicians from Java when he left there in 802 to establish the Khmer dynasty in Cambodia, and shadow puppeteers may have accompanied him as well. Another theory suggests that Cambodia received the shadow play from India by way of Malaysia, through conquest by a Malay prince in 1002. Accidents of war took Khmer dance (and perhaps shadow theatre) first to Laos, when in 1353 a prince who had been raised at Angkor established an independent Lao court at Luang Prabang. Next, it reached the Thai capital at Ayutthaya in 1431, when Angkor fell to invading Thai armies. These returned to their court with the Cambodian court-dance troupe, thereby beginning the traditions of Thai court dance and dance-drama. In 1767 the Thai court was captured, in turn, by the Burmese, who brought to Burma the Thai-modified Khmer dance and created Burmese court drama. By this time, also, Javanese shadow theatre had been taken by colonists to Bali and to Malaysia, from whence it later entered southern Thailand.

Migration of the shadow play

When Indonesia was converted to Islām and Chinese influence became strong in the northern tier of mainland states beginning in the 13th and 14th centuries, existing court dance and dramatic forms were scarcely affected. Instead, new Islāmic plays were devised in Indonesia and Malaysia for shadow presentation and for the doll-puppet theatre. Islāmic influence was very strong in Malaysia, however, and even such pre-Islāmic forms as the shadow play absorbed Islāmic prayers, characters, and themes. Bali was never converted to Islām, and its performing arts are thought to reflect, even today, an older tradition than is seen in Java.

Chinese performing arts came to dominate Vietnam during the 1,000-year rule of northern Vietnam by the Chinese. Long after the Chinese were expelled, Vietnamese kings patterned their dances and opera on Chinese models. In time, however, local Vietnamese melodies and stories took their place alongside those of Chinese origin; and play scripts, at first filled with Chinese loan words, were rewritten in more colloquial Vietnamese.

Popular theatre and Western rule. From the 19th century onward, the incursion of Western culture brought about a variety of developments. A steady decline in the power of the royal courts precipitated the death of court drama in Burma; the shifting of support for dance and drama from the court to national bureaus of education and culture in Thailand, Cambodia, and Vietnam; and the movement of the court dance-drama into the popular theatre tradition in Java. In every country, new popular forms of theatre were created. These were based on historical events, on Islāmic and Chinese stories (but romances rather than Hindu and Buddhist myths), on national heroes fighting colonial rule, and on stories about contemporary events. It was not Western drama that sparked the burgeoning of popular theatre, though these plays were largely spoken dramas interspersed with music and dances. Rather, it was more of an indirect response to colonial rule, which caused an upsurge of nationalist feelings, and to the rapid growth of cities that created large populations without access to either folk or court theatre yet eager for some form of entertainment.

DIVERSE NATIONAL FORMS AND TRADITIONS

Although most of the dance and dramatic forms of Southeast Asia are related at least in the distant past, except in Vietnam and the Philippines, they acquired a very distinctive national and local character over the centuries. An examination of a few of these myriad forms will provide a more precise picture of the dense texture of the performing arts in Southeast Asia.

Cambodia. Court performing arts that had flourished during the Angkor period (802–1431) almost ceased in the centuries following the fall of the Khmer dynasty. Whether there was an organized court life or not is uncertain because of the scarcity of records, but in the 18th and 19th centuries performances in Thai form were produced by the Thai rulers of the western provinces of Cambodia. At Phnom Penh a classical ballet troupe was established by the royal family in the 19th century. The government of the Khmer Republic (established in 1970) has continued to support this troupe and the dance school attached to it.

Court styles. The chief court forms are *nang sbek* shadow theatre, *lakon* female dance and dance-drama, and *lakon kawl* male masked pantomime. The puppets of *nang sbek* stand four to five feet in height, have no movable arms, and are manipulated from beneath by two fixed handles or sticks. The standing puppeteer either sways the puppet with his arms or he dances with it. In processional scenes, as many as ten puppeteers parade completely around the screen, front and back. An entire tableau may be carved on one puppet, including several figures, forest scenery, or palace buildings, as if to bring to life the epic scenes carved in relief on the temples of Angkor Wat. Two narrators alternate a slow chant with dialogue. During dance sections, the large *pi phat* ensemble, augmented by a large drum, is played. Only plays based on the *Rāmāyaṇa* are performed, and major puppet figures represent Rāma, his consort Sītā, the monkey Hanumān, and Rāvaṇa, a ten-headed demon king who kidnaps Sītā. Khmer peasant figures have been inserted as rustic clowns in every *nang sbek* play. Performance has religious significance, the gods being invoked and honoured, and a performance may be arranged to assure rain or to halt an epidemic. It is not certain when and how *nang sbek* originated, but it seems probable that it was taken to Thailand in the 15th century and then brought back. This would explain the details of costume and headdress of today's puppets that are in Thai style.

Puppet tableaus

The lithe *apsaras* carved in Angkor's stone show details of the *lakon* style of female dance, but neither these nor other records are evidence that their lively dance was used in relating the epic stories. The 19th-century Thai rulers of western Cambodia reintroduced *lakon* dance and dance-drama, which was indigenous to Thailand as well. At the same time, Thailand's male masked pantomime was brought to Cambodia, as far as is known for the first time, and it became known as *lakon kawl*. Both male and female dance-plays were translated into Cambodian. Recently, costumes and headdresses have been redesigned in the style of the Angkor carvings. The stories, music, dance, and dramatic styles of *lakon* and *lakon kawl* are much like their Thai counterparts.

Popular forms. *Lakon bassac*, performed by some 20 professional troupes in Cambodia, is a highly eclectic form. Musical selections, dances for female characters, and costuming are borrowed from court *lakon*. The form was created by Khmers living in the Bassac River region of Vietnam. Villains wear Vietnamese costumes and move with Vietnamese opera movements, an evidence of the historical conflicts of the two peoples. Chinese, *Jātaka*, or Khmer stories may be performed. *Pi phat* music alternates with Chinese–Vietnamese instruments and with the Western saxophone and piano. Prince Sihanouk, chief of state between 1941 and 1970, encouraged a few French dramatic productions, but such drama is scarcely known outside the Western-educated elite.

Thailand. Folk *lakon jatri*, *lakon nai* female dance and dance-drama, *khon* masked pantomime, and *likay* popular theatre are Thailand's chief performing arts.

Folk performance. *Lakon jatri* began in the south, when male dancer-sorcerers performed, in simple folk style, the *Manora* Buddhist birth story as a dance-play. A troupe of three players was usual. One played the beautiful half-bird, half-human princess, Manora; a second played the hero, Prince Suton; and the third, often masked, played clown, ogre, or animal as needed. Flute, bell cymbal, and drums provided the music. The full *Manora* cycle of plays, staged in a village in the open, could last for two weeks. Probably after the 14th century, some *jatri* troupes moved to the Thai capital, where they established commercial theatres and staged a new all-male drama, *lakon nok* (*nok*, "outside" [the palace]), that emphasized plot and an often obscene humour. Advances in dramatic form were accomplished by court writers of *lakon nok* between 1800 and 1909. *Likay* troupes succeeded and completely supplanted *lakon nok* troupes in the early decades of the 20th century, but such popular *lakon nok* plays as *Sang Thong* ("The Prince of the Golden Conch") are presented today in modified form by the Thai National Theatre.

Bawdy humour

Female court dance-dramas. The *lakon nai* (*nai*, "inside" [the palace]) female dance-drama of the court was created in the mid-18th century from a confluence of three previously separate elements: female court dance, the *lakon nok* drama, and the Javanese *Pandji* stories as subject matter. Romantic episodes from the long *Pandji* tale were ideal for staging in the elegant and delicate style of female court dance, accompanied by songs and the music of a large *pi phat* ensemble. In the unhurried court atmosphere, dance scenes lasted an hour or more, and dance figures might be repeated many times. In time, other stories came to be staged in *lakon nai* and were given other names, but the *Pandji* plays composed by the daughters of King Boromokot (1733–58), by Rama I (1782–1809), and by Rama II (1809–24) remain favourites. In this form, *lakon nai* was introduced into Cambodia within the past two centuries.

Masked mime. Until recent years, a Thai version of the Khmer *nang sbek* shadow play, *nang yai*, occupied an important place in court as a Brahmanic-related ritual performance of the *Rāmāyaṇa*. Thai scholars describe it as the source of *khon* masked pantomime, citing celebrations for King Ramathibodi II in 1515 that included a *nang yai* performance without puppets. Wearing heavy makeup, the puppeteers themselves danced the usual *Rāmāyaṇa* episode as narrators told the story and spoke dialogue. Later, masks took the place of makeup, the screen was eliminated, and *khon* was born. In present-day Cambodia, one troupe can perform both forms. A number of *lakon nai* elements entered *khon* in later years, so that today a *khon* performance mixes the vigorous, masculine *khon* with gentle *lakon nai* singing style and female dance. All of the Thai dance-drama traditions (*lakon jatri*, *lakon nok*, *lakon nai*, and *khon*) are taught at the Department of Fine Arts in Bangkok, and representative plays from them are staged, often mixing traditions, at the Thai National Theatre.

Development of the *khon*

Popular plays and puppets. The major popular theatre form is *likay*, which evolved in part out of *lakon nok*. It is now performed by more than 100 troupes in most parts of Thailand. Actors are skilled in improvising not only the dialogue and lyrics but also the plot of a play as well, weaving romantic scenes and fragments of *lakon nai* dance, set to *pi phat* music, into a story from a well-known *Jataka*, history, or court play. *Likay* plays are set to music of the Lao *khen*, a reed organ, in northeast

Marie Mattson—Black Star



Rāvaṇa, the demon king, fighting the white monkey Hanumān, in *khon* masked pantomime, Thailand.

Thailand. A type of shadow play called *nang talung*, in which a single, seated puppeteer moves small puppets of individual figures with movable arms, is very popular in southern Thailand. The performance technique undoubtedly came from Malaysia, while the plays and the identifying features of the puppet figures, mostly from the *Rāmāyaṇa*, are from Thai *khon* and *lakon nai*. A similar shadow play exists in Cambodia, suggesting that the form travelled from southern Thailand to Cambodia, perhaps in the 19th century.

Laos. From the time Laos became a kingdom in 1353, the performing arts at the relatively small Lao court at Luang Prabang followed those of the more illustrious courts to the south, Angkor in Cambodia and then Ayutthaya and Bangkok in Thailand. Today, Lao dancers study in Bangkok, and the style of dance, music, and drama of the Royal Lao Ballet, the only remaining court troupe in Southeast Asia, is almost identical with that of *lakon nai* in Thailand. It is usual to perform excerpts from the very long dance-plays, the staging of a full-length spectacle being beyond the means of the court at present. Male *khon* dance is known but seldom performed. A number of Lao folk dances are studied and performed by the royal ballet troupe.

Balladeers of Laos

Scores of popular troupes perform plays derived from Thai *likay* and set to the lively and melodic Lao folk song style known as *mohlam*. *Mohlam* balladeers, accompanied by the *khen* (a complex reed organ), have for centuries travelled the Lao-speaking countryside, which includes Laos and northeast Thailand, singing bawdy songs of physical love and weaving into their performance local gossip and bits from the epics and court plays. When *likay* troupes from Bangkok played in northeast Thailand, the *pi phat* music and court dancing were not popular, although the plays themselves were. Enterprising *mohlam* performers then set the *likay* plays to the familiar *mohlam* song style, thereby creating a new popular theatre form, *mohlam luong*, or "story *mohlam*." Of the *mohlam* troupes, a few large ones are located in major cities in the two countries, but most are small and travel from village to village, performing for a few days or weeks in each.

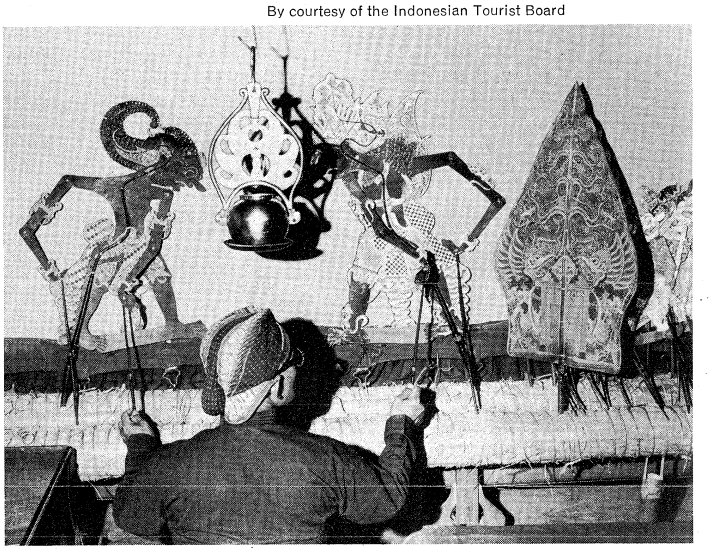
Burma. In spite of an old Burmese tradition of spirit dances stemming from animism and early contact with Indian culture, formal theatre did not begin until 1767, with the introduction of Thai *khon* and *lakon nai* to Burma following the capture and sack of Ayutthaya. Burmese courtiers and dancing girls immediately learned the two forms, and the plays were translated into Burmese. Because Rāma was viewed as a previous incarnation of Buddha, pious Burmese were reluctant to alter *khon* scripts. For a time *Jātaka* plays, including *Rāmāyaṇa* episodes, were forbidden to live actors. Instead, marionette troupes doing plays based on *khon* brought the Rāma stories to the Burmese countryside. But the *Pandji* plays were not considered *Jātaka*s, and even the first Burmese version, by U Sa under the title *Inao*, departed from its Thai model, thus setting the stage for the creation of court drama, or *zat pwe*, based on myth and legend but capable of being independently developed. The three *zat* written by U Kyin U (flourished 1819–37) portray the futility of political strife and urge a life of Buddhist renunciation. U Pon Nya (flourished *c.* 1850–66) created a freer form of dramatic verse, and his *Water Seller* is noted for its comparatively realistic treatment of court life.

Court drama ceased after 1866, when the British conquered Burma. Thereafter, drama was staged by professionals in public theatres, primarily in Rangoon. U Pok Ni in *Konmara* (*c.* 1875), U Ku in *The Orangoutan Brother and Sister* (1875), and others created a new type of drama, *pya zat*, that mixed royalty and commoners, emphasized humour, and added songs to appeal to a popular city audience. Hundreds of these works were published. Popular troupes in Burma today perform a long bill that lasts most of the night. It comprises songs and dances, a new contemporary play, and, as a final number, a classic *zat* in which remnants of old court music and dance are preserved. British touring companies in the late 19th and early 20th centuries brought examples of contemporary European melodrama and some classics to Burma. Subsequently a number of plays were written in Burmese and in English, following Western conventions and without songs or dance. Of these, *The People Win Through* (1950), by former Prime Minister U Nu, is an interesting example.

Indonesia. The sober, majestic, and profound court arts of eastern and central Java, where Javanese is spoken, include *wayang kulit* shadow theatre, *wayang orang* unmasked dance, and *wayang topeng* masked dance.

Shadow-puppet theatre. It is uncertain whether the shadow theatre is indigenous to Java or was brought from India, but the *wayang kulit* technique of having a single seated puppeteer who manipulates puppets, sings, chants narration, and speaks dialogue seems to be an Indonesian invention. Unlike most court arts, *wayang kulit* has had

By courtesy of the Indonesian Tourist Board



Wayang kulit (shadow-puppet theatre), Java.

centuries of performance in the folk tradition as well, so that today, with several thousand puppeteers active, it is the strongest traditional theatre form in Southeast Asia.

Plays are set in mythological times, some relating to indigenous animistic festivals and worship of local spirits, some directly dramatizing episodes from the *Rāmāyaṇa* and *Mahābhārata* epics, while the majority—the Pandawa (Pāṇḍav in Sanskrit) cycle of about 100 plays—are essentially Javanese creations in which the five heroic Pandawa brothers are placed in different situations. Three and sometimes four god-clown-servants and a set of ogre-antagonists who are not in the epics at all suggest how far removed the shadow plays are from the epics.

Conventions in puppet performances

The *wayang* puppeteer works within one of the world's most carefully organized performing arts, making possible a virtually solo performance without intermission, from around nine at night until the gray before dawn. Each play is in three parts, coordinated with three keys of music played by the gamelan ensemble. Certain standard scenes appear in a standard order, though some may be dropped. "Opening Audience" introduces the play's conflict, "Inner Palace" shows the king meeting his queen(s), and in "Outer Audience" the army is dispatched. In "Forest Clearing" the first battle scene occurs, and in "Foreign Audience" the antagonist kingdom, usually one of overseas ogres, is introduced. Concluding part one are "Foreign Outer Audience," in which the second army marches forth, and "Opening Skirmish," a battle scene between the two armies. The puppeteer chooses from among 150 musical selections, matched to scene type, character, mood, or action. The puppet figures are carved to indicate character type and status according to fixed patterns for nose, eyes, gaze, stance, body build, and costume. The puppeteer can choose one or another puppet of the same character, coloured gold or black or with a stern or relaxed countenance, to indicate the mood of the

figure in a particular scene. In battle scenes, he develops individual encounters between opponents, drawing upon a repertory of 119 movements that are classified for use by god, female, refined hero, muscular hero, ogre, or monkey. Formula narrative phrases describe famous kingdoms and characters, and battles are preceded by challenges couched in standard phrases. Although the puppeteer works only from a brief scenario, he is able to extemporize each performance, adding contemporary jokes for the clowns and molding the performance to suit the occasion and the audience. He and his supporting musicians and female singers are improvising within completely known, although exceptionally complex and subtle, artistic conventions.

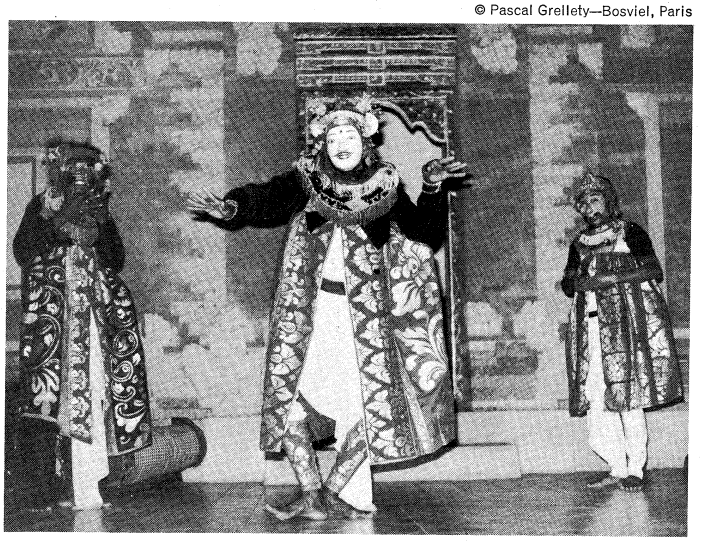
This artistic system, developed within the shadow theatre for performance of Pandawa plays, has proven to work so well that it has been widely imitated. The entire body of *wayang kulit* drama was adopted in Bali and in Malaysia. At least 25 other play cycles have been performed in Indonesia as shadow drama within this system, including the *Pandji* cycle (*wayang gedog*), Islāmic Amīr Ḥamzah plays (*wayang menak*), and plays dramatizing the revolutionary struggle against the Dutch (*wayang suluh*). The Pandawa *wayang kulit* repertory was transposed to the doll-puppet theatre (*wayang golek*) in Sunda, the western part of Java, and to dance-drama in eastern and central Java (*wayang orang*) and in Bali (*wayang wong*).

Performances are commissioned for special occasions and usually can be interpreted in religious or mystical fashion. There may be offertory plays at harvest time or animistic, ritualistic exorcisms protecting children from being devoured by the voracious god Kala. In *The Reincarnation of Rāma* the divine attributes of the god Wisnu (Viṣṇu or Vishnu in Sanskrit) reincarnate in Ardjuna (Arjuna), hero of the Pandawa cycle and ancestor of the Javanese race. The translucent screen can be interpreted as heaven, the banana-log stage as earth, the puppets as man, and the puppeteer as god, and the Pandawas can symbolize the manifold attributes of righteous behaviour.

Masked dance

Wayang topeng. Masked dance was also popular at the eastern Javanese courts (*c.* 1000–1400) and may be related to ancient animistic masked dance seen throughout the Pacific islands. Later, Indian dance style was assimilated, and sometime after the 15th century at the earliest, the *Pandji* story was dramatized. This is *wayang topeng*, widely performed as both a sophisticated and a folk art throughout Indonesia. Unlike the large-scale unmasked dance-drama, *topeng* dance focusses on interpreting character through solo dance.

© Pascal Grellety—Bosviel, Paris



Wayang topeng (masked dance), Bali.

Wayang orang. Java's spectacular dance-drama, *wayang orang*, grew out of the strong unmasked dance tradition that is illustrated in reliefs of female dancers carved on the 9th-century Borobuḍur and Prambanan temples in central Java and that produced the carefully cultivated female group dances of the Surakarta and Jogjakarta courts after their establishment in the 16th century. Of the latter dances, two stand out, the almost sacred *bedaja*, which even today is danced only in court surroundings, and the *srimpi*, in which two pairs of girls execute a delicate slow-motion duel with daggers and bows. In the middle of the 18th century, *wayang kulit*'s Rāma and Pandawa plays were set to court dance to form *wayang orang*, or "human" *wayang*. The music, narrative, and dramatic organization of the shadow play was kept largely intact, many of the actors' movements mimicking the stiff actions of the puppets, while new dance sections were added. Court performances stopped with World War II, but *wayang orang* continues to be performed by some 20 to 30 professional troupes in major cities. In popular performances, attractive actresses play the roles of such refined heroes as Ardjuna and humour and spectacle take precedence over dance.

Actresses as refined heroes

Ketoprak and ludruk. Two other types of popular theatre, *ketoprak* and *ludruk*, were performed in Java by 150 to 200 professional troupes. *Ketoprak*, created by a Surakarta court official in 1914, evolved into a spoken drama of Javanese and Islāmic history in which the clown figure is a spokesman of the common man. Whereas *ketoprak* is performed primarily in central Java, *ludruk*, a spoken drama that handles mainly contemporary subject matter, is performed in eastern Java by both amateur and professional troupes. Though *ludruk* is relatively realistic, male actors play all roles. Songs and dances, accompanied by gamelan music, are performed between acts in both forms.

Sundanese performing arts. There are three main performing arts in the Sundanese area of western Java. *Reog*, a kind of urban folk performance, can be seen especially in the streets of Djakarta: two or three men improvise popular songs, dances, and dramatic sketches for a neighbourhood audience in this type of entertainment. *Wayang golek* is a performance based on *wayang kulit* but using doll puppets without a screen. Approximately 500 Sundanese puppeteers perform *wayang golek*. Female singers, who are almost as important as the puppeteer, respond to requests and gifts of money by singing song after song and virtually stopping the play. *Sandiwara* troupes in Djakarta, Bandung, and a score of other cities perform both *wayang* stories in the form of Sundanese dance-drama and spoken historical and contemporary dramas for popular audiences. Sundanese-style court dances and *topeng* masked dances are often performed solo at festivals and for circumcision or wedding celebrations in private homes. Sundanese dance is more sensuous than Javanese and broader in style.

Balinese dance-drama. Of the many factors that have contributed to the remarkable flourishing of dance and drama on the island of Bali for more than a millennium, three are of particular note. First, Bali remained isolated from both Islām and the West. Second, there was a merging of folk and court performance styles into a single communal tradition appreciated by all. Third, dances and plays are indissolubly linked to the recurring cycles of local festivals and rituals whereby the well-being of the community is maintained against constantly threatening malicious forces in the spirit world. From the verve and brilliance of Balinese performances it is clear not only that the people like to perform but also that there exists some culturally determined compulsion to do so.

Balinese dance and dramatic forms are so numerous that only a few can be noted. Balinese villagers playing in the *barong* exorcism dance-drama are not merely actors exercising theatrical skills. The actors' bodies, going into a trance, are believed to receive the spirits of Rangda and the Barong, and it is the spirits themselves that do battle. Thus the performance is actually more a ritual than a piece of theatre. The *sanghyang* dance is usually performed by two young girls who gradually go into a state of trance as women sing in chorus and incense is wafted about them. Supposedly entered by the spirit of the nymph Supraba, the girls rise and dance, often acrobatically, though they have been chosen from among girls

Hypnotic possession of Balinese dancers

untrained in dance. The dance's purpose is to entice Supraba to the village to gain her blessing when evil forces threaten. In the *ketjak*, or monkey dance, as many as 150

Tor Eigeland—Black Star



Ketjak, or monkey dance, Bali.

village men, sitting in concentric circles around a flaming lamp, chant and gesticulate in unison until, in trance, they appear to have become ecstatically possessed by the spirits of monkeys. This performance, however, has no ritual function of altering an earthly condition.

That the Balinese *wayang kulit* may represent the older style of *wayang*, known on Java before the coming of Islām, is suggested by the less stylized shape of the puppets, by the shorter performing time of four to five hours, and by the simple music of only four *gender*, a bronze instrument similar to a xylophone with resonance chambers underneath, from the gamelan ensemble. In one type of shadow play having a special religious significance, the puppets perform before a screen during the daytime, and the puppeteer is seen in his role as a Brahman priest, bare to the waist. In the *redjang* processional dance, village women symbolically offer their bodies to their temple gods.

Because Balinese performing arts are vitally alive, they change from decade to decade, even from year to year. The *gambuh*, respected for its age, contains elements of dramatic dance, song, narrative, and characterization found in later forms. It is thought dull, however, and is seldom performed, though it is believed to have provided the model for the singing style of popular *ardja* opera troupes and the dance style of the lovely girls' *legong*. *Wayang wong* is analogous to the Javanese *wayang orang*, but masks are worn and the repertory is limited to Rāma plays. Pandawa plays are staged in identical style but are called *parwa*. It has been suggested that these forms also stem, at least in part, from *gambuh*. *Wayang topeng* masked-dance plays are ancient, being mentioned in a palm-leaf document of 1058. The Javanese chronicle of the Majapahit period (*c.* 1293–1520), the *Pararaton*, in which Ken Angrok is the hero, is a favorite *tapeng* story. This points to the strong influence exerted by Javanese on Balinese arts after the Majapahit court was transferred to Bali in the 16th century to escape Islāmic domination.

Malaysia. The Malay peninsula, in the geographical centre of Southeast Asia, has assimilated repeated intrusions of neighbouring cultures. The dances of the former princely states on the east coast show the influence of Indian nondramatic dance.

The multiform wayang. Rulers from Java in the 13th and 14th centuries and later large colonies of Javanese introduced their *wayang kulit* shadow theatre. The puppets of *wayang Djawa*, or "Javanese" *wayang*, are identical with the two-armed, long-nosed, highly stylized puppets of today's Javanese *wayang kulit*. Those of *wayang Melayu*, or "Malayan" *wayang*, have only a single movable arm and are less sophisticated in conception, which suggests that they are either descended from old Javanese puppets, before both arms were made movable, or are a degeneration of the more complex form. Rāma, Pandawa, and Pandji plays are staged. The puppets of *wayang Siam*, or "Siamese" *wayang*, though manipulated by a single seated puppeteer, represent a Thai conception of the figures from the *Rāmāyaṇa*; and costumes, headdresses, ornamentation, and facial features follow those of *khon*. The plays include Islāmic elements as well, while the chief clown figure, Pak Dogol, is thought to be a recent Malay creation that has supplanted Semar, the Javanese clown of *wayang kulit*.

In a performance, puppets of all types may appear together. Either such Thai instruments as the *lakon jatri* drum and small bell cymbals or gamelan instruments play the accompanying music. Song lyrics can be in ancient Javanese; animistic, Islāmic, and Hindu-derived invocations to the gods are offered in the Thai and Malay languages; and the play proper is in colloquial Malay. Puppeteers once performed throughout the peninsula, including the five Malay-speaking provinces of southern Thailand, but today puppeteers are found primarily in northeast Malaysia.

Multilingual performances

Chinese and popular entertainments. Chinese immigrants introduced various forms of opera during the 19th century. Troupes perform for Chinese Buddhist temple festivals, for local fairs, or on national holidays. In Singapore troupes occasionally perform in public theatres as well. Young people of Chinese descent in both Malaysia and Singapore have little interest in the opera, however, because their Chinese is limited. Occasionally troupes import star performers from Hong Kong or tour Chinese communities in Thailand.

Bangsawan was created by professional Malay-speaking actors in the 1920s as light, popular entertainment. Songs and contemporary dances were added to a repertory of dramatic pieces drawn from Islāmic romances and adventure stories. Troupes travelled to Sumatra, Kalimantan, Sunda, and Java, where their melodramatic plays found large audiences and influenced local performers of *sandiwara*, *ketoprak*, and *ludruk*. The cinema and television, however, have captured much of this audience.

Vietnam. An indication of the antiquity of the performing arts in Vietnam is a large bronze drum of the 3rd century BC found near Haiphong, in northern Vietnam, which is ornamented with instruments and musicians playing for dancers. Chinese performing arts presumably were a part of court life in northern Vietnam during the period of Chinese rule (111 BC–AD 939), and between the 10th and 13th centuries the dances and music of the Hinduized Cham peoples, living in what is now central Vietnam, were welcomed there. The melancholy Cham songs were particularly popular, and most authorities believe that the sad southern style of Vietnamese singing is derived from them.

Satirical drama. Hat cheo is a popular, satirical folk play of northern Vietnam that combines folk songs and dances with humorous sketches criticizing the people's rulers. Some scholars theorize that it is an indigenous folk art, whereas others, to show that it reached the people from the court, cite the legend of a Chinese actor who in 1005 was hired by the Vietnamese king to teach "Chinese satirical theatre" to his courtiers. *Hat cheo* is widely encouraged by the government in North Vietnam.

The opera. The classic opera, known as *hat boi, hat bo,* or *hat tuong*, is a Vietnamese adaptation of the Chinese opera long supported by kings and provincial mandarins as a court art and performed for popular audiences as well, especially in central Vietnam. The introduction of Chinese opera is attributed to the capture of a troupe of performers attached to the Mongol army that invaded

northern Vietnam in 1285. The actors' lives were spared in return for teaching their art to the Vietnamese. In 1350 another Chinese performer was engaged by the northern court as an instructor. Almost exclusively a court art in the north, *hat boi* was made a form of popular entertainment in central Vietnam by the playwright Dao Duy Tu in the 16th century. It was introduced to southern Vietnam under the Nguyen dynasty in the 18th and 19th centuries, but its future in both north and south has been jeopardized by the decades of war in the mid-20th century. The last large troupe of court musicians, dancers, and actors at Hue in southern Vietnam disbanded in 1945. Neither in the north nor in the south does the government provide *hat boi* with strong support, and the popular troupes lack audiences.

Blighting effects of war

In form and content, *hat boi* is a blend of China and Vietnam. Direct imitation of Chinese costume and acting techniques was encouraged under the reign (1847–83) of Emperor Tu Duc and it is probable that the present form of *hat boi* dates from this period. At Tu Duc's court in Hue, the playwright and scholar Dao Tan gathered 300 actors and with them wrote out texts of the standard repertory that previously had been preserved orally. He then had the texts published and distributed them to actors and troupe managers. In the 20th century there has been a movement to loosen the rigid structure of *hat boi* and to reduce the high proportion of Chinese loan words that makes the operas difficult for the ordinary Vietnamese to appreciate.

Following Chinese practice, the operas are classified as military or domestic. The former, which may be derived from Chinese and Vietnamese legend or history or may be purely fictional, concern struggles for power between kings. The Chinese novel *The Romance of the Three Kingdoms* furnishes material for many military plays. The latter, dealing with the lives of commoners, contain humorous scenes alternating with scenes of suffering that are played to the accompaniment of sad southern-style songs. The Confucian ethic of obligation to one's superior—of wife to husband, of son to father, or of subject to king—underlies plays of both types.

Hat boi staging is modelled on conventions of Chinese opera. Actors perform on a stage that is bare except for a table and two chairs. These can serve as a castle, a cave, or a bed as well as for sitting and eating. A single embroidered drop at the rear has an entrance right and an exit left. Costume and makeup indicate character type: black for boldness, red for anger or rashness, white for treachery, and gold as the colour of the gods. Conventionalized mime may be used alone or in conjunction with symbolic properties. The actor mimes stepping over an imaginary threshold or sewing without needle and thread, but he indicates riding a horse by gestures with a riding crop and travels in a carriage when a stage assistant holds flags with wheels painted on them at each side of his body. Percussion instruments accompany stage action, and songs—which may be in falsetto Chinese style, in soft southern Vietnamese style, or in a form of prose recitative—are accompanied by stringed instruments.

The popular stage. Southern-style singing is the basis of another type of theatre, *cai luong*, begun in the 1920s by popular singers who performed plays in which they sang the love lament *Vong Co*. Today, regardless of whether a historical or contemporary play is being performed as *cai luong* or which of the 75 or so current troupes is staging it, this melody will be heard throughout the play many times, underlying different lyrics. *Cai luong* stars are lionized, and the best troupes maintain high artistic standards. Among popular theatre forms in Southeast Asia, only *cai luong* plays are fully scripted and directed as they would be in the Western theatre. In contrast to the operetta form of *cai luong*, modern spoken drama is known as *kich*. It is a young dramatic form performed mostly by amateurs who are trying to put Western dramatic conventions into practice.

The love lament

The Philippines. Whatever indigenous theatrical forms may have existed in the Philippines, other than tribal epic recitations, were obliterated by the Spanish to facilitate the spread of Christianity.

The comedia. The earliest known form of organized theatre is the *comedia*, or *moro-moro*, created by Spanish priests. In 1637 a play was written to dramatize the recent capture by a Christian Filipino army of an Islāmic stronghold. It was so popular that other plays were written and staged as folk dramas in Christianized villages throughout the Philippines. All told similar stories of Christian armies defeating the hated Moors. With the decline of Spanish influence, the *comedia*, too, declined in popularity. Some professional troupes performed *comedia* in Manila and provincial capitals prior to World War II. Today it can still be seen at a number of church festivals in villages, where it remains a major social and religious event of the year. Much in the manner of the medieval European mystery-play performances, hundreds of local people donate time and money over several months to mount an impressive performance.

By courtesy of Philippine Embassy



Comedia, or *moro-moro*, folk drama based on the battles between the Christians and the Moros, the Philippines.

Styles from Europe. Dances and dramas from Spain were brought in, some of which took root. The "María Clara," a stately minuet, and the "Rigodón de Honor," a quadrille, were adopted by local European society for its formal balls. Spain's sprightly operetta, the *zarzuela*, became the favourite light entertainment in Manila and other cities. Professional *zarzuela* troupes continued to flourish in the early decades of the 20th century but had disappeared by World War II. New plays with original music were produced in profusion. A number of them based on topical themes and criticizing American colonial policies were banned.

Western drama is studied and widely performed in both English and Tagalog. There are no professional companies, but amateur university and community groups abound. Western classics and recent popular successes are staged, and in recent years many original plays have been written to celebrate the Filipino heritage.

EVALUATION AND STUDY

The most important resource for the study of dance and theatre is performance itself. Useful collections of puppets, masks, and theatrical costumes are in state museums in Bangkok, Phnom Penh, Kuala Lumpur, Djakarta, and Surakarta. The Department of Fine Arts in Bangkok and Indonesia's Habirando school for *wayang kulit* puppeteers in Jogjakarta and State Academy of Dance in Surakarta and Denpasar accept foreign pupils.

In recent decades, many plays have been published in Burmese, Thai, Javanese, and, in the Philippines, Tagalog and English. In these areas and in Bali there are manu-

scripts of plays and manuals for production that have still received little or no scholarly attention. Written primary sources in Laos, Cambodia, Malaysia, Sunda, and Vietnam are few. The need to know the language, literary traditions, music, and dance of the dramatic form, as well as its religious and cultural background, pose formidable difficulties for either the Asian or Western scholar. In European languages, theatre scholarship was begun in the last century almost as an afterthought by ethnologists and cultural historians. Only after World War II were most studies of the performing arts for their own sake begun. As yet, no single dance or dramatic form in Southeast Asia has been the subject of a comprehensive study that encompasses all aspects of its art. (J.R.B.)

V. Visual arts

GENERAL CONSIDERATIONS

Religious–aesthetic traditions. The visual arts in Southeast Asia have followed two major traditions.

Indigenous magical and animist tradition. The first is a complex inheritance of magical and animist art shared by the different tribal peoples of the mainland, where it evolved from Paleolithic origins, and of the islands. Such art gave the peoples who made it a sense of their identity in relation to the forces of their natural environment, to the structure of their society, and to time. It consists of types of potent emblem, mask, and ancestral figures broadly similar to those that hunters and early farmers the world over have used in connection with seasonal ceremonies, life and death rituals, and ecstatic shamanism (belief in an unseen world of gods, demons, and ancestral spirits responsive only to the shamans, or priests). The spiritual powers that the arts name and invoke are local and vary from group to group of the population. The rich formal artistic languages have been subject to successive episodes of influence from inland Asia, but each of the tribal groupings has developed its own artistic language on the basis of a common fund of Southeast Asian thought forms.

Indian tradition. The second major tradition was received from India during the early centuries of the Christian Era, when seagoing merchants from that subcontinent so fertile in ideas were expanding their trading activity. Into many parts of Southeast Asia—especially Burma, Thailand, and the coasts of Cambodia and Indonesia, where Indian traders settled and married into the families of local chieftains—they brought with them a script and literature in the sophisticated Sanskrit language. They also brought a highly developed conceptual system dealing with kingship, statecraft, and hydraulic engineering, integrated and authenticated by profound metaphysical ideologies of Indian pattern, both Hindu and Buddhist. These ideologies claimed to be universal, embracing all human diversity within a cosmic frame of reference. And this explains why the culture was adopted.

Why the Indian culture was adopted in Southeast Asia

For there was no Indian conquest of terrain; instead, the Indian conceptions, along with the art that expressed them, were used by dynasties in the colonial kingdoms as a method of overcoming divisions in their population, of centralizing effort, and of uniting their religions into viable states based upon cities. Although the new religious conceptions must have offered deep personal satisfaction to the general population, the architecture and sculpture in stone and bronze in which they were artistically expressed were expensive in materials, labour, and skill and were thus available primarily to patrons who were claiming for themselves a royal (*i.e.*, divine) status and using the resources of art to demonstrate that status.

The Indianizing traditions were continually refreshed by direct influences from India and Ceylon. There can be very little doubt that, during the early centuries after the birth of Christ, Indian artists and craftsmen travelled to work in the distant trading colonies of Southeast Asia, for they would have been needed to set up local traditions with proper formulae and methods. And there can be no doubt either that works of art made in India were continuously exported to the colonial kingdoms, thus keeping the local art styles in touch with developments "at home." It is also clear, however, that within a very short space of time the Southeast Asian kingdoms produced their own distinctive local versions of Indian styles; and some of their work shows skill, finesse, and invention on a colossal scale unrivalled even in India.

Although the art styles were to some extent sectarian, and sectarian partisanship played a part in political events, it was by no means unusual to find Hinduism and different forms of Buddhism flourishing side by side. In both Burma and Thailand, however, dynastic options were early exercised in favour of that particular form of Buddhism known as Theravāda (Hīnayāna), which adheres to the nontheistic ideal of purification of the self to Nirvāṇa. These countries follow the same form to the present day. It was also adopted in Cambodia and South Vietnam after prolonged and successful periods of Hindu and Mahāyāna (a theistic branch teaching compassion and universal salvation) Buddhist dominance. The strongly Sinicized population of the region around the Gulf of Tonkin, which pushed gradually down the coast of Vietnam to become the modern plains Vietnamese, began to adopt Theravāda Buddhism with its artistic types by about the 13th century AD, partly because this form could be best adapted to its self-contained and anti-dynastic cellular social structure.

Theravāda and Mahāyāna Buddhism

Relations between the two traditions. Even in those regions where Indian influence became strongly entrenched, the older layers of more primitive religion and artistic consciousness remained very much alive. Indian deities were readily identified with local spirits. The tribal populations retained, as many still do, their old animist customs, especially those connected with fertility and practical magic, often with an art (in perishable materials) in which to express them. These arts were influenced by and exercised a reciprocal influence upon the styles of officially imposed Indianized arts. In many parts of Southeast Asia, where the official Indian styles were not completely established (most of Borneo) or where they died out (colonies in Celebes), in inaccessible areas beyond the reach of dynastic influence, or on isolated islands, primitive styles have survived unmodified. Even in Indianized regions where a strict formula, say, for a necessary building type, had not been imported, a native pattern was adopted into the official canon (*e.g.*, Laos). In the Indonesian island of Bali, which has remained nominally Hindu, the Indian and the folk elements were thoroughly assimilated to each other, producing a quite individual style of both religion and art. In Sumatra and Java, whose populations were gradually converted to Islām from India during the 13th–16th centuries, the primitive cult of the ancestors was revived and encouraged by Muslim rulers, with folk versions of denatured Hindu art adapted to it. Decorative styles based on this art have flourished there and have been officially revivified during recent years. In the Philippines, notably in and around Manila, Spanish Roman Catholic art flourished after the Spanish colonization of 1571.

Artistic styles. The royal temple is the basis for the classic Indianizing styles of Southeast Asia. Each Hindu temple is centred on a shrine, symbolizing heaven upon earth, which is crowned by a roof tower representing the cosmic Indian mountain, Meru, conceived as the hub of creation. Since all the peoples of Southeast Asia already believed the natural habitat of spirits and gods to be a mountaintop, the Indian pattern was readily accepted. The temple usually stands upon a lofty terraced plinth (a block serving as a base), which itself also symbolized a mountain. Towered shrines could be multiplied on the terraces, though one of them remained the principal focus. Within the cell of this main shrine was a sacred image carved in stone or cast in bronze. The local Hindu ruler identified the subject of this image as his transcendent patron, or celestial alter ego. This was normally one of the Indian high gods, Śiva (represented perhaps by a phallic emblem, the *liṅga*) or Viṣṇu. In Mahāyāna Buddhist kingdoms a royal *bodhisattva* (a being that refrains from entering Nirvāṇa in order to save others) was sometimes adopted to fulfill the same role, a favourite form being known as Lokanātha, or Lokeśvara, Lord of the World. Subsidiary shrines, niches, or terraces

The royal temple

sometimes contained subsidiary images, including goddesses representing at the same time wives of the god and queens of the king. These images were worked in smooth, deeply rounded, and sensuously emphatic styles derived from Indian art but with varying inflections characteristic of each region and time. The whole exterior of the shrine was usually adorned with rhythmic moldings, foliage, and scrollwork, with figures representing the inhabitants of the heavens. Ideally, the building was constructed and carved in stone; but, particularly where good stone was not readily available (for example, in Burmese Pagan), it could also be brick, coated and sculptured with stucco after northeast Indian patterns. Temple complexes tended to grow as successive kings strove to outdo their predecessors with the magnificence of their buildings. Hindu rulers, influenced perhaps by vestiges of tribal custom, would sometimes retain their own family's temples and images while destroying those of earlier dynasties.

Buddhism, however, is a religion based on a doctrine of transcendent merit and sustained by an order of monks who have, ultimately, no vested interest in kings and gods. They may, however, take a great interest in the world of spirits and the operations of astrology, just as the local population does, even though they regard such matters as subordinate to the ultimate Buddhist aim of universal Nirvāṇa. Buddhist monasteries, therefore, tended to expand around stupas (domed monuments emblematic of the Buddhist truth, also called pagodas or *dāgabas*) of ever-increasing size and number; the preaching halls, libraries, and living quarters for monks were continually enlarged and repeatedly rebuilt, often as a testimony to the piety of royal patrons. Although, strictly speaking, Theravāda Buddhism has no place for a "divine ruler" whose identity an actual king may adopt, provision was made in legend and in court and monastic ritual for the ruler of a Theravāda country to assume a magical role as the dominant sponsor and patron of the Buddhist truth. His legendary prototype was usually not identified, therefore, with an icon of the enlightened Buddha but with images such as the chief disciple at the knee of the enlightened Buddha, as Prince Siddhārtha (the Buddha-to-be), or figuring in scenes of the Buddha's life that lined the monastery halls and corridors.

Prescriptive formulas for art

Both Hindu and Buddhist art were produced according to prescriptive formulas. If the formulas were not followed, the art was believed not to fulfill its transcendent function. In practice, however, there has been room for styles and types of image to change and develop fairly quickly. Hindu and non-Theravāda art recognized what could be called aesthetic values as a component in religious expression. Theravāda Buddhism, however, which might be called fundamentalist, has always attempted to preserve the closest possible connections with the Buddha's recorded original deeds and sayings; its art, therefore, has concentrated on repeating in its main Buddha figures the most exact possible imitations of authentic ancient images. This had led to a relative monotony of style in Theravāda icons (see below *Burma; Thailand and Laos*). In the subsidiary sculptured and painted figures, however, which illustrate scenes from sacred history, Theravāda art has had greater freedom of invention. In the 20th century, Theravāda Buddhism is the only form of Indian religion to survive in Southeast Asia, save for the modified Hinduism of Bali. Its architecture in recent centuries has been decorated with a vigorous, sometimes coarse fantasy and made gaudy with gilt paint and coloured glass.

General development of Southeast Asian art. Most of the works made under the inspiration of the primitive, magical, and animist tradition are in perishable materials such as wood. Because the climate is so hostile, the works that survive are relatively recent; and any that is even 100 years old generally owes its preservation to Western interest. There are, however, a large number of Neolithic stone implements and prehistoric stone monuments (megaliths), as well as bronzes, which provide a solid archaeological basis for interpretation of Southeast Asian primitive art.

For the art of the classic Indianizing civilizations, French archaeology played the major role in clearing, excavating, and reconstructing major sites in Cambodia, Laos, and Vietnam; Dutch archaeology in Indonesia; British in Burma. Old bronzes have been found in fair quantities; apart from those of the early Dong Son culture (see below *Bronze Age: Dong Son culture*), all belong to one or other of the Indianizing traditions. Many old brick and stucco buildings survive, notably the medieval work at Pagan and in central Thailand, though an enormous number are known to have perished. Little very old painting is known, save a few Indianizing medieval rock and wall paintings on plaster. In spite of the fact that Buddhist monasteries are able to act as agents for preserving their own artworks, most of the surviving Buddhist pictorial art on wooden panels or other fragile material is less than 300 years old.

The stone of dynastic buildings, of course, has survived far the best. Scholars thus know much more about Indianizing stone architecture, with its sculpture, than about any other Southeast Asian visual art. But, where good relief sculpture flourished, one can legitimately assume that vanished pictorial arts also flourished; and from details carved in stone and incised on bronze, as well as from the scattered enthusiastic references in Chinese sources, one can be sure that throughout their history the Southeast Asian peoples have been intensely creative and have lived their lives surrounded by a wealth of imaginative art in many different mediums.

Unexcavated sites

There are many sites yet to be discovered and excavated. Twentieth-century knowledge of the history of art in many parts of Southeast Asia, especially of important episodes in Burma, Thailand, and Sumatra, is still scantily documented.

Neolithic period. The earliest works in Southeast Asia that can be called art are the rectangular polished axheads of a familiar late Neolithic type that have been found at many sites in Malaya, Indochina, and Indonesia. Some of the later Neolithic (*c.* 2000 BC to early centuries AD) implements are extremely beautiful and polished with the greatest care. They include practical adzes and axes; but some, made of semiprecious stone, were clearly intended for purely ritual purposes. Even in the 20th century a few such blades are preserved and revered as sacred objects in certain Indonesian farming communities and similar objects have continued to be made in some very remote regions. These tools, with their fine edges, suggest that their owners were capable of very high quality woodworking and might well have decorated their wooden houses with designs of which we know nothing.

During the Neolithic Period, metal—both bronze and iron—came into use for implements but did not greatly alter the material culture. In many regions, notably Cambodia, Borneo, and Sumatra, numerous works of megalithic, or stone, art survive, including menhirs (single upright monoliths), dolmens (two or more upright monoliths supporting a horizontal slab), cist graves (Neolithic graves lined with stone slabs), and terraced burial mounds, all dating from the late Neolithic epoch. Some remarkable large stones are worked in relief with symbols and with images of animals and men, notably in the Pasemah region of Sumatra. Shaped stone sarcophagi and skull troughs (containers to hold the skulls of ancestors and of enemies at village shrines) are also known. These megalithic art objects suggest a highly developed cult of a spirit world connected with the remains of the dead (see below *Cambodia and Vietnam; Indonesia*).

Bronze age: Dong Son culture (c. 4th–1st centuries BC). By about 300 BC a civilization with elaborate arts based on bronzeworking existed, extending probably from the Tongking region into Laos, Vietnam, Cambodia, and Indonesia. This is called for convenience, after a major site, the Dong Son culture, though it may not have been a true cultural unity. A variety of bronze ritual works, many decorated with human and animal figures and with masks, were cast by the cire perdue method (metal casting using a wax model). The chief objects were ceremonial drums, large and small; the largest was

found in Bali and is called "the Moon of Bali" (see below *Indonesia*). Extremely elaborate bronze ceremonial axes were made—probably as emblems of power. Certain relief patterns on the bronzes suggest that "ship of the dead" designs, like those still woven in textiles in both Borneo and Sumatra, may well have been woven even then. The spiral is a frequent Dong Son decorative motif; later Dong Son art was probably responsible for transmitting—especially into Vietnam, Cambodia, and Borneo—versions of the contemporary Chinese Chou dynasty's asymmetrical squared-hook patterns.

1st to 10th century. During the 1st century AD, Indian influence began to spread through Southeast Asia in the wake of trade, both overland, through Burma and Thailand, and by sea traders settled at especially favourable spots along the inland roads and along the sea routes around the coasts and into the islands. Buddhism, which was particularly popular among the Indian merchant classes, took root at a large number of trading cities, where monasteries were set up under the patronage of local kings. Many fragmentary Buddha images based upon Indian types of *c.* AD 300–400 have been found in Burma, Thailand, and Cambodia, produced in the kingdoms of the Mon people, the chief of which, in Thailand, was called Dvaravati. By the 5th century the first Hindu kingdoms had been established in western Java and Borneo. These kingdoms produced dynastic cult images, fragments of which have been found.

The first Hindu kingdoms

Perhaps the most splendid of the earlier Indianizing kingdoms, lasting till the 9th century AD, was that of the Pyu people in the upper Irrawaddy River Valley. The Pyu were the people most directly in touch with eastern India by land routes. Only one of their enormous cities has been explored archaeologically (see below *Burma*). The remains of Buddhist buildings, east Indian Buddhist images, and Hindu sculptures of Viṣṇu have been found there.

In the 1st century AD the predominantly Hindu kingdom known as Funan (the name given it by Chinese historians) was established in Cambodia. It seems to have controlled an empire that included kingdoms in Malaya and even parts of southern Burma. Its population was probably Mon and shared the culture of the Mon in the lower Irrawaddy Basin. (The Funan kingdom really represents the earliest phase of what became, in the 9th century, the great Cambodian Khmer Empire.) Between c. 550 and 680 the kingdom retreated from the coast up the Mekong River into Laos, where it was called by the Chinese Chenla. This joint Funan–Chenla tradition produced some of the world's most magnificent stone cult images. Though Buddhist icons are known, these images principally represent Hindu deities including Viṣṇu, his incarnation Krsṇa, Śiva, and a combined Śiva-Viṣṇu figure called Harihara. The images were housed in wooden or brick shrines, now vanished.

During the Chenla retreat the Theravāda Buddhist kingdom of Dvaravati flourished in southern Thailand, on the lower reaches of the Mae Nam Chao Phraya; the kingdom lasted until the 11th century, when it was captured by the Khmer. What little of its art is known is close to that of eastern India and provided the basis for later Buddhist art in the Khmer Empire, as well as for some of the later forms of Thai art.

Almost contemporary with Chenla was the rise of the central Javanese kingdom. Soon after AD 600 the earliest surviving Hindu temples were built. In *c.* 770 the Śailendra dynasty began its long series of superb stonecut monuments both Hindu and Buddhist, which culminate in two enormous symbolic architectural complexes: the Mahāyāna Buddhist Borobuḍur (*c.* 800) and the Hindu Lara Jonggrang, at Prambanam (*c.* 900–930). These monuments were decorated in an individual and exceptionally accomplished style of full-round and relief sculpture. Many small bronze religious images have survived. The art of the Śailendra dynasty testifies to the imperial and maritime power of the central Javanese kingdom, which seems to have influenced politics and art in Khmer Cambodia. It also took over the possessions of a major Theravāda Buddhist kingdom called Śrivijaya, which had flourished in Malaya and Sumatra and was centred at Palembang. The Javanese Śailendra ruled most of Malaya and Sumatra and installed themselves there in the mid-9th century, when their home terrain in Java was taken over by the Mataram dynasty, heralding the eastern Javanese period, which began in 927. Śrivijaya, under Śailendra rule, declined in the mid-11th century, and most of its remains still await discovery.

The central Javanese kingdom

In Vietnam *c.* the 2nd century AD the predominantly Hindu kingdom of Champa was founded. Its capital was at My Son, where many temples have been found. This kingdom suffered much from attacks by the Chinese, and, after it began to lose the north to the Sinicized Vietnamese in 1069, the Cham capital moved in 1069 to Vijaya (Binh Dinh), in the south. There it was involved in continual warfare with the Khmer, who finally annexed southern Vietnam in 1203. The art of the northern Vietnamese as a whole has always been so strongly under the influence of China that it can best be characterized as a provincial Chinese style.

10th century to the present. In Cambodia the Khmer Empire succeeded to the old territories of Funan–Chenla. About 790 the first major Khmer ruler, Jayavarman II, who was related to the old Funan royal family, came to Cambodia from the Śailendra court in Java. In 802 he set up a religious capital on a hill at Phnom Kulen; he seems to have called in artists from Champa and Java, thus giving to Khmer art a distinct new impetus. At another site, Sambhupura (Sambor), he built temples with sculpture based upon the old Funan–Chenla tradition. At Amarendrapura, about 800, he built a brick pyramid—an artificial mountain—to support a quincunx of temples.

The Khmer Empire

It was Indravarman I (877–889) who laid the foundations of the fabulous temple complex known as Angkor. His plan was based on a rectangular grid of reservoirs, canals, and irrigation channels to control the waters of the river system. Later kings elaborated this original design to a colossal scale. Indravarman built the first great works of Khmer architecture: the Preah Ko, at Roluos, and at Angkor his temple mountain, the Bakong, ornamented with sculpture. Successive kings built their own temple mountains there, including the Bakheng (*c.* 893), Pre Rup (*c.* 961), the Ta Keo (*c.* 1000), and the Baphuon (*c.* 1050–66) and culminating in Angkor Wat, built in the first half of the 12th century by Suryavarman II. After a disastrous invasion by the Cham, Jayavarman VII undertook the most ambitious scheme of all, the Mahāyāna Buddhist Angkor Thom and Bayon (*c.* 1200). Thereafter, for a variety of reasons, including conquest by the Thai, no more large-scale work was done by Angkor and the country became Theravāda Buddhist. The modern dynasty has adapted remnants of traditional splendour, and the craftsmen of Cambodia have remained capable of work in the same vein as but often superior to the Thai.

Hindu Javanese art continued to be made under the eastern Javanese dynasties (1222–14th century), although their structures are not nearly as ambitious as the central Javanese works. There are many temple enclosures and volcanic bathing places with modest stonecut architecture. Some of the stone sculptures from these sites, however, are now world famous. In the 20th century the east Javanese tradition still survives, modified by folk elements, in Bali, to which the east Javanese Hindu kings retreated in the 16th century to maintain their religious independence in the face of Muslim expansion. Muslim monuments in the form of mosques and tombs are found in various parts of Indonesia. They adapt older forms of Indonesian art.

In 1056 the great Burmese king Anawrahta decreed Theravāda Buddhism to be the religion of his country, replacing earlier cults. He removed the Mon monks and artists from the capital of the old Mon kingdom in southern Burma, transporting them to his own northern capital, Pagan. There they built a city, with many large brick and stucco temples (pagodas) based on Indian patterns, that remains one of the most impressive sites in Asia. The Mongol invasion of 1287 put a stop to work there.

The Mon kingdom, Dvaravati, of southern Thailand,

was annexed to the Khmer Empire in the 11th century and Khmer imperial shrines were built there. After the decline of the Khmer and the Mongol invasion of 1287, a powerful alliance of racially Thai kings established the first major Thai empire, retaining Theravāda Buddhism as the state religion. Thailand was divided into two principal regions, northern and southern, with capitals, respectively, at Chiengmai and Ayutthaya, possession of the trade city of Sukhothai being an issue between them. In all the Thai cities, brick and stucco temples were built on variants of Indian and Burmese patterns. Many fine bronze Buddha figures, large and small, were cast in canonical Theravāda Buddhist styles. Most of these figures were accommodated in monastery halls built in impermanent materials.

In both Burma and Thailand a very large number of monasteries, usually surrounding one or two principal pagodas, were constructed during the later Middle Ages and into modern times. The major cities of Rangoon, Mandalay, and Bangkok contain the most elaborate examples, although there are many elsewhere. Because the pagodas were repeatedly enlarged and redecorated and the wooden monastic buildings and their many smaller stupas continuously reconstructed and renovated, no absolute chronology has been established for the arts of this epoch.

In Laos and Vietnam Theravāda monasteries, with brick stupas, were similarly built and rebuilt of wood. An outstanding stupa is the That Luang at Vientiane, in Laos, founded in 1566 but much restored in the 18th–19th centuries. In Vietnam local variants of Chinese styles were adapted during the Middle Ages to the planning and decoration of palaces and of Confucian, Taoist, and Buddhist temples.

The primitive styles that prevailed in the Philippines were modified by the conversion of various groups—the Moro people, especially—to Islām in the 15th–16th centuries. When, in 1571, the Spanish took control, Manila became the capital of a Spanish colony, and Roman Catholic Spanish art was adopted via Mexico. A local school of church architecture and figure sculpture flourished until the 20th century, when Manila became the centre of a modern commercial society, with its attendant architecture and art.

BURMA

One date is crucial in the art history of Burma: AD 1056. In that year King Anawrahta of Pagan decreed Theravāda Buddhism to be the state religion of all Burma. This signalled the unity of what had been a divided country, consummating tendencies apparent in earlier Burmese history.

6th to 11th century. The only major Burmese art known to scholars is based upon Indian and Ceylonese Buddhist art. In the period preceding Anawrahta's decree there had been three major historical eras in what is now the country of Burma, the first two of which produced Indianized art known to scholars only fragmentarily: the rule of the racially Mon kingdom of the lower Irrawaddy (*c.* AD 500–11th century), the contemporaneous dominion of the Pyu people in central and Upper Burma, and the subsequent decisive incursion of racially Burmese people from the northeast (11th century).

Major historical eras

The earliest concrete evidence of Indian culture in Burma is a Buddhist inscription from Prome dated *c.* AD 500. This and later inscriptions from the same area were cut probably in the western Mon kingdom, which followed Theravāda Buddhism and was confederated with the Theravāda Buddhist eastern Mon kingdom of Dvaravati (see below *Thailand and Laos*) in southern Thailand and part of Cambodia (AD 6th–12th centuries).

During this same period, in Upper Burma, the people called Pyu, speaking a Tibeto-Burman language and perhaps originating in Central Asia, built cities whose magnificence was known to contemporary compilers of the Chinese Tang dynasty history. In the 8th century one city was recorded as being some 50-odd miles (80 kilometres) in circumference, containing 100 Buddhist monasteries lavishly painted and decorated with gold and silver. The Pyu were in direct contact with northeast India, where various forms of Mahāyāna Buddhism, which embraced philosophies and rituals unacceptable to the Theravāda, flourished; their Ari priesthood was later proscribed by Anawrahta. Their capital city, Śrī Kṣetra (modern Hmawza), which was once larger than even Pagan or Mandalay, has been partly excavated. Three huge Buddhist stupas—one 150 feet high—survive there. They illustrate the pattern from which all later Burmese stupas were developed. Enshrining revered relics of Buddhist saints, they consist of tall, solid brick cylinders mounted on shallow, circular, stepped plinths and crowned by what was probably a tapering bell-like pinnacle. Other excavated halls, one on a square plinth with four entrance doors, follow Indian examples. A few Hindu fragments survive as well.

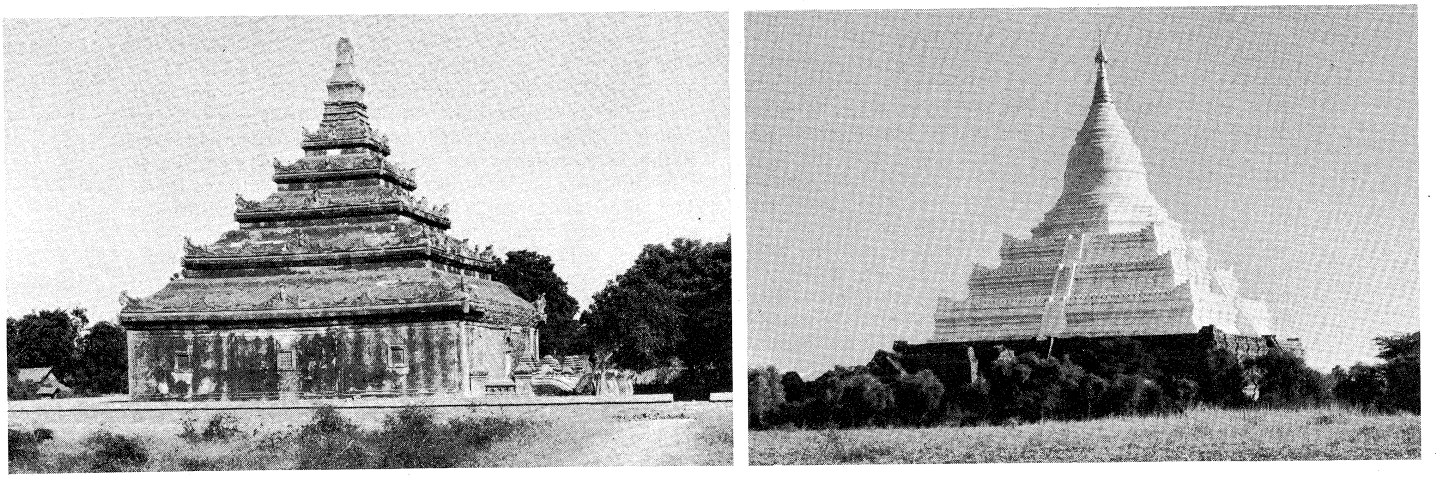
The Pyu were conquered by a neighbouring kingdom, probably before AD 900. During the following century their terrain and cities were infiltrated by the racially Burmese people. These people were of common tribal stock with the Thai and northern Vietnamese and were probably on the move under pressure of the Chinese colonization of their home terrain around the Gulf of Tonkin. They were converted to Buddhism by the Pyu and later by the western Mon; but they never completely abandoned their own original cult of nature spirits, known today as the *nat*s. The *nat*s are a mixed collection of the spirits of trees, rivers, snakes, ancestors, and people who have met a violent or tragic death; they act supernaturally, each according to its character. Their number was fixed at a canonical 36, the Buddha being included as the 37th. The *nat*s were worshipped with orgiastic ceremonies and trance rites of spiritual possession. Certain mountaintops were sacred to them. Even in the 20th century the *nat*s exert a powerful influence on the lives of the ordinary people; every village has its own *nat* house—a fragile pavilion built into a tree, after the pattern of the tribal house, and adorned with shreds of coloured cloth, glass, and other offerings. The Buddhist temple in Burma is conceived essentially as an enormous *nat* house, a section of the domain of the spiritual located upon earth. And, since the Buddha was adopted as the last and greatest of the *nat*s, the same symbols of supernatural splendour that adorn the *nat*s adorn the Buddha's images, and a *nat*-like spirituality attaches to the ubiquitous monks in whom the presence of Buddhism is experienced as an everyday reality.

The *nat*s

11th century to the present. When King Anawrahta came to the throne, he is said to have captured an old Mon city (perhaps Thaton or Nagara Pathama, now in Thailand) and carried off its royal family, many skilled craftsmen, and most of the Theravāda monks to his own northern city of Pagan. The king recognized the superior culture of the Mon captives; he established their main form of Buddhism by decree and gave them the task of organizing and civilizing the new united Burmese kingdom and producing for it a Buddhist art. Under Anawrahta's successor, links with the Buddhist homeland were forged. Embassies were sent to Bodh Gayā, in Indian Bihār, and the great Mahābodhi temple there—marking the spot where the Buddha achieved enlightenment—was restored with Burmese money and somewhat in Burmese taste. A smaller copy, with its large rectangular block crowned by the characteristic pyramidal, storied tower, was built at Anawrahta's Pagan. It is here that the greatest achievements of western Mon art—a splendid profusion of architecture and decorative work—are probably to be found. After 1287, when Burma was sacked and garrisoned by the Mongols, new construction at Pagan was virtually abandoned.

Pagan

In Pagan (founded *c.* 849), architecture is the dominant art; except for the big brick icons, mostly ruined, sculpture and painting play a subordinate role. Pagan contains the largest surviving group of buildings in brick and plaster of the many thousands that once stood in various parts of South Asia. The remains at the site, all religious buildings of one kind or another that must once have been surrounded by dense building in perishable materials, are in varying states of preservation. The inscriptions they



Burmese architecture.
(Left) Library, Pagan, *c.* 1058. (Right) Shwesandaw *cetiya*, Pagan, 11th century.
Louis Frederic—Rapho Guillumette

bear indicate that royal devotees often turned their palaces over to religious use; so it is likely that palace and monastic architecture were very close in style. A few structures still stand that belong to the period before Anawrahta, some of them inspired by Mahāyāna Buddhism and one—the Nat Hlaung Gyaung (*c.* 931)—by Hinduism. Flanking the Sarabha Gate is a pair of small *nat* shrines with pointed, open windows—the earliest in Burma, perhaps 9th century.

The library, built about 1058 to house the books of one of the Buddhist monasteries, is one of the most important buildings in Pagan. It is rectangular, with a series of five stepped-in, sloping stone roofs crowned by a rectangular tower finial. The concave contours of the roofs are characteristic of much Burmese architecture. The eaves and corners of all the tiers are adorned with the typical Pagan flame ornament, or antefix.

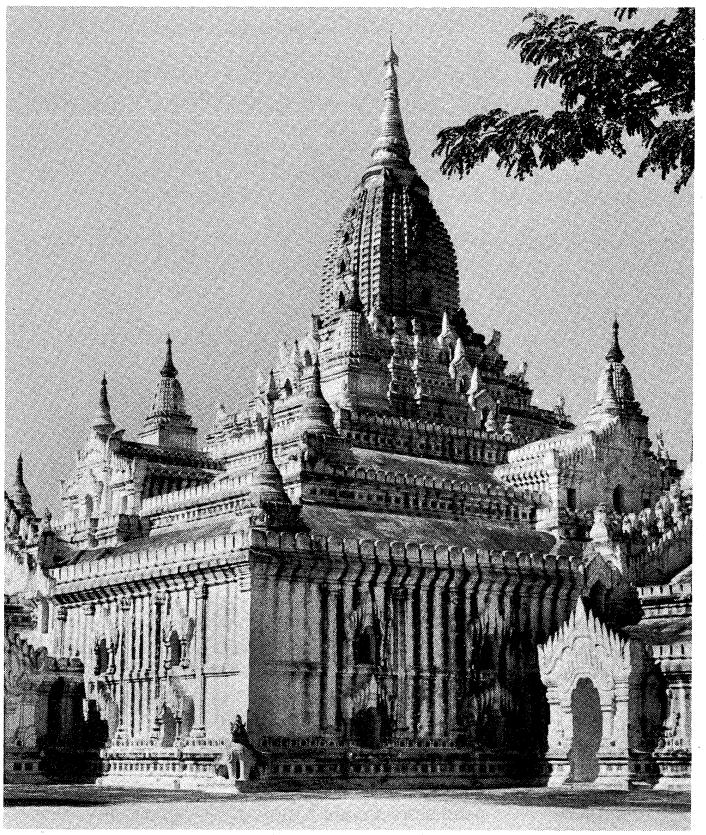
The *cetiya*

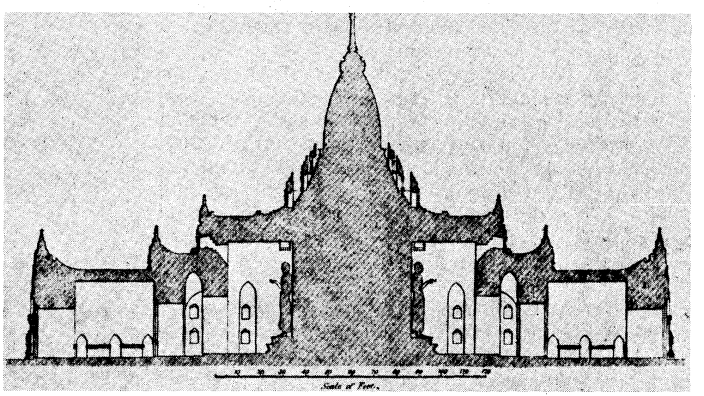
There are other buildings of the same general type among the ruins of Pagan. Far the most numerous and important, however, are the buildings—called *cetiyas*—that combine the attributes of stupa and shrine. These have a history and a line of evolution of their own, which can be traced from the Pyu stupa to the huge structural temple. The normal stupa, derived from the early medieval Indian form, is a tall structure consisting of a solid dome set on a tiered square plinth (often with miniature stupas at the corners) around which the faithful may perambulate. The dome is surmounted by a *harmikā*, which resembles the small railed enclosure found on the oldest Indian stupas. In Burmese stupas, however, the *harmikā* becomes a decorated cubical die, above which is a circular pointed spire; in memory of its distant origin in India, the spire is horizontally flanged (rimmed) with moldings in a series of honorific umbrellas of decreasing size. In later practice *harmikā* and umbrella spire become a single architectural unit. The Burmese stupa dome, based on the tall, cylindrical Pyu prototype, has a spreading concave foot resembling a bell rim. The Lokananda and Shwesandaw at Pagan are two well-known examples. Because in recent times they have been coated with plaster, the finely detailed brick carving characteristic of early Pagan architecture has been obscured. Such carving is beautifully exemplified in the Seinnyet temple at Myinpagan (11th century).

Anawrahta's type of *cetiya* followed the general form of the early Pyu stupa. The main point of evolution was in the progressive elaboration of the terraced plinths on which the dome stands. The plinths became virtually sacred mountains, with a series of staircases running from terrace to terrace up each of the four sides. Perhaps inspired by vanished work in contemporary late-11th-century India, the Burmese began to open up the interior of the terraced base of the stupas with wide corridors and porticos, converting it into a roofed temple. The cylinder of the stupa dome was carried down through this temple space to its floor. Four large Buddha icons were added to the lower part of the dome, facing the four directions. Once this conception had evolved, it was possible to create around the central stupa a broad circuit of roofed enclosures, which from the outside would still suggest the traditional pattern of the stupa standing on its raised terraces, while the interior could be used for ceremonial, as in a true temple. Sculpture and painting, decorating the internal halls, corridors, and doorways, recounted the life of the Buddha and presented the example of his previous virtuous incarnations. The most famous example of this type of *cetiya* is the great Ananda temple at Pagan (dedicated 1090). It is still in use, unlike most of the old temples there, and so is kept in repair; it is painted a blazing white with lime stucco—which has, of course, obscured the finer detail of its old architecture. Its plan is square, with a broad, four-pillared porch hall added to all of the four doors in the four faces of the square. Its tower is a curvilinear pyramid resembling eastern Indian Hindu temple towers, and its enormous brick mass is pierced with two circuits of vaulted corridors. The sloping, curved terrace roofs have an elegant overall concave profile and flame antefixes along all the eaves.

As time went on, Burmese brick and stucco architecture developed principally through the stiffening of masses into rectangular blocks and through the elaboration and often the coarsening of its ornament. The 13th-century Gawdawpalin temple at Pagan, for example, consists of a rectangular hall with a large closed entrance porch; the hall is surmounted by a tall but narrow second story whose decoration repeats that of the lower story; the whole building is crowned by a four-faced tower with a curved profile. Multiple moldings and decorative motifs are used as outlining elements and the doors are framed in elaborate upward-flaring hooded porches.

Until the Mongol conquest in 1287 much excellent work seems to have been done at Pagan. It is, however, impossible to form an adequate idea of the older styles of temple architecture at other sites in Burma, such as Rangoon or Mandalay. Whereas most of the temples of Pagan were abandoned early on, so that even though they may be ruined they show their original characteristics, temples in modern cities have been repeatedly and drastically restored. Old stupas may have as many as eight successive casings of brick and stucco; temple walls and doors are constantly torn down and rebuilt; and stucco surfaces may be renewed almost annually. Such attentions to a religious building are popularly regarded as acts of merit; thus, revered architectural monuments suffer continually from well-intentioned but disastrous renovation. At the big stupa sites huge numbers of pagodas are constantly falling into decay and new ones are being built at great speed. Among them are variants, whose evolution cannot at present be traced, on the basic pattern of the long tapering bell, with a variety of transverse moldings, standing perhaps on a recessed plinth. Many are covered quickly with extravagant and gross stucco ornament. Or-





(Top) Ananda temple, Pagan, Burma, dedicated 1090. (Bottom) Section of the Ananda temple.
(Top) Louis Frederic—Rapho Guillumette, (bottom) from P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London

nate flaring porches and flame finials are added to gates, wall ends, and eaves corners. A tapering slenderness is the outstanding characteristic of all the different types.

Stupa sites in Rangoon and Mandalay

The monastic architecture—patterned on the hall, with its elaborate doors—that surrounds the great stupa sites of Rangoon and Mandalay is mainly in wood, built by simple pillar and architrave construction. The roofs are steeply gabled, with multiple gables riding over each other on immense carved pillars in the larger halls. The angles between pillar and architrave and the edges of roof gables, tiers, and terraces are filled with flamboyant cartouches (scroll-shaped ornaments) of pierced work, often lacquered and gilt; thus, the whole building may be smothered in repetitive ornamental curlicues. All this ornament has an otherworldly or spiritual significance. Other stupa sites in Burma, where less money has been spent and less ornament added to the buildings, may be more beautiful to the modern eye, with only a few flamboyant antefixes pointing the gables and punctuating the eaves. All over Burma similar buildings can be found; but, while many have been listed, they have been scantily surveyed, and no real study of their complex history has yet been attempted. There may well be a substantial Chinese influence in the construction of some of the wooden halls and pavilions.

Paintings and sculpture in Theravāda Burma do not seem to have reached the same heights of achievement as in other countries of Southeast Asia. They do not show the same originality and sense of life. The temples of Pagan contain the best examples, although even these are highly schematic, reminiscent in design of eastern Indian Buddhist manuscript styles. A number of early buildings at Pagan contain fragmentary terra-cotta (fired clay) reliefs or scraps of wall painting whose individual figures display some of the sensuous charm of their Indian prototypes (it is quite likely that Indian artists worked there). The 12th-century terra-cotta panels from the elaborate facings of the Ananda temple, however, show the beginnings of the petrifaction that overtook later Burmese figurative art. Both in reliefs and in wall paintings, the figure compositions are reduced to schematic groups of the minimum number of standard human and celestial types needed to tell a moral story, without any infrastructure of significant form and execution. The colossal Buddha images enshrined in the temples were usually built of brick and finished in stucco, gilded and ornamented. Such work is still done at high speed today. The technique is not a flexible one, and the emphasis of Theravāda Buddhism on the exact imitation of ancient Buddha images gave the Burmese no aesthetic incentive to develop the expression of their figures or compositions. The repeated heavy gilding and repainting of older icons has almost entirely destroyed any formal vitality they may once have had.

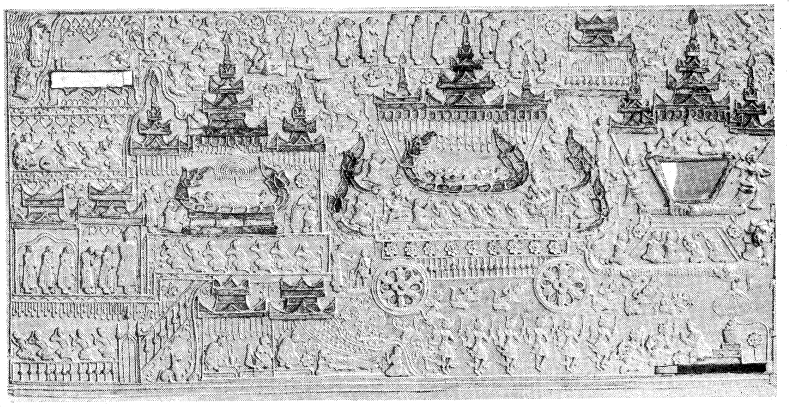
From about 1700 to 1850 Burma excelled in the decorative arts, whose forms continually recall those of Theravāda Ceylon. Burmese woven silks and embroideries are well-known. The carved wooden screens, panels, and brackets used inside temple halls, many devoted to representing the *nat*s and the population of the spirit world, have benefitted from being outside the strict canon of Theravāda Buddhist orthodoxy. The figure types follow fluid, slightly "boneless" conventions derived from classical Indianizing dance postures. The decorative goldwares and silverwares, which use much stereotyped decorative scrollwork, are also based on standard Indianizing iconography. Perhaps the most aesthetically satisfying works of the Burmese sculptors are the reliefs ornamenting the *sūtra* chests that were used in monasteries to store the sacred texts of Buddhism. The gilt gesso (paste used for making reliefs) facings of these chests carry the schematic style of relief sculpture beyond its normal aesthetic limits, by the way they are compelled to set off their figures as an intelligible scheme of thin raised lines and inlay against a plain ground. The forms of the fine lacquer bowls and boxes used in monasteries, decorated only superficially with painted ornament, show the underlying formal sense of the Burmese at its clearest.

Regional Burmese art

Interesting regional types of Burmese art are those of the Shan and Karen peoples, who live in the relatively remote northern hills. These areas have often produced extremely beautiful types of domestic and religious architecture, made of wood, on stone bases. They are a simpler and more austere version of the ancient pattern that underlies the halls and pavilions of the more sophisticated recent southern temple buildings, with their steep, gabled roofs. The peoples of the north also produce a variety of decorative arts. Notable among them are the textiles, which are characterized by banding, checkering, and triangular counterchanging of brilliant colours set off against black. The woven shoulder bags, particularly, are well-known in the West.

THAILAND AND LAOS

Dvaravati Mon kingdom: 6th to 11th century. Archaeology has recovered in central Thailand substantial glimpses of the magnificent early layer of Indianized culture, which includes a religious art produced between the 6th and 11th centuries by the eastern Mon kingdom of Dvaravati. The art predominantly served Theravāda Buddhism. Remains of Dvaravati architecture so far excavated include stupa bases: notable



Front view of a Burmese *sūtra* chest, wood and gesso, probably 18th century. In the Gulbenkian Museum of Oriental Art, the University of Durham, England. 63.5 cm × 120.6 cm.
By courtesy of the Gulbenkian Museum of Oriental Art, the University of Durham, England

examples are the Wat Phra Meru in Nagara Pathama (Nakhon Pathom) and others at Ku Bua and U Thong, some of which have elephants supporting their bases, following a pattern from Ceylon. The plinths of Buddhist assembly halls, which existed near the solid monumental structures, have also been discovered. Many terra-cotta and stucco fragments of decorative surface designs and celestial figures have also been found. The Wat Pra Meru, on a plan similar to that of the Ananda temple at Pagan in Burma (see above *Burma*), probably antedates the latter's foundation (*c.* 1090). It is likely that many other ancient monuments are encased in later stupas still in religious use, for it was probably customary not to destroy an old sacred monument but to encase it in a new shell, maybe several times over, and perhaps to construct a small external replica of the encased original alongside.

At many sites, especially Lop Buri, Ayutthaya, and U Thong, fine Dvaravati sculptures have been found among the architectural remains. Particularly important are the seated and standing Buddha figures in stone and bronze. Many of the faces have characteristic Mon features, with lips turned outward (everted) and downward-curved eyelids marked by double channels. Some of these Dvaravati images may well have furnished models for later Khmer art in Cambodia.

Dvaravati sculpture

Dvaravati sculpture shows close relations with several Indian styles, notably those of Amarāvatī, Gupta, post-Gupta, and Pāla Bihār. It also was probably influenced strongly by the art of the enigmatic kingdom of Śrivijaya in Sumatra, as well as by central Javanese types (see below *Indonesia*). One outstanding masterpiece from Chaiya, of Dvaravati date, may well be a work produced in Śrivijaya. It is a bronze torso and head of a *bodhisattva*, for which a mid-8th-century date is suggested. The body and face are modelled with a plastic and delicate sensuousness; and the elaborate necklaces, crowns, earrings, and armlets are beautifully chased (decoratively indented by hammering). The Śrivijaya origin is made more likely by stylistic reminiscences of the sculpture of contemporary Indonesia, which was also under Sumatran inspiration.

Khmer conquest and Tai immigration: 11th to 13th century. In the 11th century Dvaravati was captured by the Khmer of Cambodia and became a province of their empire. A number of Khmer shrines, probably intended as focuses of the Khmer Hindu dynastic cult, were built in Siam (Thailand). At Phimai (Bimaya) was the most important full-fledged Khmer temple, where one of the personal cult statues of the Khmer king Jayavarman II (see below *Cambodia and Vietnam*) has been found, together with bronze images, some of Tantric Buddhist deities. At Lop Buri the Phra Prang Sam Yot is perhaps the best surviving example in brick and stucco of Khmer provincial art in Thailand, its tall towers having complex rebated (blunted) corners and its porticoes high, flamboyant pediments (the triangular space used as decoration over porticoes, doors, and windows). Wat Kukut, at Lamphun, built by a Dvaravati Mon king *c.* 1130, represents an adaptation of the Khmer stepped-pyramid temple base as pattern for the temple itself. The niches on its terraces are filled with images in a deliberately archaistic revival of the old Mon style.

Domination by the Tai peoples

During the period when the Khmer were taking over the southern Mon region of Thailand, the northern region was falling under the domination of immigrant racially Tai peoples. The Tais were a branch of the migrating population who invaded Burma as the Burmese and of the Sinicized Vietnamese who were then pushing southward into what is now Vietnam. The Tais seem to have professed an animist nature religion, resembling the early form of the Burmese cult of the *nat*s (see above *Burma*). This whole group of peoples originated most probably as a tribal population in the region of Tongking and Canton. In the course of their southward migrations they probably played an important role, as yet unclear, in a kingdom called Nanchao, in what is now the Chinese province of Yunnan. The rulers of this kingdom seem to have followed a Mahāyāna form of Buddhism, including the cult of a *bodhisattva* as personal patron of the king. Several smallish bronze icons of a *bodhisattva* with a nude torso and a strap round the upper belly are known from Nanchao, in a style reminiscent of the later Pallava art of the east coast of peninsular India. The date of these images is still uncertain. Tai kingdoms were gradually established further and further south. Some of their tribes gained experience of administrative techniques by living within the boundaries of the Khmer Empire, with their own chieftains under Khmer officials. When the Khmer power was broken in the 13th century, the Tai moved into central and southern Siam, intermarrying with the Mon.

The Tai people normally built in perishable materials, wood and bamboo in particular. Their animist religion, which has no canonical group resembling the Burmese *nat*s, is still very much alive today. The spirits of trees need to be pacified, and the ancestors can be powerful helpers. Shamans, in a state of trance, make contact with the spirit world to perform good or evil magic. In the wooden high-gabled houses of the northern Tai (Chiengmai province), even today, ornate lintels are carved with floral relief designs to sanctify and potentiate the inner domestic part of the house where the domestic spirits live. The animist religion gave ground partially to Buddhism, which was gradually assimilated among the people, and at some date, as yet uncertain, was adopted by the greater Tai kings as a dynastic religion. With the spread of Buddhism a special religious architecture in brick and stucco was established.

Division of Siam

The Thai kingdom: 13th to 17th century. During most of its history, Thailand has been divided into two fairly distinct regions, a northern and a southern, the capital of the north at Chiengmai, the capital of the south at Ayutthaya. Between the two lies the great trade-route city of Sukhothai, possession of which fluctuated between the north and the south. Sukhothai seems to have been the principal focus and source of Buddhist culture in Siam, for it retained direct touch with Ceylon, which,

after the decline of Buddhism in India in the 12th century, became the principal home of Theravāda Buddhism. By the 15th century the difficult art of casting large-scale Buddha figures in bronze had been mastered in the north of Siam, as well as in the south.

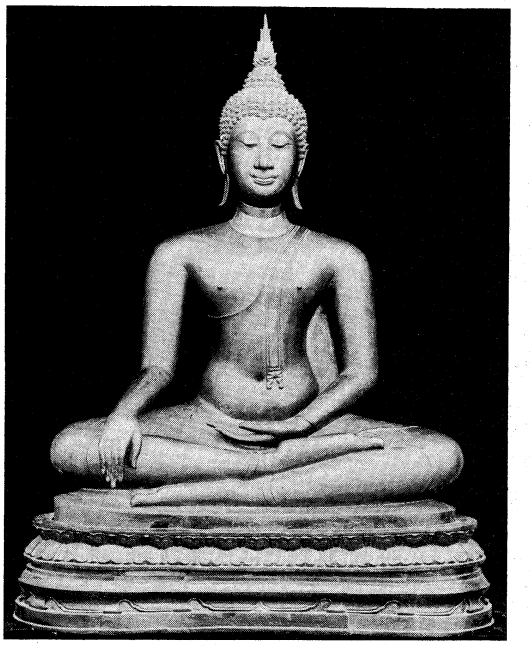
Sculpture. The Thai kings made repeated attempts to "purify" their conservative Theravāda strain of Buddhism, importing patterns of art along with texts and learned monks from Ceylon and trying to wean their people from worship of the spirits. To retain the greatest spiritual potency, Buddha icons in Thai temples had to be as close in type as possible to a great original prototype that Buddhist tradition erroneously believed had been made during the lifetime of the Buddha; in practice, this meant the types the local craftsmen knew as the oldest and most authentic. There were at least three major successive efforts by Thai kings to establish and distribute an "authentic" canon for the Buddha icons, which were their prime artistic concern. Each type that became canonical and was known to be magically effective was imitated repeatedly. For it was regarded as an act of merit simply to multiply images of the Buddha, whether they were to be installed in temples or not; hence, in addition to icons, enormous numbers of small images—made of many materials, from bronze, silver, stone, and wood to terra-cotta—were kept in temple storehouses. The images followed canonical patterns established for the major temple icons.

Since their work had to be as similar as possible to the oldest sacred images of which they knew, the Buddhist sculptors in Siam adhered to strict formulas and diagrams; artistic development was never a part of their purpose, though of course gradual change did occur. There is no tradition in Theravāda Siam in any way resembling the traditions of Mahāyāna art in, say, Cambodia or Indonesia, which encouraged artists to explore the possibilities of their mediums to express developing religious conceptions. Thus, Thai Buddhist sculpture consisted almost entirely of careful repetitions of the standardized types, which tended naturally, despite the artist's desire to capture an authentic sense of style, to lose their older vitality. It also happened that the three main canonical patterns often lost their individuality, blending into each other. Perhaps the best works were made in the 15th century, but work of high quality was still being done in the 16th and early 17th centuries.

Canonical types of Buddha image

The first canonical types were the Sukhothai, which seem to have been evolved in the trade-route city of Sukhothai as an attempt to capture the quality of early-medieval Ceylonese images and elements from Dvaravati sculpture. The developed versions of these types are marked by an extremely smooth, rounded modelling of the body and face, without any clearly defined planes. The outlines of hair, eyebrows, lips, and fingers are elegantly recurved, or S-curved, and the head is crowned by a tall, pointed flame finial. The entire figure gives an impression of great elegance. Full-fledged Sukhothai images of the full-round walking Buddha—an original Sukhothai invention—emphasize a kind of swaying, sinuous, boneless grace in the execution of the legs and arms. One of the most impressive colossal images of the type is the brick and stucco icon at the Wat Mahathat, Sawankhalok, another Sukhothai technical forte, dating probably to the 14th century. This type of image remained the most popular in Siam; an enormous number of imitations, of all dates, are preserved, many in Western collections.

Perhaps the Buddha types most successful aesthetically were those called after U Thong. They were produced originally in the southern capital of Ayutthaya, which took over Sukhothai in 1349, and represent a fusion of the Sukhothai types with vestiges of Khmer and Theravāda Dvaravati traditions, whose Buddha types had been marked by a strong Mon sense of squared-off design and cubic volume. The latter may have been influential because they seemed to incorporate an older and more authentic tradition, since they were based upon patterns developed in eastern India, the true homeland of Buddhism. In the U Thong style the sinuous linear curves, loops, and dry ridges of the pure Sukhothai patterns are



"Buddha Calling the Earth to Witness," Thai, bronze, Sukhothai high classical style, 14th century. In the collection of Prince Chalermbol Yugala, Bangkok. Height 94 cm.
By courtesy of the Breezewood Foundation, Monkton, Maryland

suppressed, and genuine modelling, with clearly defined planes and volumes, appears. In the northern kingdom a crude version of the Sukhothai type gained currency in the late 14th century. When, in the middle of the 15th century, King Tiloka of the northern kingdom re-established contact with Ceylon, images seem to have been imported directly from that country. They must have shown clearly how far the Sukhothai types had departed from the type used in the Buddhist homeland, because the third Siamese icon pattern, known as the lion type, attempted to recapture the stern simplicity of the genuine Sinhalese images. Most of the best examples were made between 1470 and 1565. Limbs and bodies are given a massive cylindrical strength, and the Sukhothai elegance is eliminated. It seems, however, that the native Thai genius is for the sinuous and unplastic curve, which may have expressed for them the same spiritual unworldliness as it did in Burmese ornament. Thus, in later examples reminiscent of the lion type, the curvilinear patterns of the Sukhothai style reassert themselves with more or less emphasis; and by the end of the 16th century the lion type had lost its distinguishing features and merged into the run of Sukhothai patterns.

Architecture and painting. There are as yet few results of authenticated research available concerning the history of architecture during the early period of Thai supremacy. Many monasteries contain stupas, or *cheddis*, that either originated or were renewed in this period; but most of the monasteries themselves have been repeatedly overworked. Building complexes seem to have developed by accretion, rather than by the studied working out of space articulations. The oldest building in Ayutthaya, dating from the early 13th century, is the Wat Bhuddai Svarya, a towered shrine, approached by a columned hall. From the late 14th century onward, Sukhothai influence seems to have predominated everywhere. The architectural types included a bell-shaped reliquary stupa with a circular flanged base and onion finials, reminiscent of combined Ceylonese and Burmese patterns; a stupa raised upon a cylindrical shrine as its drum; and a shrine with a plinth faced with images (usually later additions) above which rise one or more pyramidal towers reminiscent of the tower of the Mahābodhi temple at Bodh Gayā in Bihār, India. An example of the third architectural type is King Tiloka's late-15th-century Wat Chet Yot at Chiengmai, which has one large and four smaller pyramids mounted on a main block. The Thai kings also adopted something of the personal funeral cult of Khmer Angkor

14th-century architectural types



Wat Chet Yot, Chiengmai, Thailand, late 15th century.
Louis Frederic—Rapho Guillumette

(see below *Cambodia and Vietnam*), for a custom grew of building bell-shaped brick stupas—which had earlier been used only for the relics of Buddhist saints—as the kings' tombs, each approached by a colonnaded hall and surrounded by smaller stupas or shrines. In many of the brick and plaster or wooden monastic buildings of more recent centuries, such as the Wat Po in Bangkok, one can trace the distant influence of the Khmer styles of Angkor. Tall, gabled roofs, with steps and overlaps, the gables adorned with flame finials, are typical, exemplified by the Water Pavilion at Bang Pa-in.

Thai painting of the early period (13th–16th centuries) demands a great deal more research and study than it has yet received. Although it is, of course, devoted to the canonical iconography of the Theravāda, its fluent and relatively unschematic outline shows that it retained much of the original inspiration visible in the earlier work at Burmese Pagan (see above *Burma*). The oldest examples of Thai painting are the much-ruined frescoes in the Silpa cave, Yala, and some engraved panels from Wat Si Chum, Sukhothai, dated to 1287. Later paintings (dating to the 1420s) in the inner chambers of the Wat Rat Burana and Wat Mahathat at Ayutthaya show strong Chinese and perhaps Khmer influence in their high perspectives and landscape backgrounds with animals, combined with the native Thai clear outlines and bright, flat colours. By the 17th century at, for example, the Wat Yai Suwannaram at Phet Buri, large mural compositions —such as an elaborate scene of demons worshipping the Buddha—were being undertaken. In this later painting, theatrical stereotypes from the Thai dance-drama exerted a strong influence in the rendering of figures.

Luc Bouchage—Rapho Guillumette



Water Pavilion, Bang Pa-in, Thailand, 1294.

18th century to the present. In the 18th century the Burmese invaded and conquered Siam. The Burmese king —in expiation, it is said, of his war guilt—ordered the construction of many Buddhist buildings in the current Burmese style (see above *Burma*). These made their impact on Thai art, and the gaudy gilding and inlay characteristic of late Burmese ornament were widely adopted. When the capital was moved to the present Bangkok, in 1782, no substantial artistic development took place, though large pagodas were built and filled with rows of images, many in gilt wood. A highly ornate interpretation of older, airily flamboyant Burmese decorative styles, featuring curved "oxhorn" projections, blunted the edge of architectural and sculptural quality. Without exception, the new large-scale icons were dull and inferior works of art; and the monstrous guardian figures of spiritual beings and lions decorating the major shrines are fantastic rather than aesthetically valuable. In the painting of wooden panels, some of them votive, and of historical manuscripts, the Thai retained a good deal of their older vigour. The figures illustrating legend and history are based upon the unworldly stereotypes of the court dance.

Burmese influence on Thai art

In addition to the incorporation of European motives, many buildings and their ornamentation in Bangkok have a strongly Chinese flavour. This is attributable partly to the influence of the large expatriate Chinese population living there and partly to the influence of earlier expatriate Chinese craftsmen. The early-20th-century Pathamacetiya at Nagara Pathama (Nakhon Pathom), which is entirely orange, is a fine example of the many *cheddis*. Some tiles were certainly imported from China, but others were descendants of the fine pottery (of basically Chinese inspiration) that was produced at the kilns of Sawankhalok during the 14th and 15th centuries by expatriate Chinese craftsmen. This pottery imitated in its own materials Chinese Yüan dynasty (1279–1368) Tzu-chou and celadon wares (stonewares and porcelain with a glaze developed by the Chinese) with underglaze ornament and blue or brown painted decoration. Similar wares were made in the 15th century at kilns at Sukhothai and at Chiengmai. Some of these pieces are, in their own idiom, as fine as native Chinese work. Later, during the 18th and 19th centuries, somewhat garish, flamboyant Ayutthaya figure designs in polychrome were applied to rice bowls and other vessels.

Sawankhalok, Sukhothai, and Chiengmai pottery

Laos. The kingdom of Lan Xang (Laos) was founded in the mid-14th century and ruled by Buddhist Thai. At the northern capital, Luang Prabang, the influence of the northern Thai city of Chiengmai predominated; in the southern capital, Vientiane, a mixture of Ayutthaya and Khmer motives prevailed. In Laos there is no stone and little brick architecture. The most impressive single monument, the brick and stucco That Luang in Vientiane, founded in 1586 but much restored, is a stupa, shaped as a tall four-faced dome on a square plinth enclosed in a court; the dome is crowned with an ornate spire and encircled by a row of similarly shaped spires. The architecture of monastic halls also follows the Thai pattern; very steep multiple-gabled roofs, gently curved and over-



Fresco of the Preaching Buddha at the Wet-kyi-in, Ky-byauk-kyi, Pagan, *c.* 1113.



Terra cotta relief at the Petleik-Matangjakata temple, Pagan, late 11th century.

Burmese art



Vessel and cover in the shape of a sacred bird, gold decorated with filigree work and inlaid with rubies and imitation emeralds, 19th century. In the Victoria and Albert Museum, London. Height 41.5 cm.

Shwe Dagon pagoda, Rangoon, *c.* 15th century.

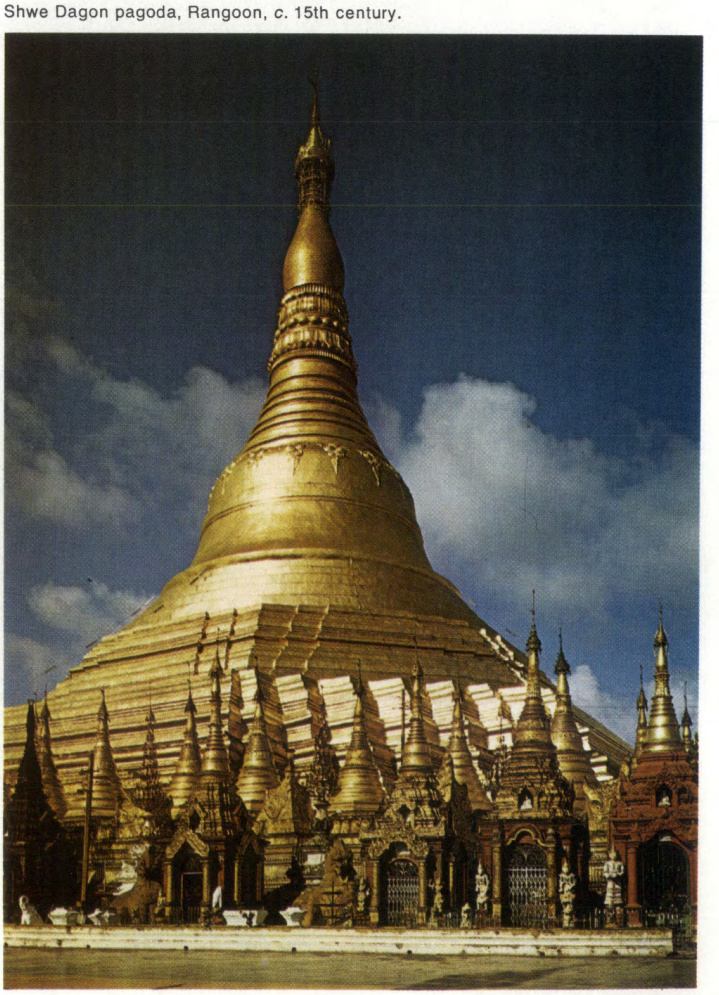


Plate 1: By courtesy of (bottom left) the Victoria and Albert Museum, London; photographs, (top left, top right) J.A. Lavaud, Paris, (bottom left) A.C. Cooper Ltd., (bottom right) R. Manley—Shostal



That Luang *stupa*, Vientiane, Laos, 1566, restored 18th and 19th centuries.



Bodhisattva from Nanchao, ancient Tai kingdom (now in the Chinese province of Yunnan), bronze, 13th century. In the British Museum. Height 44 cm.

Thai and Laotian art

Thai painted lacquer panel of a court scene, Bangkok style, middle of the 19th century. In the collection of Prince Piya Rangsit, Bangkok. Height 50 cm.





"The Great Departure of Bodhisattva," detail from "Episodes from the life of Buddha," Thai painting on silk panel, 17th-18th century. In the Musée Guimet, Paris. Detail 93 × 93 cm.

Plate 2: By courtesy of (top right) the trustees of the British Museum; photographs, (top left, bottom left) Holle Bildarchiv, Baden-Baden, (bottom right) Ciccione—Rapho Guillumette
Plate 3: (Top left, centre) Holle Bildarchiv, Baden-Baden, (top right) © Pascal Grellety-Bosviel, Paris, (bottom) R. Manley—Shostal



Library, Banteay Srei, Angkor, Cambodia, 10th century.



"Siva and Umā," sandstone sculpture from Banteay Srei, Angkor, Cambodia, late 10th century. In the Musée National, Phnom Penh, Cambodia. Height 60 cm.



Cham sandstone panel of a pedestal altar showing an ascetic playing a flute, from Mi-son E 1, Vietnam, second half of the 7th century. In the Da Nang museum, Vietnam. Height 60 cm.

Cambodian and Vietnamese art

Gate at Angkor Thom, Angkor, Cambodia, *c.* 1200.





Pura Besakih, Gunung (mount) Agung, Bali, 14th century. In the centre is the eleven-storeyed *meru* dedicated to Siva.



Gold kris, embossed scabbard and grip, from southern Celebes. In the Royal Tropical Institute Museum, Amsterdam. Overall length 40.5 cm.

Indonesian art



Indonesian textiles. (Top) Javanese batik textile accented with gilding. In the Royal Tropical Institute Museum, Amsterdam. (Bottom) *Ikat* cloth from Sumba Timur, Lesser Sunda Islands. In the J. and R. Langewis Collection, Castricum, The Netherlands.



Main entrance to the Pura-desa Pasaban, Bali.

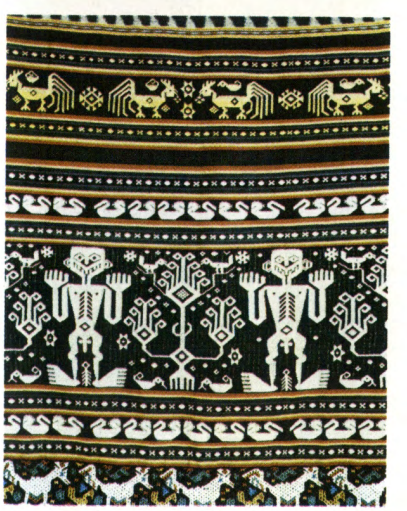


Plate 4: By courtesy of (centre left, centre right) the Royal Tropical Institute, Amsterdam; photographs, (top, bottom left, bottom right) Holle Bildarchiv, Baden-Baden

hung with long eaves, are carried on brick or wooden pillars and adorned with flame finials. Buddha figures, preserved in some of the monasteries, are based on northern Thai versions of Sukhothai types; some may be as early as the 17th century. The schematic paintings on monastery walls are in versions of the later Thai styles. In the northwest a strong influence from late Burmese art can be found in Buddhist images made to serve a religion that was far closer to the original Thai animism than to true Buddhism.

CAMBODIA AND VIETNAM

Paleolithic tools similar to types found in India have been found in Cambodia and Vietnam; and it is possible to trace the movement of population or culture groups, some of whom probably migrated onward by sea from Southeast Asia into the islands. The important group of speakers of Mon–Khmer languages may conceivably have been the people who produced the megalithic monuments in Cambodia and Laos, which include colossal stone burial urns, dolmens, and menhirs, perhaps associated with the many circular earth platforms as yet unexcavated (see above *General development of Southeast Asian art*). Probably contemporaneous, at least in part, with the Neolithic Mon–Khmer culture is the culture known by the name of its richest, most northerly site, Dong Son, on the coast of the Gulf of Tonkin in North Vietnam. It seems probable that the chief influences on this culture came from southern China. Many sites, ranging in date from about the 4th to the 1st century BC, stretch southward from the coast of Vietnam, as far as northern New Guinea. The islands of Indonesia and parts of Malaya may have been the principal location of the Dong Son culture.

Bronze drums

The most impressive bronze objects produced by this culture are large drums, which seem sometimes to have been buried with the dead. Splendid examples have been found in Java and Bali (see below *Indonesia*). These and many other bronze objects, such as superb funeral urns with relief ornament based on squared hooks, lamp holders, dagger hilts in the form of human figures, and other weapons, are of extremely high quality. Their ornament was produced by the Chinese casting technique of incising the patterns into the negative mold that was to receive the molten bronze; much of it suggests a parallel version of contemporary Chinese ornament of the Ch'in period (221–206 BC). From the figures and objects represented in this bronze work, it seems that the Dong Son culture had much in common with that of some of the peoples of the Melanesian islands today. The culture knew large seagoing canoes, houses similar in structure to those still common among peoples of Melanesia, and ceremonies that the Melanesians might recognize. It is probable that one group of their descendants, which retained its identity, is known to the history of this region as the Cham (see below *Cambodia and Vietnam*).

By courtesy of the Fine Arts Conservation, Ministry of Culture, Hanoi



Lamp holder from Lach Truong, North Vietnam, bronze, Dong Son culture. In the National Museum, Hanoi. Height 33 cm.

Although many peoples isolated in the densely forested uplands also retained a tribal identity, by far the most important art was produced in the two Indianizing empires: Khmer, in Cambodia, with its linear predecessors the kingdoms of Funan and of Chenla (names they were given by Chinese historians), and the Cham, in Vietnam.

Cambodian kingdoms of Funan and Chenla: 1st to 9th century. Funan, which was in existence by the 1st century AD, was the earliest of the kingdoms that arose along the lower reaches of the Mekong River in response to Indian ideas. Its influence probably extended over long stretches of the coast of the Gulf of Siam, even as far as southern Burma, and corresponded with the range of the Mon peoples. Lying on the natural focus of land and sea routes linking eastern India and southern China to the islands of the South Seas, its geographical situation was ideal for a kingdom whose wealth was based on trade. At Funan sites even Roman, Ptolemaic Egyptian, and Sassanian Persian objects have been found, giving an idea of the extent of its trading interests.

The founder was probably a Brahmin trader from western India; for a local legend describes how the first king, a Brahmin, married the daughter of a local serpent deity, so establishing the ruling family. Serpents (*nāgas*) in Indian mythology are the spiritual patrons of water; and the basis this kingdom laid for later kingdoms in the same area was an elaborate system of waterworks, canals, and irrigation channels controlling and distributing the waters of the Mekong River. Contemporary Chinese accounts refer to cities with splendid wooden buildings, carved, painted, and gilded. But nothing remains, save a few foundation piles. Probably during the 6th century AD the kingdom called Chenla was established in the upper-middle reaches of the Mekong River, in what is now Laos. The kings who ruled in Chenla were descended from the kings of Funan and took over much of the Funan domain. It seems that disastrous floods had finally ruined Funan, which had previously suffered from Indonesian aggression, and that the shift of power to Chenla represented a recognition of temporarily insuperable geographical difficulties.

Stone sculptures

Culturally, Funan and Chenla are continuous. Their artists produced some of the world's greatest stone sculptures, most of which are large, freestanding icons, carved in sandstone. Intended to be installed in brick-built shrines, none of which survive, they usually represent the two major deities of Hinduism, Śiva and Viṣṇu. Sometimes both deities are combined into a single figure called Harihara; the right half of the body is characterized as Śiva, the left as Viṣṇu. A few examples of other figures are known, including some magnificent images of goddesses. The style of these sculptures is marked by an extremely smooth, continuously undulating surface, given strength by a system of clear, broad frontal planes and side recessions related to the foursquare block. Such images were meant to demonstrate the power and charm of a heavenly prototype to whom an earthly king appealed for his authority. The earliest images belong to the 6th century, and the series continues into the 9th.

In later Khmer times each king and sometimes each member of a royal house had statues of himself or herself in the guise of a patron deity set up in the family temple precinct. That the same custom prevailed in 6th-century India, particularly in the southeast, suggests that some of the early Funan and Chenla sculptures may have served the same function. A number of figures are Indian in style —some more markedly than others, which is probably more than a matter of date; for it is quite likely that Indian craftsmen occasionally travelled into this region to work. The style of the greatest of these early sculptures, however, is not Indian at all.

Similarly non-Indian are the magnificent sandstone lin-

Sandstone lintels

tels made for the doorways of the vanished brick shrines. Although distantly related to Indian prototypes of the 1st and 2nd centuries AD, they appear as full-fledged Indochinese inventions and may well have been developed in combination with a native conception of the lintel as a special attribute of the spirit shrine (see above *Thailand and Laos*). They are carved in relief with designs based on a pair of monsters, one at each end, which are linked by an ornate arched or lobed beam. The beam is adorned with figures inside foliate plaques, a long sequence of elaborately carved swags of jewels hanging beneath them.

Among the Funan–Chenla sculptures are a few Buddhist icons executed in sandstone, markedly less sensuous than the Hindu figures and close to the styles of Dvaravati (see above *Thailand and Laos*), though a number of small Buddhist bronzes representing *bodhisattvas* approach the delicacy of the Hindu work.

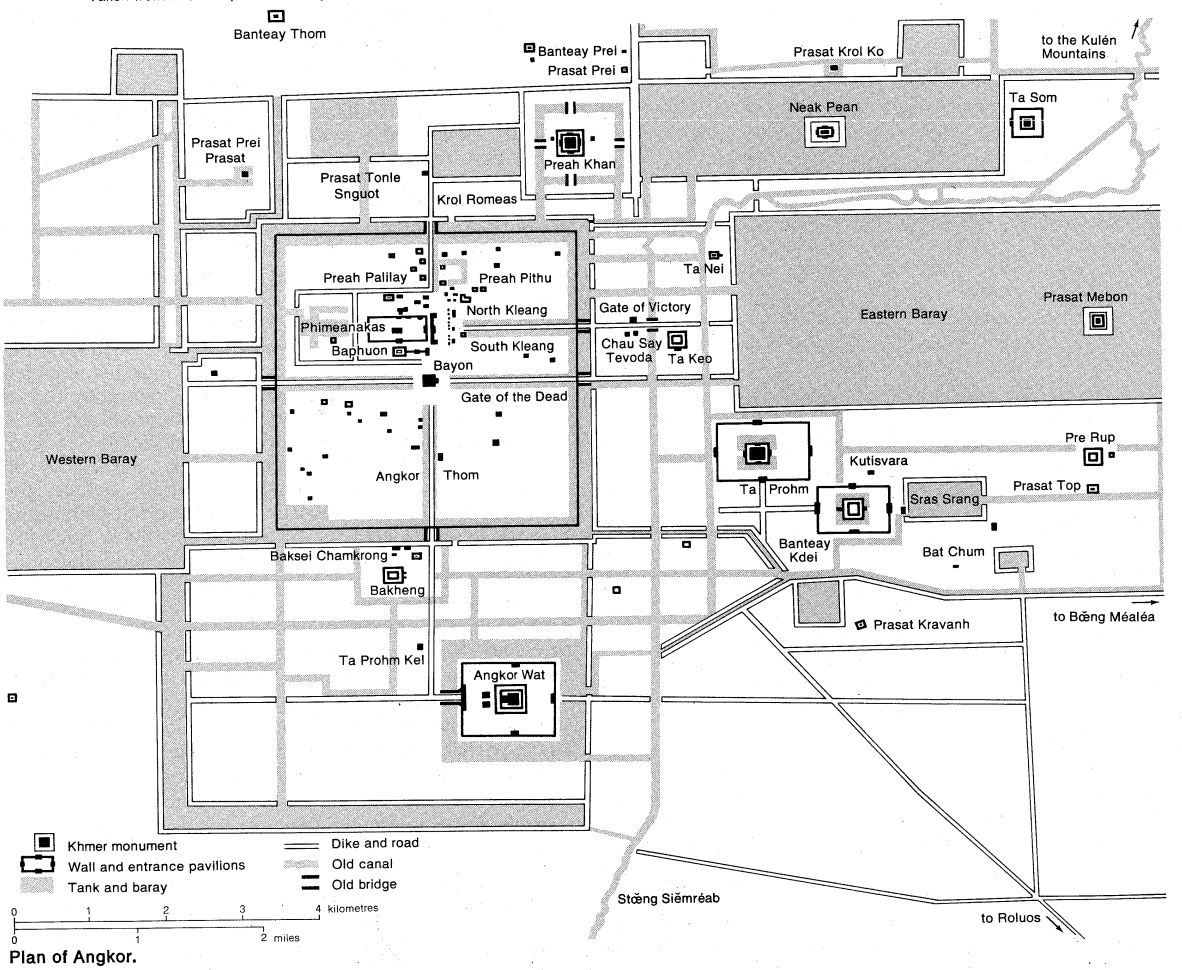
Kingdom of Khmer: 9th to 13th century. Late in the 8th century the kingdom of Chenla declined politically, perhaps because of dynastic disputes with the rising power of Indonesian kings, who were themselves also descended from the original royal dynasty of Funan. It seems that the Indonesians gave some assistance in establishing a new kingdom in the northern part of what had been the territory of Funan. In 802 a Khmer king, who took the title of Jayavarman II, established his capital near Phnom Kulen, about 20 miles (30 kilometres) from Angkor. It was a rather unsuitable place for an administrative capital, but it was a mountain, and the peoples of Southeast Asia have always believed that gods and spirits dwell on mountaintops. The image of the sacred mountain thereafter remained the inspiration for all the later architecture of the Khmer around Angkor. Jayavarman, who built other temples in the vicinity, seems to have revived the Chenla style. A distinctively Khmer art, however, began to emerge under Indravarman I (877–889), who expanded the boundaries of the Khmer kingdom and finally settled its administration. Most important of all, he developed the initial plan of the colossal city of Angkor, whose mysterious ruins, lost in dense jungle until very recently, have tantalized Western travellers for three centuries.

Technological significance of Angkor

Angkor is not only a city; more important, it was an immense technological achievement, from which the agricultural prosperity of the whole Cambodian plain derived. This plain was well watered naturally, but its rivers were subject to strong seasonal fluctuations. Controlled, they were capable of producing an enormous increase in fertility. Angkor was thus essentially an elaborate system of artificial lakes, canals, and radiating irrigation channels that watered a huge acreage of rice paddy; and it was the basis for the strength and prosperity of the Khmer Empire. Since Angkor itself was the technical source of the life-giving agricultural water controlled by the king, it was regarded by the Khmer with religious reverence. Its temples and palaces were an expression of that reverence and at the same time an essential part of its supernatural mechanism. Royal intercession by numerous ceremonies, some of which re-enacted the primal marriage of Hindu divinity and native earth spirit on the pattern of ancient folk cult, ensured the continuing gift of the waters of heaven. The king, an earthly image of his god, was the intermediary who ensured that his kingdom would continue to receive divine benevolence in the form of water in controlled quantities. Courtiers played roles at once religious and administrative for the king, who believed that after his death he would be united with his patron deity. Dedicatory statues were often set up in his chief temple to commemorate his divinization.

In order to conform with mountain mythology, the Khmer kings built themselves a series of artificial mountains on the Cambodian plain at Angkor, each crowned by shrines containing images of gods and of themselves, their family, and their ancestors. The huge platforms of

Taken from *The Art of Indochina* by Bernard Philippe Groslier, © Holle & Co. Verlag, Baden-Baden, Germany. Used by permission of Crown Publishers, Inc.



Plan of Angkor.



Angkor Wat, Angkor, Cambodia, mid-12th century.
Asie Photo—De Wys

earth on which these buildings were founded probably consist of the soil excavated in forming the lakes, moats, and channels that not only divided up the city but also provided an easy means of transport. The temple mountains, like the city itself, are oriented east to west, the main gates facing east. Each king strove to outdo his predecessor in the height, size, and splendour of his temple mountain. The earlier ones, therefore, are relatively small, though beautiful, while the later ones, such as Angkor Wat and the Bayon, are of stupendous size.

The basic pattern of the Khmer temple mountain

In the basic pattern of the Khmer temple mountain the principal overall enclosure, which is square or rectangular, is at ground level. Within it the artificial mountain rises through a series of terraces and at least one further enclosure wall toward a flat summit. On the summit stands either a single shrine or a group of shrines, often a quincunx—five shrines, one at each corner and one in the middle of a square. Arranged along the terraces or within the enclosures there may be further shrines, whose arched doorway pediments refer to the rainbow bridge between heaven and earth. There may be other long buildings, perhaps used as libraries or administrative offices. A principal staircase runs directly up from the east gate to the summit, and sometimes subsidiary staircases run up from other gates at the cardinal directions.

The architecture of the shrines themselves is relatively simple; it is based upon patterns invented in India, though the ornament of the shrines is often highly developed and characteristically Cambodian. Fundamentally, each shrine consists of a cell whose internal space is cubic and whose external walls are marked by moldings at top and bottom. The shrine is roofed by a pyramidal tower composed of a series of similar but diminishing tiers, each of them a compressed version of the exterior pattern of the main shrine volume. Depending on which Indian pattern is followed, the cell has one main door with an elaborately carved portal or, if the plan is cruciform, four entrances. The earlier shrines were built of brick, most commonly with stucco ornament and figures on the outside. The later shrines were built of stone, with all their ornament and figurative sculpture carved in relief. The moldings on the roofs of the shrines and the decoration of the roofs of many of the subsidiary buildings are extremely elaborate. There are long panels of dense foliate ornament, and the niches in which the sculptured relief figures of celestials are set are framed in flamboyant ogival (contoured like a pointed arch) moldings crowned by no less flamboyant foliate ornament; the smaller architectural features, such as niche pilasters, are elaborately carved and molded. The figures themselves wear gorgeous jewelry and chignons. The massive stone icons that survive in some of the shrines and on the terraces do not have the subtlety or strength of the Funan–Chenla sculptures. Instead, they have an inflated massiveness, intended, no doubt, to make them awe-inspiring. Among the lesser relief figures of celestials, which decorate the walls of the shrines, one finds a more sensuous touch; for many of these celestials represent *apsarases*, the celestial girls of Indian mythology.

Relief panels

On some of the temple mountains there are also relief panels illustrating various aspects of the royal mythology. Episodic relief sculpture first appears on Banteay Srei (10th century). The relief revolves around a series of Indian legends dealing with the cosmic mountain Meru as the source of all creation and with the divine origin of water. The chief artistic achievement of its architecture is the way in which it conceives and coordinates the spaces between the walls of the enclosures, the faces of the terraces, and the volumes of the shrine buildings. A most sophisticated architecture of full and empty space, it seems to have been influenced by that of the Hindu Pallava dynasty in southeastern India.

The earliest more or less complete example of a shrine complex devoted to deifying the ancestors of a king is the Preah Ko at Roluos, near Angkor, completed in 879. The earliest surviving temple mountain at Angkor itself is the Bakong, probably finished in 881. In the central shrine at the summit was a *liṅga*, the phallic emblem sacred to Śiva. Around the base of the terraced pyramid stood eight large shrines inside the main enclosure, with a series of moats, causeways, and auxiliary sculptures guarding the approaches to the exterior. The Bakheng, begun in 893, had an enormous series of 108 tower shrines arranged on the terraces around the central pyramid, which was crowned by a quincunx of principal shrines. The whole was intended to illustrate a mystical conception of the cosmos, very much on the lines of the great temple mountain at Borobuḍur in Java (see below *Indonesia*). Pre Rup, dedicated in 961, was probably the first of the temple mountains intended as a permanent shrine for the divine spirit of a king after his death. It, too, has a quincunx of principal shrines, but it is distinguished by the large number of auxiliary pavilions arranged along both sides of the inner enclosure wall.

From about the same period is perhaps the most beautiful—and most beautifully preserved—of the early Khmer temples, Banteai Srei. This was actually a private foundation, built some 12 miles from Angkor by a Brahmin of royal descent. Its auxiliary buildings, all of sandstone, are adorned with a profusion of elaborate ornament and relief figure sculpture. The roof gables, in particular, are treated with antefixes of fantastic invention. Its principal icon, a huge sandstone sculpture of the god Śiva, seated with his wife Umā on his left knee, is perhaps

the most impressive full-round sculpture from the whole Khmer epoch. It differs from 10th-century Khmer official sculpture, which began to take on a conventional and relatively insensitive massiveness.

The Baphuon temple mountain (1050–66) is unfortunately almost completely destroyed. It was a vast monument 480 yards (440 metres) long and 140 yards (130 metres) wide, approached by a 200-yard (180-metre) causeway raised on pillars. Its ground plan shows that it was no mere assemblage of buildings but a fully articulated structure. In this it must rank as the immediate prototype for the great Angkor Wat. Built by Suryavarman II in the early 12th century, Angkor Wat is the crowning work of Khmer architecture, the culmination of all the features of earlier styles.

Angkor Wat

The enormous structure of the Wat is some 1,700 yards (1,550 metres) long by 1,500 yards (1,400 metres) wide. Surrounded by a vast external cloister, it is approached from the west by a magnificent road, which is built on a causeway and lined with colossal balustrades carved in the likeness of the cosmic serpent, associated with the sources of life-giving water. The Wat rises in three concentric enclosures. The western gate complex itself is nearly as large as the complex of central shrines, and both are subdivided into smaller, beautifully decorated courts. Only five of the original nine towers still stand at the summit; although they follow the basic pattern of the Khmer roof tower composed of diminishing imitative stories, the contour of the towers is not rectilinear but curved, so as to suggest that the stories grow one out of another like a sprouting shoot. All the courtyards, with their molded plinths, staircases, porticoes, and eaves moldings, are perfectly articulated enclosed spaces. The symbolic meaning of the Wat is clear. Its central shrine indicates the hub of the universe, while its surroundings —the gate complex, the cloister, the city of Angkor itself, and, finally, the whole visible world—represent the successive outer envelopes of cosmic reality. That it is oriented toward the west—and not to the east, as was customary—indicates that its builder, Suryavarman II, intended it as his own mortuary shrine; for, according to Indochinese mythology, the west is the direction in which the dead depart.

The sculptors who worked at the Wat demonstrate little ability in carving in the full round. Such full-round figures as there are—the guardians on the terraces, for example—lack life. The relief sculpture, however, is

Holle Bildarchiv, Baden-Baden



Bas-relief of a battle scene, Angkor Wat, Angkor, Cambodia, early 12th century.

magnificent and full of vitality. The open colonnaded gallery on the first story contains over a mile of relief carving six feet (two metres) high. Much of it was originally painted and gilded, which strongly suggests that there must have been a Khmer style of painting of which nothing is known. The subject matter of the carvings is taken principally from the Hindu epics, but there are also many scenes representing Suryavarman's earthly glory. Working in relief only about an inch deep, the sculptors were able to depict an extraordinary complex of scenes of figures in vigorous action, full of complex overlaps to suggest deep space. The solid bodies are created mainly out of groups of convex curves; and everywhere there is the typical regional feeling for decorative spirals. Perhaps the most interesting group of figures are the *apsarases*, carved in relief, either singly or in groups, on the plain walls of the courtyards. These celestial beauties, whom Indian tradition describes as rewarding with their charms the kings, heroes, and saints who attain heaven, are carved with sinuous sensuality; but the most important part of their charm is their elaborate clothing, jewelry, and hairdressing or ornate, towering, jewelled crowns. Apparently, deep, downward-drooping curves standing far out from the body represented the height of Khmer chic. Skirts, stoles, and the long sidelocks of the hair all follow these curves, laid out flat on the ground of the relief. Symbolizing the erotic joys that are essential attributes of heaven, the *apsarases* were natural possessions of the king.

The *apsarases*

In many senses the Wat was the end of the road for Khmer art. The effort demanded of the people in constructing this colossal stone monument, along with its four miles (six kilometres) of stone-lined moat 200 yards (180 metres) wide, was too great. The irrigation system itself may well have been neglected in favour not only of shifting the building stone—as much as there is in the Pyramid of Khafre in Egypt—but also of dressing, carving, and ornamenting it. After Suryavarman's death, the Cham, from the neighbouring kingdom of Champa (see below *Cambodia and Vietnam*), seized and sacked Angkor for the first time in its history (1177), thus shattering the confidence of the Khmer people in their Hindu deities. When Suryavarman's son, Jayavarman VII, came to the throne he inherited a ravaged kingdom. In 1181 he succeeded in driving out the Cham; he invaded their country and seized their capital, making Champa a province of the Khmer. Then, over 60 years old, he embarked on a series of campaigns that extended the borders of the Khmer Empire further than ever before—into Malaya, Burma, and Annam.

The ruler of this empire naturally believed himself to be the greatest of the Khmers, and he set about demonstrating the truth of his belief by building his own city, Angkor Thom (*c.* 1200), and, at the centre of it, the biggest temple complex of them all—the Bayon (*c.* 1200). Breaking with all previous Khmer traditions, he took as his patron deity not one of the Hindu gods but one of the Buddhist *bodhisattvas*. Although Buddhism had flourished for several centuries in the whole of Indochina, it had not been adopted by the Khmer as an imperial cult. Now that the Hindu gods had been discredited by defeat, Jayavarman placed himself under the patronage of Mahāyāna Buddhism. The mythology according to which the Bayon was designed was thus another version of the old mythology of the celestial mountain and the divine origin of water. Only the central figure of his mythology, Lokeśvara, Lord of the World, was specifically Buddhist. The colossal masks that look out over the four directions of the world from the towers of the Bayon and from the gates of Angkor Thom are there to demonstrate the compassionate, all-seeing power of Lokeśvara and the king.

The Bayon

When Jayavarman VII set out to create Angkor Thom, he had to raze the fine older work of his predecessors, for the site at Angkor had become choked with nearly four centuries of grandiose temple building. Within Angkor Thom's ten miles (16 kilometres) of moats he constructed huge complexes of building and made his city the focus of a final system of canals and irrigation, with additional lakes.

Unfortunately, the innumerable new shrines that surround Angkor Thom, the towers that crowd the Bayon, and the vast stone terraces faced with relief that surround

the royal palace are in general much inferior in execution to the work of the earlier kings. Thus, today Angkor is dominated by the overwhelming presence of Jayavarman's immense but relatively unrefined architecture. The King's ambition was satisfied by size and quantity rather than artistic quality. Because sculptors were obliged to produce such vast quantities of work so fast, their standards deteriorated, and the huge vistas of narrative relief show signs of haste and slipshod workmanship. The real achievement lay with Jayavarman's scholastic architects, who conceived and laid out a complex of mythical imagery in massive architectural symbols. Their stupendous overall plan illustrates the creation of the world, a cosmos spreading outward from the central mountain tower. The two roads leading from the tower are lined with mile-long rows of gigantic deities who are pulling on the body of the serpent *nāga*. According to Hindu legend, the gods use the magical mountain Meru, symbolized by the mountain tower, as a churning stick and the body of the cosmic serpent as a churning rope to churn the world out of the milk of nothingness. Lake-sized fountains represent the healing waters of the Buddhist paradise, and allegories of salvation are realized in carved architecture. Perhaps the most impressive works of art associated with this last period of Angkor are some stone icons, such as the famous "Leper King," in the Angkor Thom complex. Many excellent smaller bronze figures of deities have also been found among the ruins.

Decline of Angkor

13th century to the present. After the death of Jayavarman VII, *c.* 1215, possibly as late as 1219, Angkor declined. The Thai population of Siam gradually pushed the Khmer down toward the Mekong Delta. Theravāda Buddhism became the religion of the people, and the grandiose vision of a cultural unity based on sacred kingship disappeared. In the 15th century, Angkor was retaken from the Thai, and a few buildings were restored by the ancestors of the modern (now abdicated) Cambodian kings. Some of the buildings were used as monasteries, but the city, with its essential irrigation system, had fallen into ruin.

Vietnam kingdom of Champa: c. 2nd to 15th century. The kingdom of Champa existed alongside the Khmer kingdom, sometimes passing under its rule, sometimes maintaining a precarious independence. From the north it was continually subject to the pressure of the advancing Vietnamese, a people racially related to the Burmese and Thai, who were themselves under pressure from the Chinese. The Hinduizing dynasties who ruled Champa from the 6th century were obliged to pay heavy tribute to the Chinese Empire. After 980 they were forced by the Vietnamese to abandon their northern sacred capital, My Son; thereafter, except for a brief return to My Son in the 11th century, their southern capital at Vijaya (Binh Dinh) became their centre. Under such disruptive circumstances, it is perhaps surprising that the Cham succeeded in creating and maintaining a dynastic art of their own. It was, however, always on a relatively modest scale, devoted to a conception of divine kingship similar to but far less ambitious than the Khmer.

The evolution of Cham art falls naturally into two epochs, the first when the capital was in the north, the second when it was removed to the south.

Art of the northern capital: 4th to 11th century. The form of the earliest temple at My Son, built by King Bhadravarman in the late 4th century, is not known. The earliest surviving fragments of art come from the second half of the 7th century, when the king was a descendant of the royal house at Chenla. The remains of the many dynastic temples built in My Son up until 980 follow a common pattern with only minor variations. It is a relatively simple one, with no attempt at the elaborate architecture of space evolved by the Khmer. Each tower shrine is based upon the central rectangular volume of the cell. The faces are marked by central porticoes that are blind on all but the western face, where the entrance door is situated. The blind porticoes seem to have contained figures of deities—perhaps armed guardians standing in a threatening posture. The porticoes are set in a tall, narrow frame of pilasters (columns projecting a third of their width or less from the wall), crowned with horizontally molded capitals that step out upward. They support a tall, double-ogival blind arch, crowned by another stepped in behind it. The arches are based on an Indian pattern and are carved with a design of slowly undulating foliage springing from the mouth of a monster whose head forms the apex of the arch. The faces of the walls are formed of pilasters framing tall recesses. The pilasters are carved with foliate relief, and elaborate recessed and stepped-out horizontal moldings mark their bases. The height of the pilasters and recesses gives a strong vertical accent to the body of the shrine. The principal architrave is carried on stepped-out false capitals to the pilasters. The roof of the tower is composed of three diminishing, compressed stories, each marked by little pavilions on the faces above the main porticoes. Inside the tower is a high space created by a simple corbel vault with its stepped courses of masonry. The chief portico was extended to include a porch, and the whole structure stood upon a plinth whose faces bore molded dwarfed columns (small columns) and recesses.

Pedestal altars

These temples have one distinguishing internal feature: a pedestal altar within the cell, upon which statues were set, sometimes, it seems, in groups. The pedestals themselves are often beautifully adorned with reliefs, and some of the best Cham sculpture appears upon them. The subjects are usually based on Indian imagery of the celestial court. The fact that the pedestal altars carried their sculptures in the space of the cell, away from the wall, meant that the Cham sculptors could think in terms of three-dimensional plasticity as well as relief.

The glory of Cham art is the sculpture of the whole of the first period. Much of what survives consists of lesser figures that formed part of an architectural decor: heads of monsters, for example, which decorated the corners of architraves, and figures of lions, which supported bases and plinths. These figures reflect the heavy ornateness of the Cham decorative style at its most aggressive; and many of them effloresce into the solid, wormlike ornament that is the Cham version of Indo-Khmer foliage carving and carries strong reminiscences of Dong Son work. The remaining fragments of the large icons suggest a double origin for Cham art traditions. On many of the capitals and altar pedestals are series of figures carved in relief in a sensuous style, which is nevertheless strictly conceptualized. This sophisticated work is reminiscent both of late Chenla art (see above *Cambodia and Vietnam*) and of Indonesian decoration, especially during the 11th-century return. Other figures are more coarsely emphatic in style, with the crudely defined cubic volumes and the heavy faces of Melanesian sculpture. It is thus probable that artists trained in the sophisticated Cambodian tradition worked for the Cham kings at one time or another, while Champa's own native craftsmen emulated the work of the foreigners in their own fashion.

Apart from My Son there are one or two other sites in north and central Vietnam where Cham art was made in quantity. The most important of these is Dong Duong, in Quang Nam. It is a ruined Buddhist monastery complex of the late 9th century, conceived on the most beautifully elaborated plan of structured space in Champa. The architectural detail is distinguished from the My Son work by its greater emphasis upon the plasticity of architectural elements such as angle pilasters and porticoes. The circuit wall was about half a mile (one kilometre) long and once contained many shrines dedicated to Buddhist deities. It is possible that, when this complex of brick courts, halls, and gate pavilions was intact, it may have resembled very closely the contemporary Buddhist monasteries of northeastern India.

Art of the southern capital: 11th to 15th century. After 980, when the northern provinces were taken over by the Vietnamese and the Cham capital established at Binh Dinh in 1069, the kings maintained a gradually diminishing splendour. After the Khmer attack of 1145 they could claim little in the way of royal glory.

Although the Cham kings made a brief return to My Son from 1074 to 1080, most of their artistic effort was spent on shrines at Vijaya (Binh Dinh) and a few other

sites in the south. The early-12th-century Silver Towers at Binh Dinh are simplified versions of the older northern towers, with corner pavilions added to the roofing stories and arches of pointed horseshoe shape. Throughout the 13th and early 14th centuries the architecture of successive shrines gradually declined. The plasticity of the old pilasters and architraves was suppressed into simple moldings, and the beauty of the buildings became largely a matter of proportion. By the mid-14th century even the temples erected at Binh Dinh amounted to little more than piles of crudely cut stones articulated only by reminiscences of the classic Cham style.

Decline of shrine architecture

Sculpture shows a parallel decline. One or two reliefs at the Silver Towers do convey a sense of tranquillity and splendour, but an indigenous style of rigid cubical emphasis came progressively to dominate the iconic Hindu figures at southern sites. The curlicued design of earlier figures was gradually converted into a style of massive, scarcely carved blocks that convey, at their best, an impression of barbarous strength but without the refinement of first-class primitive art.

As was the case in Cambodia, this decline in art by the mid-14th century may be attributed to the people's loss of confidence in the concept—and, with it, the imagery—of divine kingship. Theravāda Buddhism, as a popular religion based upon numerous small, local monasteries, adopted probably from the Tai, was spreading all over the region. The northern Vietnamese, who had originally been organized in self-contained kingdoms without any concept of royal divinity, owing an intermittent administrative allegiance only to the distant Chinese emperor, found this ultimately suitable as a state religion after the final eclipse of Confucianism in the 17th century. They did incorporate echoes of older Hindu architecture, however, in details of the flamboyant ornament used on eaves and gables of their wooden monastery buildings.

Vietnam: 2nd–19th century. The great achievement of Vietnamese art, at least during the Le period (15th–18th centuries), seems to have been in architectural planning, incorporating Confucian, Taoist, or Buddhist temples into the landscape environment. The plans themselves include halls for a multitude of images in South Chinese vein and provision for a variety of rituals. There are no intact monuments of early Vietnamese architecture that are unrestored. Numerous fragments exist, however—either isolated stone bases, columns, stairways, and bridges or carved wooden members incorporated into later buildings—all of which are influenced to some degree by Chinese styles.

Tombs of the Chinese type

Tombs of generically Chinese type from the 2nd to 7th century contain bronze furnishings, in many of which, such as lampstands, the influence of the Dong Son style is clearly visible. There are no spirit images so typical of Six Dynasties (3rd–6th centuries) and T'ang (7th–10th centuries) Chinese tombs. The Chua Mot-cot, Hanoi, has vestiges of a stone shrine probably dated 1049. The only old paintings, on rock, at Thien Khe (Tuyen Quang, 9th century) represent the Buddha, *bodhisattva*s, and donors. The Van-mieu at Hanoi (built 1070 but frequently restored) contains ritual bronzes in "barbaric" Chinese style.

Perhaps the most interesting early sculptures to survive are the stone fragments from the Van-phuc temple (9th–11th centuries), which are based on Chinese Buddhist imagery but in a style strongly Indianized, perhaps by Cham influence. The most important piece of old work still virtually intact is the portable octagonal wooden stupa kept in the hall of the But-thap, at Bac Ninh, east of Hanoi. It has wooden panels carved in a flamboyant 14th-century Chinese style; part of it bears a representation of the Buddhist paradise of Amitābha. Incorporated in many Buddhist temples of the Le period (15th–18th centuries), as well as in stone terraces, bridges, and gateways, is extremely elaborate carved and coloured woodwork in a style based upon the coiling dragon-and-cloud decoration of Ming (1368–1644) and Ch'ing (1644–1911) China, but with a characteristically Vietnamese exaggeration of weight and curve.

At Tho Ha there was a potters' village, where the glazed ceramic figures used on many types of Chinese temple were manufactured. The remains of many tombs, palaces, bridges, and Confucian and Taoist temples decorated in similar vein are known everywhere.

19th and 20th centuries. The beautiful imperial palace of Hue (final plan before 1810) contained vestiges of older architecture and many works of Sinicized art before its devastation in 1968. It consisted of a series of simple, rectangular, one-story pavilions, laid out among trees inside a group of courts. These buildings and their decoration were southern Chinese in basic conception.

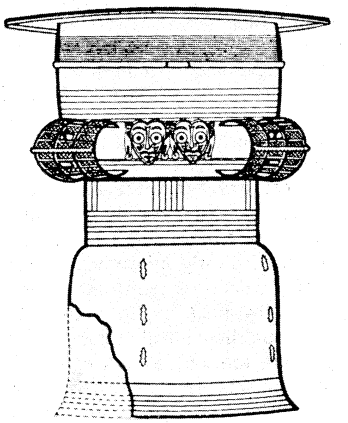
Elsewhere in Vietnam, both religious and secular buildings have been constructed in the 20th century in provincial versions of Chinese styles. There was little demand for the sculptor's art beyond the carving of stereotyped Buddha icons, monsters, and guardians. In modern times southern Vietnam has adopted a decorative style partly derived from the active traditions of Bangkok. Religious sects abound, with hybrid native European and Chinese elements used in their iconic and decorative art. Because of political turmoil, no clear and individual modern Vietnamese artistic tradition has been able to emerge.

INDONESIA

The islands that at the present day compose Indonesia probably once shared in the complex Neolithic heritage of artistic tradition, which also spread further, into the islands of Melanesia and Micronesia. Beautifully ground Neolithic axes of semiprecious stone are treasured still in some countries. In many parts of Indonesia there are quantities of megalithic monuments—menhirs, dolmens, terraced burial mounds, stone skull troughs, and other objects. Some of these are undoubtedly of Neolithic date, but megaliths continued to be made in much more recent times. One stone sarcophagus in eastern Java, for example, is dated post-9th century. On Nias island, megaliths are still revered, and they are still being erected on Sumba and Flores islands. Thus, in Indonesia especially, different layers of Southeast Asian culture have existed side by side. The most impressive and important collection of megaliths is in the Pasemah region, in south Sumatra, where there are also many large stones roughly carved into the shape of animals, such as the buffalo and elephant, and human figures—some with swords, helmets, and ornaments and some apparently carrying drums.

Megalithic monuments

These drums immediately suggest the drums characteristic of the mainland Southeast Asian Dong Son culture, which flourished *c.* 4th–1st centuries BC (see above *General development of Southeast Asian art*). This culture may well have helped to diffuse throughout the region styles related to Chinese Chou and pre-Han ornamental work. Certainly, the Dong Son influence is clear in many of the ceremonial axes, as well as many of the ornamented bronze drums, that have been found in the islands. The bronzes were cast by a cire perdue process resembling that used in parts of the Asian mainland. The largest and most famous drum is "the Moon of Bali," found on that



"The Moon of Bali," bronze drum found on an island near Pedjeng, Bali. In the Panataram Sasih temple, Pedjeng, Bali. Height 1.87 m, diameter 1.60 m.
From P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London



Indonesian sculpture.
(Left) "Buddha" from Celebes, bronze, Amarāvatī style, 3rd–5th century. In the Djakarta Museum. Height 75 cm. (Centre) "Viṣṇu" from Bali, stone. In the Royal Tropical Institute Museum, Amsterdam. Height 80 cm. (Right) Ancestor figure from the Tanimbar Islands, Indonesia. In the Royal Tropical Institute Museum, Amsterdam. Height 38 cm.
By courtesy of (centre, right) the Royal Tropical Institute, Amsterdam; photograph, (left) Louis Frederic—Rapho Guillumette

island near Pedjeng. It has molded flanges, and cast onto its faces is extremely elaborate relief ornament consisting of stylized masks with ears pierced and lengthened by large earrings. Such drums were probably originally used in ritual—by the rainmaker, perhaps—and they may have been buried with the distinguished dead. No one knows the exact age of these bronzes; "the Moon of Bali," for example, could be anywhere between 1,000 and 2,000 years old. Similar small drums were used quite recently as bride prices; and many of the islands today produce textile designs and ceremonial bronzes that are strikingly reminiscent of Dong Son ornament.

Central Javanese period: 7th to 13th century. Sometime between the 3rd and 6th centuries AD, Indianized principalities existed in Java. The chieftains who lived in their *kratons* (fortified villages) seem to have derived great inspiration, prestige, and practical assistance from the skills and ideas imported from India. In Sumatra there was the important but so far enigmatic Indianized kingdom of Śrivijaya, which, from its strategic position on the Strait of Malacca, exercised a powerful artistic influence in the whole region. Its great Buddhist centre, Palembang, might have had direct connections with the monasteries of southeastern India; for fine bronze Buddhas and *bodhisattvas* in a style reminiscent of Amarāvatī (2nd century AD) have been found in many regions where the influence of Śrivijaya might have been felt, including Mon Dvaravati (see above *Thailand and Laos*) and distant Celebes. Elsewhere among the islands were Indianized kingdoms still unknown to history.

The Indianized art of Śrivijaya

The local dynasties of the *kratons* competed among themselves for power, and eventually the principal dynasties known to history came to the fore. The earliest major cultural assimilations from India took place probably during the 7th century, when the Hindu Pallava form of southeast Indian script was adopted for inscriptions in west Java. Thereafter, a central Javanese dynasty that worshipped Śiva made the oldest surviving artworks in stone. The last king of this dynasty retreated to east Java in the face of the rising power of another central Javanese dynasty, The Śailendra (AD 775–864.) The Śailendra were followers of Mahāyāna and Tantric forms of Buddhism, although Hinduism, as manifested in the worship of Śiva and Viṣṇu, was by no means eliminated. This dynasty created far the larger part of the immense wealth of first-class art known today in Java.

In Indonesia, the word *tjandi* refers to any religious structure based on an Indianized shrine with a pyramidal tower, the type on which virtually all the stone Indianizing architecture of Southeast Asia was originally based. The Javanese, like the Khmer, evolved an elaborate architecture of their own around the basic Indian prototype.

Central Javanese stone architecture did not use structural pillars, nor did its major stone monuments conceptualize hollow space in the way Khmer architecture did. Like Indian stonework, central Javanese stonework is fundamentally conceived as a solid mass, serving as a vehicle for figurative and symbolic sculpture. Its temples are centralized, with enclosures radiating around the central shrine. In eastern Java and Bali, however, the pattern of the shrine was influenced by older traditions and was usually conceived as an enclosure, the walled area of ground being the sacred element, while the buildings in it were of secondary importance. Old wooden buildings do not survive; but representations of wooden architecture in stone reliefs and the recent architecture of Bali show that eastern Indonesia was influenced by the ancient Southeast Asian tradition of constructing wooden pillared halls with tiered, sloping, and gabled roofs.

Because there are no inscriptions to supply dating points, the exact dates of the earliest Indonesian architectural monuments are not certain. The group of shrines generally believed to be the earliest is situated on the Dijeng Plateau. This is a high volcanic region, about 6,000 feet (2,000 metres) above sea level, where there are sulfur springs and lakes. The whole mountain seems to have been sacred to the Hindu deity Śiva, for all temples on the Dijeng are dedicated to him. There can be little doubt that during the 8th and 9th centuries the Javanese, who traditionally had interpreted the volcanic turbulence of their landscape as a manifestation of divine power, identified this power with the terrifying Siva. On other Javanese volcanic mountains, also, groups of shrines are dedicated to him.

Hindu *tjandis*



Tjandi Mendut, near Borobuḍur, Java, *c.* 800.
Louis Frederic—Rapho Guillumette

The temples on the Dijeng are single-cell shrines, roofed with diminishing stories. The exteriors of the temples are relatively plain; only around door frames and window frames are there distinctive passages of central Javanese ornament. Around the niches of Tjandi Puntadewa are perhaps the earliest surviving examples of the characteristic Javanese doorframe: across its lintel is carved a mask of the Indian Kāla monster, which represents time; and down the jambs, as if vomited from his open mouth, run string panels of foliage. The foot of each jamb terminates in an elaborately carved scrollwork cartouche, which is itself a *makara* (water monster) head seen in profile. This *tjandi*, like others on the Dijeng, has a single approach stairway rising between curved balusters. A few stone images of Śiva from these temples have been found. In broad, vigorous forms they express the dangerous power of the god.

Two of the very finest early Javanese sculptures—virtually in the full round—come from yet another Śiva temple, Chandi Banon, near Borobuḍur (see below). One, representing the god Viṣṇu (no stranger in syncretic Javanese temples of Śiva), has the extremely smooth, faintly amorphous suavity, the absolute convexity, and the lack of definition between planes characteristic of the classical central Javanese sculptural style; while the garment he wears, with its assortment of girdles, is closely reminiscent of late Pallava–early Cōḷa Hindu styles of southeast India. Another icon, sometimes called Agastya but more likely the third deity of the Hindu trinity, Brahmā, represents the god in the form of a bearded Brahmin sage. He has a large and, to Oriental eyes, splendid potbelly. This icon was indigenous to southeast India. The great depth of the side recessions of these figures, although perhaps not so clearly defined as in the great Funan–Chenla style (see above *Cambodia and Vietnam*), gives them a bland massiveness. The lack of movement in the figures and the regularity of the designs, the impassive faces, and the slowness of the lines must have been part of the central Javanese conception of transcendent glory.

Buddhist *tjandis*

The Hindu temples of central Java are conceived simply as shrines to contain icons of deities for worship. The Mahāyāna and especially the Tantric Buddhist *tjandis*, however, were called upon to do far more. They were designed to express complex metaphysical theories. The challenge this presented to the central Javanese architects was met in a series of splendid monuments, completely original in conception. The culminating work of the series, Borobuḍur, is a highly evolved architectural image, whose subtlety and refinement were never matched, even at Angkor in Cambodia.

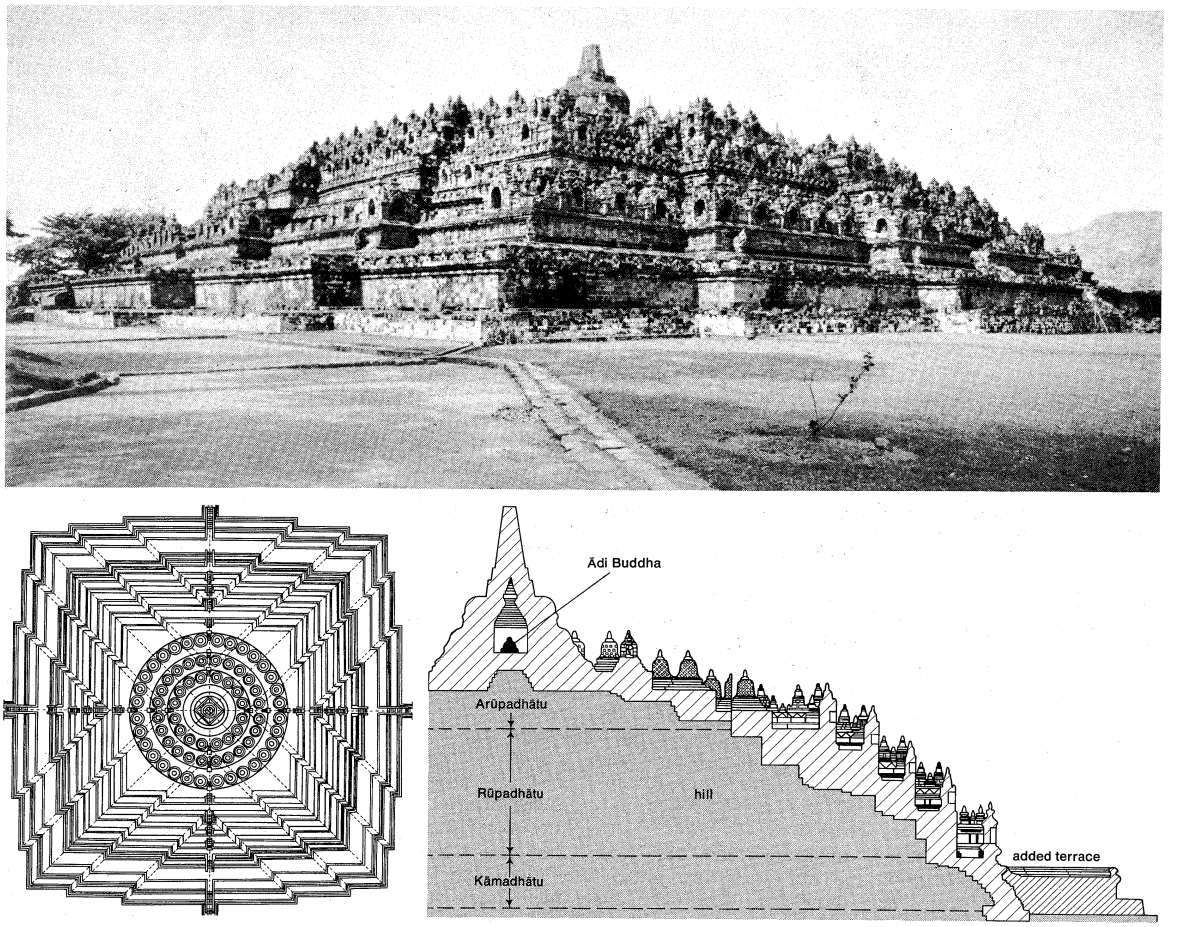
The first work of this Buddhist series is Tjandi Ngawen, near Muntilan. This *tjandi* consists of five shrines facing east, 12 feet (four metres) apart in a row from north to south. Each shrine contained one of the five Buddhas who, according to Tantric Buddhist theory, presides over one of the five major psychological categories under which ultimate reality reveals itself. The shrines themselves are based on but more developed than those used for Hindu deities elsewhere in Java. Roughly square in plan and roofed with diminishing stories, they have pilastered projections on three faces and a portico on the east. Along the architrave are small triangular antefixes and reliefs of Kāla monsters vomiting floral scrolls hood the niches and portals.

The group of five Buddhas is familiar in the art of Tibet, Japan, and northeast India. Among them they compose what is called the *vajra-dhātu*, which means, roughly speaking, "the realm of total reality." According to the old Javanese theology, above this group is another, called the deities of the *garbha-dhātu*. *Garbha* means "womb" or "innermost secret," and its three deities personify the most esoteric realms of Buddhist speculation. At the centre of the group is the image of the single, undivided Buddha nature, which symbolizes the ultimate reality of the entire universe. From his right side emanates the *bodhisattva* Lokeśvara (Lord of the World), who is both compassionate and possessed of all power. From the left emanates the *bodhisattva* Vajrapāṇi, who is the personification of the most secret doctrines and practices of Tantric Buddhism. One of Java's greatest monuments, Tjandi Mendut, is a shrine expressly created to illustrate the combined doctrine of *garbha-dhātu* and *vajra-dhātu*.

Tjandi Mendut

Mendut dates from about 800 and is thus, generally speaking, contemporary with Borobuḍur. It is formed as a single large, square chamber, roofed with the usual diminishing stories, and mounted on a high, broad plinth, which is approached on its northwestern face by a staircase with recurved balustrades. The exterior is in every way more ornate than that of any shrine so far discussed. In addition to floral diaper (an allover pattern consisting of one or more small repeated units of design connecting with or growing out of one another) and scrolls, there are numerous figures in relief representing male and female deities, the subsidiary principles of the combined doctrine of *garbha-dhātu* and Vajrapāṇi. Cut into the fine ashlar (squared-stone) masonry are many relief panels with scenes from Buddhist literature, each panel self-contained and placed with consummate aesthetic judgment. Some represent mythical ideas, such as the wish-granting tree, others narratives from Buddhist legend.

The principal images were placed inside the cell chamber. Apparently, there were originally seven huge stone icons, but only three remain: the central Buddha, who also represented the ultimate Buddha nature of the *garbha-dhātu*, and his two emanations in the *garbha-dhātu*, Lokeśvara, and Vajrapāṇi. When complete, the interior



(Top) Borobuḍur, Java, *c.* 800. (Bottom left) Plan of Borobuḍur and (bottom right) section of Borobuḍur.

(Top) Holle Bildarchiv, Baden-Baden, (bottom left, bottom right) from P. Rawson, *The Art of Southeast Asia*, Thames and Hudson Ltd., London

of Mendut must have been an even more awe-inspiring and spiritually moving place than it is now. The three great statues are seated on elaborate thrones, backed against walls, but the figures are carved virtually in the full round. The inflated, gently inflected forms of the figures give them a majestic presence. The types and carving technique, as well as the monumental scale of the figures, are reminiscent of contemporary work in the cave temples of the western Deccan in India.

On the west–east road from Tjandi Mendut to Borobuḍur stands a small, relatively plain temple called Tjandi Pawon, dedicated to the god of wealth. Pawon was probably a kind of anteroom to Borobuḍur, catering to the more worldly interest of pilgrims. The outside has fine reliefs of female figures, and the roof bears towers of small stupas. On the reliefs are wish-granting trees surrounded by pots of money, and bearded dwarfs over the entrance pour out jewels from sacks.

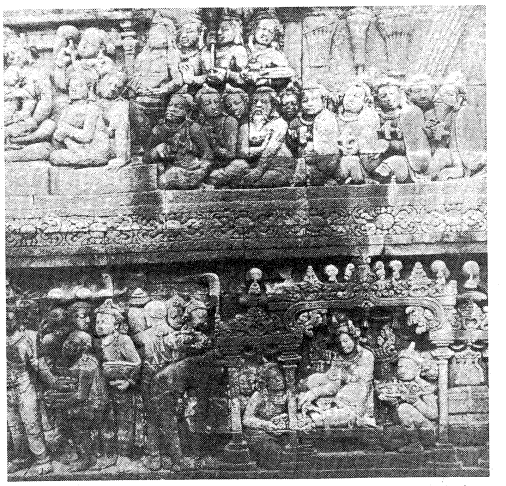
Borobuḍur

Borobuḍur is one of the most impressive monuments ever created by man. It is both a temple and a complete exposition of doctrine, designed as a whole, and completed as it was designed, with only one major afterthought. It seems to have provided a pattern for Hindu temple mountains at Angkor (see above *Cambodia and Vietnam*), and in its own day it must have been one of the wonders of the Asian world. Built about 800, it probably fell into neglect by about 1000 and was overgrown. It was excavated and restored by the Dutch between 1907 and 1911. It now appears as a large, square plinth (the processional path) upon which stand five terraces gradually diminishing in size. The plans of the squares are stepped out twice to a central projection. Above the fifth terrace stands a series of three diminishing circular terraces carrying small stupas, crowned at the centre of the summit by a large, circular, bell-shaped stupa. Running up the centre of each face is a long staircase; all four are given equal importance. There are no internal cell shrines, and the terraces are solid; Borobuḍur is thus a Buddhist stupa in the Indian sense. Each of the square terraces is enclosed in a high wall with pavilions and niches along the whole perimeter, which prevents the visitor on one level from seeing into any of the other levels. All of these terraces are lined with relief sculptures, and the niches contain Buddha figures. The top three circular terraces are open and unwalled, and the 72 lesser, bell-shaped stupas they support are of open stone latticework; inside each was a huge stone Buddha figure. The convex contour of the whole monument is steepest near the ground, flattening as it reaches the summit. The bottom plinth, the processional path, was the major afterthought. It consists of a massive heap of stone pressed up against the original bottom story of the designed structure, so that it obscures an entire series of reliefs—a few of which have been uncovered in modern times. It was probably added to hold together the bottom story, which began to spread under the pressure of the immense weight of earth and stone accumulated above.

Symbolism of Borobuḍur

The whole building symbolizes a Buddhist transition from the lowest manifestations of reality at the base, through a series of regions representing psychological states, toward the ultimate condition of spiritual enlightenment at the summit. The unity of the monument effectively proclaims the unity of the cosmos permeated by the light of truth. The visitor was meant to be transformed as he climbed through the levels of Borobuḍur, encountering illustrations of progressively more profound doctrines the nearer he came to the summit. The topmost terrace, whose main stupa contained an unfinished image of Buddha that was hidden from the spectator's view, symbolized the indefinable ultimate spiritual state. The 72 openwork stupas on the circular terraces, with their barely visible internal Buddhas, symbolize incomplete states





Sculpture at Borobuḍur.
(Left) Stupas on one of the circular terraces. The figure in the stupa in the foreground is a Buddha. (Right) Bas-relief on the exterior wall of one of the circular terraces.
By courtesy of (right) the Indonesian Tourist Board; photograph, (left) Josephine Powell, Rome

of enlightenment on the borders of manifestation. The usual way for a pilgrim to pay reverence to a Buddhist stupa is to walk around it, keeping it on his right hand. The vast series of reliefs about three feet (one metre) high on the exterior walls of the terraces would thus be read by the visitor in series from right to left. Between the reliefs are decorative scroll panels, and a hundred monster-head waterspouts carry off the tropical rainwater. The gates on the stairways between terraces are of the standard Indonesian type, with the face of the Kāla monster at the apex, vomiting his scrolls.

The reliefs of the lowest level illustrate scenes that show the causal workings of good and bad deeds through successive reincarnations. They show, for example, how those who hunt, kill, and cook living creatures such as tortoises and fish are themselves cooked in hells or die as children in their next life. They show how foolish people waste their time at entertainments. From these scenes of everyday life, one moves to the terraces above, where the subject matter becomes more profound and metaphysical. It illustrates important Mahāyāna texts dealing with the self-discovery and education of the *bodhisattva*, conceived as being possessed by compassion for and devoted wholly to the salvation of all creatures. The reliefs on the uppermost terraces gradually become more static. The sensuous roundness of the forms of the figures is not abated; but, in the design, great emphasis is laid upon horizontals and verticals and upon static, formal enclosures of repeated figures and gestures. At the summit all movement disappears, and the design is entirely subordinated to the circle enclosing the stupa.

The iconography of Borobuḍur suggests that the legend of the royal *bodhisattva* recounted in many of the reliefs was meant to "authenticate" some king or dynasty. Yet it hardly seems possible that Borobuḍur was the focus of a specific royal cult, as there is no provision at all for the performance of royal ritual. It must have been, then, in some sense a monument for the whole people, the focus for their religion and life, and a perpetual reminder of the doctrines of their religion.

A considerable number of bronzes, some small, some large, have been found in Indonesia in a style close to that of the sculptures of Borobuḍur and Mendut. One fine, large standing image comes from Kotabangun in Borneo; but some come from Java. Many small cult images of the Buddha and Buddhist deities exist. Some are close in type to the early Pāla images of Indian Bihār, the homeland of Buddhism, with which the Javanese must have maintained close touch. A few small but extremely fine gold figurines of undoubted Javanese workmanship have also turned up. For all their small size they must rate as first-class works of art. As well as images there are many beautiful bronze ceremonial objects, such as lamps, trays, and bells. These objects are decorated with the same kinds of ornament, although on a miniature scale, as the architectural monuments: scrolled leaves, swags, and bands of jewels.

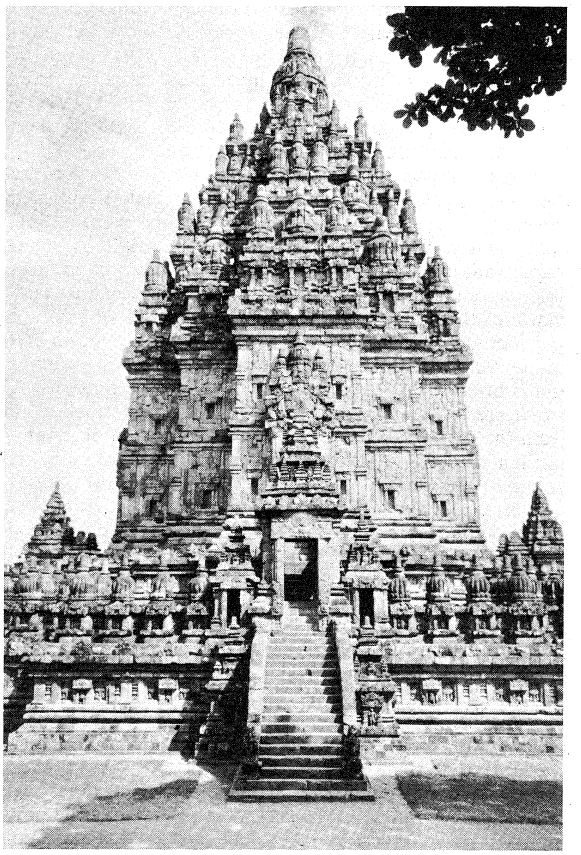
Post-Borobuḍur *tjandi*s illustrate the Buddhist doctrine in different ways. Kalasan, for example, built in the second half of the 8th century, was a large, square shrine on a plinth, with projecting porticoes at the centre of each face. The roof was surmounted by a high circular stupa mounted on an octagonal drum, the faces of which bear reliefs of divinities. Topping each portico was a group of five small stupas, and another large stupa stood at each disengaged corner of the main shrine. The moldings were restrained and elegantly profiled. Each section of the exterior wall contains a niche meant for a figure sculpture. The decorative scroll carving is especially fine.

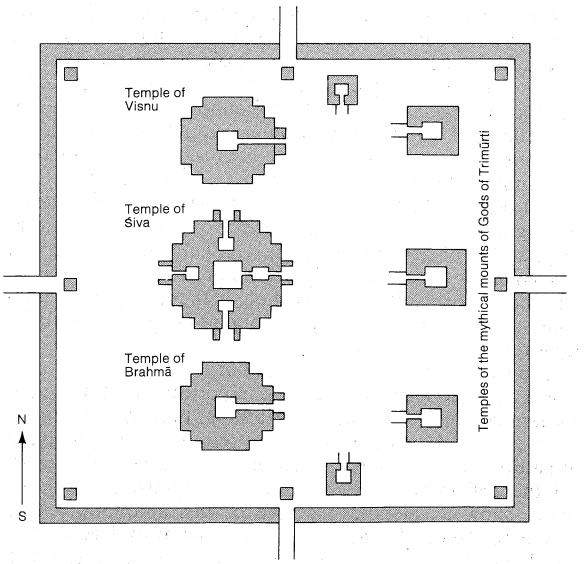
Another shrine from this period, Tjandi Sewu, consisted of a large cruciform shrine surrounded by smaller temples, only one of which has been restored. All of the temples seem to have had roofs in the form of tiered stupas, compressing the overall Borobuḍur scheme into the scope of a storied shrine tower. From Tjandi Plaosan come many beautiful sculptures, donor figures, and iconic images of *bodhisattva*s.

Tjandi Sari

Perhaps the most interesting of the post-Borobuḍur Buddhist shrines of the 9th century is Tjandi Sari. It is an outstanding architectural invention. From the outside it appears as a large, rectangular, three-storied block, with the main entrance piercing the centre of one of the longer sides. The third story stands above a substantial architrave with horizontal moldings and antefixes. Two windows on each short side, three on each long, open into each story, though at the rear they are blind. The windows are crowned by large antefix-like cartouches of ornamental carving based on curvilinear pavilions hung with strings of gems. The uppermost windows are hooded with the Kāla-monster motif. The roof bears rows of small stupas, and perhaps there was once a large central stupa. Inside, Tjandi Sari contains a processional corridor around three interior shrines that were possibly intended for images of the *garbha-dhātu* deities, as at Tjandi Mendut.

The last great monument of the central Javanese period, Lara Jonggrang at Prambanan, is indeed a colossal work, rivalling Borobuḍur. It was probably built soon after 900. Not Buddhist but Hindu, the shrine represents the cosmic mountain. There were originally 232 temples incorporated into the design. The plan was centred on a square court with four gates containing the eight principal temples. Facing east, the central and largest temple, some 120 feet (40 metres) high, was devoted to the image of Śiva. To the north and south it is flanked by slightly smaller temples devoted to the two other members of the Hindu trinity, Viṣṇu and Brahmā. The smaller shrines





(Left) Temple of Śiva, the central temple of the Lara Jonggrang complex, Prambanan, Java, *c.* 900. (Right) Plan of the Lara Jonggrang complex.
(Left) Holle Bildarchiv, Baden-Baden, (right) from F. Wagner, *Indonesia*, Holle and Co. Verlag, Baden-Baden

contained many subsidiary images. The whole complex was enclosed, far off-centre, in an extremely large walled courtyard.

Although these are Hindu buildings, their high-terraced shrine roofs bear tiers of elongated and gadrooned stupas. The reliefs are especially beautiful. One series, representing the guardians of the directions, integrates the ornamental motifs with the plastic forms of the bodies in a most original way. The balustrades and inset panels abound with lively reliefs portraying various deities or scenes taken from the great Hindu classics, especially the *Rāmāyaṇa*.

East Javanese period: 927 to 16th century. During the east Javanese period a very large number of monuments were produced at the eastern end of the island (after 1222) and in Bali (after *c.* 1050). Few single structures, however, are as impressive and as comprehensively planned as are the monuments of Borobudur or Lara Jonggrang.

Around the strange natural mountain with tiered peaks cut and built in stone called Mount Penanggungan there are 81 structures (10th-century) of different kinds (now mostly in ruins). Prominent among these structures are bathing places. This mountain was identified by the people with the sacred Mount Meru, and its natural springs were believed to have a magical healing power and a mystical purifying capacity. Another such bathing place is Belahan (11th century). Made of brick, it, too, has extensive ruined temples. Belahan is supposed to have been the burial place of King Airlangga, who probably died about 1049. One of the greatest east Javanese icons formed the central figure against the back wall of the tank. Carved of red tufa (a porous rock), it shows the god Viṣṇu seated at peace on the back of his violently dramatic bird-vehicle, Garuḍa. It is said that the image represents the King himself in divine guise. Beside this image was a sculpture of a type associated with many of these sacred bathing sites. It is a relief of a four-armed goddess of abundance, her two lower hands holding jars pierced with holes, her two upper hands squeezing her breasts, which are also pierced; through the holes the sacred water flowed into the basin. There are many variants of this idea at the springs of Mount Penanggungan. On Bali the same kind of fountain sculpture appears at the Goa Gadjah, at Bedulu, in a spring-fed tank below a cave.

Four-armed goddess of abundance

In both Java and Bali there are many rock-face relief carvings from this period (there are no secure dates). Some represent legendary scenes; others represent *tjandis*; the shallow chambers of others are thought to be royal tombs.

The structure that gives the best ideas of what the typical east Javanese shrine of the mid-13th century was like is Tjandi Kidal. The nucleus of the building is a square cell, with slightly projecting porticoes each hooded by an enormous Kāla-monster head. But the cell itself is dwarfed both by the massive molded plinth upon which it stands and by the huge tower with which it is surmounted. The tower stands above an architrave stepped far out on tiered moldings. It is no longer composed of diminishing stories, as earlier towers were, but is conceived as a massive pyramidal obelisk made up of double bands of ornament spaced by stumpy pilasters and bands of recessed panels. The architectural projections and moldings distinguish Tjandi Kidal from earlier Javanese architecture, with its plain wall surfaces.

Many masterpieces of sculpture belong to the east Javanese period. Among them are some superb icons of Śiva and of a goddess of Buddhist wisdom from Singhasāri and a splendidly "primitivist" image of the elephant-headed god of wealth from Bara, Blitar.

From the late 13th century onward a whole series of *tjandis* was created in eastern Java. As time went on the *tjandis* lost their monumental scale and became simply shrines within a series of courtyards on a pre-Indian pattern. From Tjandi Djago through Tjandi Panataran at Blitar (14th century) and Tjandi Surawana can be traced the line of descent of the modern Balinese temple enclosures.

Influence of shadow puppets on relief figures

By the end of the 14th century, the figures in the relief sculpture at these shrines had come more and more to resemble the shadow puppets of the popular *wayang* drama. They adopt the stiff profile stance that presents both shoulders, while the trees and houses resemble the stereotype silhouette leather and wood cutouts used as proper-

ties in the shadow plays. The art of carving in the near-full round, however, did not follow the same course of evolution as the reliefs. Such work did become softer and more delicate in style, with accretions of broad floral forms, but well into the 15th century the icons retain something of the strength of older sculptural conceptions. Another plastic tradition that seems to have escaped domination by the *wayang* formula produced beautiful small terra-cotta figures as part of the revetment (stone facing sustaining the embankment) of the east Javanese capital city of Majapahit. Like the reliefs, the many small excavated bronzes of Hindu scenes are under the *wayang* influence, three-dimensional though they may be. Curlicues proliferate, and the plasticity of bodies is virtually ignored.

16th century to the present. *Islām.* When Islām arrived in Indonesia, it used the repertoire of traditional ornament for its mosques and tombs; but, in conformity to a puritan Muslim custom, the representation of living creatures was excluded on religious buildings. The gates of the 16th-century mosque at Sendangduwur, Badjanegara, show a splendid example of this adaptation. The wings of the old Hindu Garuḍa, a colossal bird-vehicle of the high god Viṣṇu, frame the gate; the body and head are suppressed. Above the lintel are abstract tree-clad mountain forms recalling the imagery of the cosmic Meru; and legendary snakes hood the jambs. The 16th-century mosque at Kudus even has a gate based on the split-*tjandi* pattern used in Bali (see below). Tombs such as that of Ratu Ibu at Airmata, on the island of Madura, add to their simple volumes elaborate but abstract variants of the scroll-filled antefixes of older architecture and of the petal-shaped aureoles of the larger east Javanese icons. In Sumatra the Muslim rulers encouraged a revival of the pre-Indian ancestor cult, along with its ancient and characteristic arts.

Bali. The rajas of eastern Java finally retreated before the Muslim invaders during the 16th century and departed to Bali, where they remained. The old Javanese Indianized culture they brought with them survived and combined with animist folk elements. In Bali today that culture has bred a widespread popular art. There are now many hundreds of temples in Bali of varying age. Each family group has its own temple, dedicated to the ancestors; each village, too, has its temple, in which special attention is paid to a rich fertility goddess identified with the ancient Indian goddess of bounty, Śrī. Special temples dedicated to the goddess of death stand near the cremation ground. There are numerous major temples—many associated with volcanic peaks—dedicated to different deities and spirits; they range in size and importance from Besakih on Mount Agung (where a megalith is incorporated as a phallic Śiva-emblem) to Panataram Sasih of Pedjeng (where the bronze drum called "the Moon of Bali" is preserved).

Balinese temples

Balinese temples are conceived as multiple courts raised on terraces. The tall stone or brick and plaster gates are shaped like a *tjandi*-tower split down the centre; they are usually encrusted with ornament based upon deep multiple curlicues interspersed with simplified, two-dimensional relief figure sculpture. Fantastic three-dimensional guardians sometimes stand at the foot of the access staircase. Beyond the gates are one or two courts within which various ceremonies (including sacrifices and cockfights) may take place. The rearmost court backs onto the mountain, whence spirits descend temporarily when invoked. The court has no icons; at most, there is a seat for invisible deities. The structures in the court, mostly of wood and thatch, may be of many stories. (Such structures are called *merus*.) Sometimes the treasuries are ornamented with carving; and a few older stone *meru* towers in local shrines are carved with mythological figures.

Temple ceremonials, especially the cremation of distinguished people, evoked elaborate ritual art objects in precious metals, as well as in wood or fabric. All were characterized by exuberant and repetitive curvilinear floral ornament and by figures based on Indian legend, especially the *Rāmāyaṇa* and parts of the *Mahābhārata*. In the villages today, music, dance, sculpture, and painting are focussed on the shrines and are practiced with an intensity unknown elsewhere in the world. Art is woven intimately into the life of the people. The masks carved of wood for the dances are specially refined, sometimes ornate versions of the masks used in the animist rituals of other Southeast Asian peoples. In the 20th century there are numerous village sculptors and painters, who sell to tourists work based upon the old ceremonial arts. During the 1930s an outside impetus to develop their traditional legendary imagery in Western formats came from a German painter, Walter Spies, who lived on Bali. A landscape tradition was evolved, and the painters have been able to communicate something of the extravagant visual charm of their island, giving glimpses of luminous village and mountain landscape. The style of both sculptors and painters, however, is based upon gently undulating curves and is often highly ornamental, with repeated patterns. A repertoire of posture and gesture has been abstracted from the *wayang*. The work thus tends to prettiness rather than vigour; the sculptors create no truly intelligible volumes, and the painters fill their surfaces with naïvely structured shapes.

Java: 20th century. A conscious revival of traditional art has been attempted in the 20th century, especially in Java, the main territory of modern Indonesia. There has been government support for the resuscitation of old crafts—silverwork, for example. A number of artists have adapted Westernized figure drawing to their own decorative compositions. The best known painter of Indonesia is the Javanese Affandi. He has used oil paint to execute pictures of Indonesian subjects in a vividly coloured Expressionist impasto (thick application of pigment to the canvas). This European brushwork technique, however, contains a strong element of the sweet sinuosity of Javanese tradition. As yet, Affandi is the only artist from Southeast Asia to have attained a personal worldwide reputation.

Indonesian painting

THE PHILIPPINES

The population of this island group contains a number of different ethnic strata, the oldest of which shares in the general folk culture and its associated folk arts of the islands of Southeast Asia (see above *Indonesia*, with an emphasis on geometric simplification. An element in the Tagalog (a people of central Luzon) is perhaps descended from the oldest level of immigrants with a Paleolithic background. The Moro are Muslims, converted to Islām during the 15th and 16th centuries. Today they produce a decorative art in which old Muslim geometric motifs are combined with strong Chinese decorative influences (from Sung times, Chinese ceramics and textiles were imported). The decoration is applied primarily to textiles, weapons, and containers to hold the betel nuts that are chewed throughout Southeast Asia.

Influence of Spain upon Philippine art

The most important departure in Philippine art was the result of the Spanish conquest of 1571. Thereafter, the bishopric of Manila and all of Luzon became the focus for an elaborate development of Spanish colonial art, primarily devoted to the construction and decoration of Roman Catholic churches in the current flamboyant, highly ornate, and colourful colonial style. There is good colonial architecture in other islands, including Bohol and Cebu. A large quantity of religious sculpture of the canonical Christian subjects was imported from Mexico and from Spain itself. Sculptors and missionary painters also immigrated, and a powerful local school developed under the direct influence of the 17th-century Spanish artists Murillo and Alonso Cano. Local arts were encouraged in 1785 by the remission of taxes for religious artists. Because of the close colonial ties, the stylistic developments corresponded substantially with those elsewhere in the Spanish Empire, and European prints served as models for local artists. Of the major early churches for which this sculpture and painting was executed, only San Agustin (1599–1614), in Manila, still stands; it was designed by Fray Antonio de Herrera, son or nephew of the great Spanish architect Juan de Herrera. During the 19th century the neo-Gothic style was imported, mainly through the Philippine architect Felipe Roxas,



San Agustin church, Manila, the Philippines, 1599–1614.
Bruce Coleman

who had travelled in Europe and England. San Sebastian in Manila is a notable example of this style. The Spaniard Hervas, Manila's municipal architect from 1887 to 1893, favoured neo-Byzantine forms; *e.g.*, Manila Cathedral (1878–79).

It was only in the later 19th century that any secular art flourished at all. Schools of fine art modelled on the European schools were set up between 1815 and 1820, and a number of painters began to work in versions of European academic styles, painting landscapes, portraits, and classical subjects. The best known among them are Juan Luna, Felix Resurrección Hidalgo, Antonio Malantic, and the genre painter Fabian de la Rosa. After the transfer of rule to the United States in 1898, industrialization began in earnest; the methods of the art schools were adapted, as in Europe, to the needs of modern commercial society. In the 1930s a substantial modern experimental school of Philippine architects began the remodelling of the industrial environment in terms of 20th-century architectural and design conceptions. Prominent names are Pablo Antonio, Carlos Arguelles, and Cesar Concio. Beginning in the 1930s but especially after World War II, artists in Manila adopted the Abstract and Expressionist styles current in the United States. After the devastation wrought by that war, Manila and other cities and towns were rebuilt, virtually anew, in local variants of the international business style.

Modern Philippine architects

FOLK ARTS

The arts of many regions in Southeast Asia remained either untouched or only slightly influenced by the Indianized arts of other regions. Such influence is found especially in regions where the gold trade flourished. In Sarawak (Bonkisam), for example, the remains of buildings similar to late Tantric east Javanese *tjandis* have been discovered. Among a few people (*e.g.*, the Meo of highland Vietnam), vestiges of Indian erotic temple imagery have been adapted to local fertility ceremonies, and most of the religious ideas of the region show at least faint traces of Indian influence.

Save for the megaliths and Dong Son bronzes, most of the known folk-art objects are relatively recent, although their inspiration and types belong to traditions far older and geographically more far-reaching than the Indianized traditions.

The two main non-Indian art styles in the whole region have been provisionally named the "monumental" and the "ornamental–fanciful." They coexist virtually everywhere, though they probably represent two evolutionary phases. The principal manifestations of the monumental style are the megalithic monuments, although there is great variety among the megalithic customs of the many different populations in Sumatra, Laos, Indonesia, Borneo, and the Philippines. The influence of the ornamental–fanciful style, which is characterized especially by the scrolled spiral, insinuates itself even into many of the decorative arts, particularly in the curvilinear and flamboyant inflection given to ornamental motives in the major Indianizing styles.

The link between the two styles is probably the ubiquitous squatting ancestor figure, cocked knees supporting elbows, carved in soft wood or woven in cane or fibre. These figures may be either male or female. Under special social circumstances in recent times, very large wooden versions of the figure have been used as substitutes for more conventional, standing megalithic ancestral monuments (Sumatra and Sabah). The custom is probably an old one. There can be little doubt, for example, that the Theravāda Buddhist images of Burma, Thailand, and Laos were accepted as special modifications of the ancestor image. The transition from revering numinous ancestor images whose identity had been forgotten to worshipping an Indianizing icon was easy for the tribal populations.

The squatting ancestor figure

The complex significance of the original squatting ancestor figure enabled it to be used in a variety of contexts. It might have combined associations of the fetus, the fetal burial position, and female birth and intercourse positions, as well as a ceremonial posture assumed by the living. It came to be used primarily in wooden sculpture on all scales, but also in woven textiles (*e.g.*, Iban), to represent the continuing power informing human existence, both in the purely ancestral sense of family continuity and identity and in the sense of the fertility of the land. Its earliest recorded appearance may be on Chinese Yang-shao painted pottery (*c.* 2000 BC); but it appears in essentially the same form over a range of territory including Sumatra, Nias and Sunda islands, Java, Borneo, New Guinea, Taiwan, the Philippines, and out into northern Australia and Melanesia. It may be used purely as an ancestral image in a family shrine house or as a motif added to any one of a variety of implements to potentiate them; for example, large bowls (Sumatra), kris or sword handles (Java, Sumatra, Borneo), spoons (Timor, the Philippines), musical instruments (Borneo), and magicians' staves (Borneo).

The treatment of such figures may be invested with more or less of the characteristics of the ornamental–fanciful style, in those regions where this style prevails (*e.g.*, Batak, Dayak). There are also special versions of the squatting figure that seem to belong especially to important magical crafts, such as the Javanese kris handle, on which miniature carvings can give an extraordinarily monumental effect. Sumatran Dayak hereditary magical staves may be carved with a "tower" or "tree" of such ancestor figures. On Nias, for example, along with the squatting figure, a standing figure in the bent-knee posture common in Polynesia also appears as a variant. In the Philippines similar variants are sometimes interpreted as vestiges from a remote Indian mythology, adopted probably for the sake of their cultural prestige. In southern Borneo the figure appears carved in the full round and as a pattern for woven textiles; it often has a protruding tongue and, sometimes, antlers—a combined motif known in the Ch'ang-sha art of southern China (*c.* 300 BC). Antlers also appear on certain Sumatran knife-hilt figures. A variety of designs, some of them "abstract," are based on this figure. Among the Jarai of Vietnam, for example, a pattern of lozenges represents an abstraction from a group of these figures. Especially in the textiles of Sumba and other Indonesian islands, similar patterns, often referred to as decorated triangles, represent the same phenomenon. When, as in textiles, the anthropomorphic reference of the abstract pattern is lost, the male genitals may remain to assert the ancestor significance.

Cult of the skull

The association between the squatting figure and the widely practiced cult of the skull is manifested in the combined cult of ancestors, headhunting, and head worship. Among the Wa of Upper Burma, for example, the squatting figure in a lozenge abstraction decorates the chests in which the severed heads of enemies are stored. Virtually everywhere among the early farmers of Southeast Asia, such heads were regarded as repositories of great spiritual power. The cult of the skull has produced a version of the squatting figure that is commonly known by the Indonesian word *korvar*; it is a figure with an ancestral skull in place of a carved head. Such figures are especially common in the more easterly island cultures. The ghostly power of the deceased ancestor can thus become present and available to his descendants—to give oracular advice, for example. A related idea is incorporated in the masks used in a wide variety of rituals and dance-dramas throughout Southeast Asia; for example, among the Batak of Sumatra and the Dayak of Borneo, where especially fine examples are made. There can be little doubt that the same idea (blended with imagery from the imported Hindu epics) underlies the range of elaborate masks once used in the Javanese and now in the Balinese *wayang* dances. It is possible that the flamboyant flame skull protuberances and winglike flanges ornamenting the head in so much of the Buddhist art produced in Burma and Thailand reflect a persistent but submerged interest in the cult of the skull.

Another major motif is the snake, which (even in areas where direct Indianizing influence was not strong) is frequently combined with imagery derived from the cult of the powerful, magical Hindu *nāga*; often many-headed, this serpent is the patron and guardian of water and

By courtesy of the Royal Tropical Institute, Amsterdam



Brass receptacle in the shape of the mythical serpent, *naga*. From Krui, Sumatra. In the Royal Tropical Institute Museum, Amsterdam. Height 5 cm.

treasure, both material and spiritual. The snake motif has also been blended with images of the Chinese dragon, going back perhaps to Chinese Han ornamental designs. Outstanding examples are found on the elaborate relief-carved doors of Sumatran Batak houses; "flying" roof finials in many parts of Indonesia; and in much Borneo Dayak ornament, from tattoos to carved bamboos and bronze body ornaments. The snake is the magico-mythical creature that gives both its bodily shape (either straight or undulant) and its metaphysical power to the kris. Distributed from Malacca to Celebes, these swords (the earliest known dated 1342) reached their high point of artistic development in Java. A variety of other motifs originating on the mainland of Asia is found in many of the surviving folk arts of Indonesia. Among them are the "man in the embrace of an animal" (Dayak kris handles) and animals "stacked" one above the other (Timor, Indonesia).

The ornamental–fantastic style. The styles in which these variations on basic motifs are carried out vary principally according to the preponderance of the sinuous curves and spirals of the ornamental–fantastic style. This style serves as the basis for decoration and as a method of artistic phrasing. It may have made its way into Southeast Asia as late as the 1st millennium BC, being formally related to the spirals used in Chinese Neolithic, Shang, and Chou bronze art. (It should, perhaps, be mentioned that Chinese art, certainly until well into the Christian Era, was itself far more "tribal" than later Chinese tradition recognizes.) Probably connoting spirituality, the spiral imagery appears in Southeast Asian magical art at all levels, from the textiles of Java and the incised bamboo implements or carved doors of Dayak Borneo to the ornament on the costumes of sculptured dancers or deities at every major city site. Given a fiery upward inflection, it appears in the finials on major Indianized stone architecture and on the carved wooden gables of Burmese and Thai Buddhist halls. There is not always complete stylistic consistency within any one cultural group. For example, the fantastic snake-dragon creatures carved in deep relief on the housedoors of the Batak may be extravagantly sinuous, with many spirals, while their figure sculpture adheres to the sterner plastic idiom, virtually without any linear sinuosity. Among the Dayak of Borneo the fantastic style may be confined entirely to surface ornament. On Indonesian islands, ancestral figures may be relatively static and foursquare, while the decorative carving and textiles may display considerable linear fantasy. A special version of the ornamental–fantastic style characterizes the surviving Indianized arts of Bali and Java, intruding even into sculptural inventions derived from strongly three-dimensional medieval Indianizing patterns. Thus, the decoration on the *wayang* cutout leather puppets, with its somewhat stereotyped curlicues, has proliferated at the expense of the three-dimensional sense (see above *Indonesia*). Balinese *wayang* masks may be carved entirely out of curling surfaces and completed in paint with sinuous eyebrows and moustaches. In many parts of Southeast Asia, including Thailand, Vietnam, Burma, Sumatra, and Indonesia, designs originally based upon Indian flowering-scroll patterns can be found in architecture, textiles, theatre costumes, musical instruments, and wooden utensils, all efflorescing with extravagant curling ornament. It is unfortunately true that only in a few of its most serious manifestations does this kind of ornament display substantial artistic invention, with carefully varied, asymmetrical, complementary, and counterchanged curves. Usually, it degenerates into repetitive space filling, without variety or formal meaning.

Use of curves and spirals

Textiles. Perhaps the types of folk art best known in the West are the textiles, especially batik and *ikat*. Both names refer to techniques practiced by different groups of people, who must have learned it from each other. Essentially Javanese but known in other islands, batik may have resulted from the imitation with dyes of South Indian painted cloths, probably before 1700. The essence of the technique is that melted wax is poured from a small metal kettle onto areas of a plain cotton cloth, which is then dyed, only the unwaxed parts taking the colour. The process can be repeated with several different colours. The oldest basic colours are indigo and brown; red and yellow were used later. The possible patterns range from lozenges and circlets through a large repertoire of cursive animal and plant forms. The batik technique can produce sumptuous and complex designs that not even the most elaborate weaving techniques can duplicate. It was encouraged by the Muslim rulers as a major element of social expression in garments and hangings.

Batik and *ikat*

Ikat is known among the Batak, in Cambodia, and especially among the dispersed Dayak people. It, too, probably originated in India. The extraordinarily difficult *ikat* textiles (woven cotton and occasionally silk, especially in Cambodia) are made primarily for use in important ceremonials and were regarded by their makers as major works of art. Before being woven, the thread is tightly tied at carefully calculated points in the hank (coiled or looped bundle); this is then dyed, the tied parts not taking up dye. The process may be repeated for different colours. As a consequence of the predyeing, designs appear as the thread is woven. In most *ikat* only the warp (the series of yarns extended lengthwise in the loom and crossed by the weft) is so treated; but in southern Suma-

tra a tie-dyed floating weft is added to the plain weft. Naturally, *ikat* designs tend to be static and more or less rectilinear. In the finest *ikat*, however, birds and animals, spirits and houses, and, in Cambodia, a vestigial iconography of royal Buddhism may be formalized into extremely beautiful banded compositions. (P.S.R.)

BIBLIOGRAPHY. REGINALD LE MAY, *The Culture of Southeast Asia* (1954), deals with the art and architecture of Southeast Asian peoples. GEORGE COEDES, *Les États hindouisés d'Indochine et d'Indonésie*, new ed. (1964; Eng. trans., *The Indianized States of Southeast Asia*, 1968) and *Les Peuples de la peninsule indochinoise* (1962; Eng. trans., *The Making of Southeast Asia*, 1966), are standard works.

Literature: (*Burma*): MAUNG HTIN AUNG, *A History of Burma* (1967), contains a general survey of Burmese literature. The same author's *Burmese Folk-Tales* (1948), *Burmese Law Tales* (1962), *Burmese Monk's Tales* (1966); and MAUNG MYINT THEIN, *Burmese Folk-Songs* (1970), provide a good survey of Burmese oral literature. Dramatic literature is treated in MAUNG HTIN AUNG, *Burmese Drama* (1937). (*Thailand*): P. SCHWEISGUTH, *Étude sur la littérature siamoise* (1951), is the standard work on Thai literature. The *Journal of the Siam Society* (various issues) contains studies on Thai literature. (*The Philippines*): LEONARD CASPER, "The Philippine Literature: The Unexplored Potential," *Asia*, 9:80–87 (1967); ALBERT RAVENHOLT, *The Philippines: A Young Republic on the Move* (1962). (*Viet Nam*): NGUYEN NGOC BICH (ed.), "The Poetry of Viet Nam," *Asia*, 14:69–91 (1969). (*Malaysia*): RICHARD WINSTEDT, *The Malays: A Cultural History*, 6th ed. (1961); and OLIVER RICE and ABDULLAH MAJID (eds.), *Modern Malay Verse, 1946–61* (1963), are important works to consult. (*Indonesia*): T.G.T. PIGEAUD, *Java in the 14th Century*, 5 vol. (1960–63), is a special study of Javanese literature in the Majapahit period. See also A. TEEUW, *Modern Indonesian Literature* (1967). TAKDIR ALISJAHBANA, *Indonesia in the Modern World* (1961), contains good accounts of modern Indonesian literature against the political, cultural, social, and historical background.

Music: WILLIAM P. MALM, *Music Cultures of the Pacific, the Near East, and Asia*, ch. 2 and 5 (1967), discusses the music of the Southeast Asian region; LAWRENCE PICKEN, "The Music of Far Eastern Asia and Other Countries," in *The New Oxford History of Music*, vol. 1, *Ancient and Oriental Music*, pp. 83–194 (1957), discusses intercultural and musical relationships between Far Eastern and Southeast Asian countries. (*Burma*): U KHIN ZAW, "Burmese Music: A Preliminary Inquiry," *Journal of the Burma Research Society*, pp. 387–466 (December 1940), appends several pages of music notations to a discussion of music history and the structure of Burmese scales. (*Cambodia and Laos*): ALAIN DANIELOU, "La Musique du Cambodge et du Laos" (1957), is a brochure that discusses with the help of illustrations traditional musical instruments of both countries. (*Indonesia*): JAAP KUNST, *Music in Java*, 2 vol. (1949), an important work with a comprehensive bibliography including works of many Dutch scholars, treats matters regarding history, vocal and instrumental music, structure, notation, and tonal systems of the East-Central and the West-Javanese gamelan; MANTLE HOOD, *The Nuclear Theme As a Determinant of Patet in Javanese Music* (1954), modal structure of patet is analyzed according to basic elements in themes of several gamelan pieces of music; COLIN MCPHEE, *Music in Bali* (1966), gives detailed descriptions and specific musical examples of repertoire played in many gamelan ensembles; WALTER KAUDERN, "Musical Instruments in Celebes," in *Ethnological Studies in Celebes*, vol. 3 (1927), a detailed study with a long list of names of musical instruments, 130 figures, and 19 maps showing the geographical distribution of musical instruments; JAAP KUNST, *Music in Nias* (1939), describes both vocal and instrumental music, classifies musical instruments (Hornbostel-Sachs divisions), and illustrates their distribution in 7 maps; *Music in Flores* (1942), lists in a table the native names of 54 instruments of the five divisions of the archipelago; CHARLES S. MYERS, "A Study of Sarawak Music," *Sammelbände der Internationalen Musikgesellschaft*, 15:296–308 (1913–14), 13 gongs from different cultural groups and music of some instruments are analyzed with the help of musical examples; HENRY L. ROTH, *The Natives of Sarawak and British North Borneo*, 2 vol. (1896)—vol. 1, ch. 9 deals with feasts, festivals and dancing, while vol. 2, ch. 26 discusses music in general; EDWIN H. GOMES, *Seventeen Years Among the Sea Dyaks of Borneo* (1911), terms for songs and names of musical instruments that the author cites are still known in Sarawak; IVOR EVANS, *Among Primitive Peoples in Borneo*, ch. 14 (1922), devoted exclusively to a discussion of musical instruments, music, and dancing; JAAP KUNST, *Music in New Guinea* (Eng. trans. 1967), comprises three works first published in 1931 and 1950 that treat of vocal and instrumental music of the Papua in the north and the central range of mountains, of songs in the north and the West, and of music in the West Central range, Southwest, North and West coasts (many musical examples and a distribution map of musical instruments). (*The Philippines*): JOSE MACEDA, "The Music of the Magindanao in the Philippines," (1963, Ann Arbor University Microfilms), discusses instrumental and vocal music of the whole culture with copious musical examples and two long-playing records available at Folkways Records, New York; and "Drone and Melody in Philippine Musical Instruments," paper read at an International Conference on Traditional Drama and Music of Southeast Asia, Kuala Lumpur, August 27–31 (1969), discusses age, spread, and variety of combinations of drone and melody as a root structure. (*Thailand*): DAVID MORTON, "The Traditional Instrumental Music of Thailand," (1964, Ann Arbor University Microfilms), a study that covers both the historical and structural aspects of Thai music. (*Viet Nam*): TRAN-VAN-KHE, *La Musique Vietnamienne traditionelle* (1962), detailed and important information about history, musical instruments, and musical theory (references abound with criticism of some works); and *Vietnam* (1967), a clear and concise presentation, containing valuable new historical material.

Dance and theatre: JAMES R. BRANDON, *Theatre in Southeast Asia* (1967), a general survey of major theatre forms, incorporating firsthand observation. (*Burma*): MAUNG HTIN AUNG, *Burmese Drama* (1937), a detailed history and translations of plays, four complete and eight in excerpts; KENNETH SEIN and J.A. WITHEY, *The Great Po Sein* (1965), a lively chronicle of the Burmese stage of the last century, cast in the first person narrative, by a famous contemporary actor; U POK NI, *Konmara Pya Zat* (1952), a complete English translation of a 19th-century play with interpretive notes. (*Cambodia*): JACQUES BRUNET, "Nang Sbek, Danced Shadow Theatre of Cambodia," *World of Music*, 11:18–37 (1969), a brief and excellent description and history of Cambodian shadow theatre; SAMDACH CHAUFEA THIOUNN, *Danses cambodgiennes*, 2nd ed. (1956), the most complete history and analysis of plays, music, and dance of female and male dance-drama. (*Indonesia*): BENEDICT R.O'G. ANDERSON, *Mythology and the Tolerance of the Javanese* (1965), major *wajang* characters are described and their significance as behavioral models is treated; JAMES R. BRANDON (ed.), *On Thrones of Gold: Three Javanese Shadow Plays* (1970), translations, with description of action, music indication, and photographs, of three *wajang kulit* plays; CLAIRE HOLT, *Art in Indonesia* (1967), excellent chapters on *wajang kulit*, dance, and dance-drama in all parts of Indonesia; JAMES L. PEACOCK, *Rites of Modernization* (1968), the plays, structure and content, of *ludruk* in Java as seen through the eyes of a modern anthropologist; W.H. RASSERS, *Pañji, the Culture Hero* (1959), a disputed but brilliant theoretical discussion of the origin and meaning of Indonesian theatre; H. ULBRICHT, *Wayang Purwa: Shadows of the Past* (1970), useful for its extended translations of synopses of Pandawa plays found in J. KATS's famous Dutch work *Het Javaansche tooneel*, vol. 1, *Wajang Poerwa* (1923); BERYL DE ZOETE and WALTER SPIES, *Dance and Drama in Bali* (1938), an authoritative and encyclopaedic pre-World War II description of Bali's performing arts. (*Malaysia*): JEANNE CUISINIER, *Le Théâtre d'Ombres à Kelantan*, 2nd ed. (1957), Malaysian shadow theatre described and illustrated, with a partial play translation; RICHARD WINSTEDT, *The Malays: A Cultural History*, 6th ed. (1961), contains background information on Malaysian drama, including translations of pre-performance invocations. (*The Philippines*): JEAN EDADES (ed.), *Short Plays of the Philippines* (1950), a collection of recent one-act plays in English; ALBERTO S. FLORENTINO, *Outstanding Filipino Short Plays* (1961), short plays with two appendixes on traditional and modern drama in the Philippines; FRANCISCA TOLENTINO, *Philippine National Dances* (1946), descriptions and brief histories of many folk dances. (*Thailand*): UBOL BHUKKANASUT (trans.), "Manohra," in *Traditional Asian Plays*, ed. by JAMES R. BRANDON (1972), a translation with stage directions of the *lakon jatri* play "Manohra"; H.H. BRIDHYAKORN, "The Nang," 3rd ed. (1956), a short booklet on *nang yai* shadow play; and with DHANIT YUPHO, "The Khon," 3rd ed. (1956), a short booklet on *khon* masked pantomime; DHANIT YUPHO, *Classical Siamese Theatre* (1952), 52 folk and clas-

sical dance sequences described and illustrated; and *The Khōn and Lakon* (1963), synopses, commentary, and illustrations for 32 classic dance plays as performed by the Bangkok Department of Fine Arts between 1947 and 1960. (*Vietnam*): SONG-BAN, *The Vietnamese Theatre* (1960), brief descriptions of theatre in North and South Viet Nam.

Visual arts: PHILIP S. RAWSON, *The Art of Southeast Asia: Cambodia, Vietnam, Thailand, Laos, Burma, Java, Bali* (1967), a comprehensive survey with many illustrations and plans; BENJAMIN ROWLAND, *The Art and Architecture of India*, 3rd ed. rev. (1967), sets the art of the region in relation to South Asian, or Indian art; WIM SWAAN, *Lost Cities of Asia* (1966), a pictorial study of Ceylon, Pagan, and Angkor; detailed articles in the *Encyclopedia of World Art* (1960–67): GEORGE COEDES, "Burmese Art," "Khmer Art," and "Cham Art"; A.B. GRISWOLD, "Siamese Art"; MADELEINE HALLADE and ROBERT HEINE-GELDERN, "Indonesian Art"; LOUIS BEZACIER, "Vietnamese Art"; and J.M.R. RIVIERE, "Philippine Art"—all these articles have extensive bibliographies. (*Burma*): A.B. GRISWOLD, CHEWON KIM, and P.H. POTT, *Burma, Korea, Tibet* (U.S. title, *The Art of Burma, Korea, Tibet;* 1964), the only survey of Burmese art; BURMA RESEARCH SOCIETY, *Fiftieth Anniversary Publications*, vol. 2 (1960), source material including important articles in English by G.H. LUCE and W.B. SINCLAIR. (*Champa*): The basic studies are PHILIPPE STERN, *L'Art du Champa (ancien Annam) et son évolution* (1942); and LOUIS BEZACIER, *Relevé des monuments anciens du Nord Viêt-nam* (1959). (*Siam and Laos*): Three basic sources include SILPA BHIRASRI, *Thai-Mon Bronzes* (1957), and *The Origin and Evolution of Thai Murals* (1959); and L. BORIBAL BURIBHAND and A.B. GRISWOLD, "Sculpture of Peninsular Siam in the Ayuthya Period," *Journal of the Siam Society*, 38:1–60 (1951). PIERRE DUPONT, *L'Archéologie mône de Dvāravatī* (1959), is the most authoritative study of this topic so far. The chief illustrated source for dating Buddhist art is A.B. GRISWOLD, *Dated Buddha Images of Northern Siam* (1957). (*Indochina*): JEAN BOISSELIER, *La Statuaire khmère et son évolution*, 2 vol. (1955), an exhaustive study of the development of Khmer sculpture; GEORGE COEDES, "Le Culte de la Royauté divinisée . . . ," *Série Orientale*, conference vol. 5 (1952), a basic iconographic study; PIERRE DUPONT, *La Statuaire préangkorienne* (1955), the authoritative book on pre-Angkor sculpture; LOUIS FREDERIC, *Sud-Est Asiatique: ses temples, ses sculptures* (1964; Eng. trans., *The Temples and Sculptures of Southeast Asia* (U.S. title, *The Art of Southeast Asia: Temples and Sculpture;* 1965), a well-documented pictorial survey; BERNARD P. GROSLIER, *Indochine: carrefour des arts* (1961; Eng. trans., *Indochina: Art in the Melting Pot of Races*, 1962), the most comprehensive survey in English; and with JACQUES ARTHAUD, *Angkor, hommes et pierres* (1956; Eng. trans., *Angkor: Art and Civilization*, rev. ed., 1966), a thoroughly documented pictorial survey. (*Indonesia*): F.A. WAGNER, *Indonesia: The Art of an Island Group* (1959), a comprehensive survey of Indonesian art in English; A.J. BERNET KEMPERS, *Ancient Indonesian Art* (1959), a comprehensive illustrated book. Major works on Borobudur include T. VAN ERP, *Beschrijving van Barabudur*, vol. 2, *Bouwkundige beschrijving* (1931); N.J. KROM, *Barabudur: Archaeological Description*, 2 vol. (1927); and PAUL MUS, *Barabudur, esquisse d'une histoire du bouddhisme fondée sur la critique archéologique des textes* (1935), which covers the subject of the role of Indian philosophy and theology as a background to Borobudur.

(M.H.Au./Jé.Ma./J.R.B./P.S.R.)

Southern Africa, History of

The subcontinent of southern Africa is a vast region that includes the independent states of Angola, Botswana, Lesotho, Malawi, Mozambique, Rhodesia, South Africa, Swaziland, Transkei, and Zambia and the United Nations Trust Territory of South West Africa (Namibia).

Until the 1960s the history of this area was treated only in terms of the political struggles of white elites during the colonial period. Since then, however, it has become clear that the history of the region must also take into account the historical experience of all the peoples of southern Africa, over a period far longer than the colonial era. Although the realization poses a good many problems in methodology, interpretation, and organization and there are still many facts to be ascertained, major advances have nonetheless been made in showing the evolution, interaction, and ultimate interdependence, as well as the conflicts, of the diverse peoples who inhabit the subcontinent. This article is divided into the following sections:

I. Southern Africa to 1800
- Early man and the Stone Age
- The coming of the Iron Age
- The Portuguese
- The Dutch at the Cape

II. Southern Africa in the 19th century
- The Mfecane and its effects
- The slave and ivory trade north of the Zambezi
- British occupation of the Cape
- The Great Trek
- The Transvaalers and their neighbours
- The Orange Free State and Lesotho
- The British colonies of the Cape and Natal
- The era of mineral discoveries and confederation
- The scramble for southern Africa
- Angola and Mozambique
- The Germans in South West Africa

III. Southern Africa, 1900–45
- White settlers south of the Limpopo
- White settlers north of the Limpopo
- Land, labour, and colour bars
- The African responses

IV. Southern Africa since 1945
- South and South West Africa
- The High Commission Territories
- British Central Africa
- Angola and Mozambique

I. Southern Africa to 1800

EARLY MAN AND THE STONE AGE

The story of man in southern Africa may go back some 3,000,000 years—that is, close to the beginnings of man himself. Some of the most exciting fossil discoveries connected with early man have been discovered in the subcontinent, suggesting that for hundreds of thousands of years it stood in the forefront of human development and technological innovation. With East Africa, southern Africa has seen the evolution of hominids into the erect being who, deliberately fashioning tools to a regular and set purpose, is commonly defined as man.

The hominids' way of life

For a long time, small groups of manlike creatures roamed the open savanna lands of southern and Central Africa, collecting fruits and berries—and perhaps roots—and also perhaps catching small wild animals to supplement their diet. During the Early Stone Age, the hominids, using first pebble tools and, later, all-purpose stone axes and natural traps, gradually became more skilled as hunters, and by the second half of this period (about 200,000 to 100,000 years ago) they had also become more mobile, making their camps not only in watercourses in the dry savanna and coastal areas to which they had previously been restricted but also in the more heavily wooded country in what are now Zambia and Angola. This extension in the range of suitable habitats was made possible by the controlled use of fire, as is evidenced by living floors at Kalambo Falls dating from 200,000 years ago.

After hundreds of thousands of years of relatively undifferentiated tools, the Middle Stone Age, 80,000 to 100,000 years ago, was one of experimentation, regional development, and diversification. It was at the end of this period that early *Homo sapiens* gave way to modern man. Gradually, too, the large, crude stone tools were replaced over most of the area with a new, broadly similar range of small, light cutting and projectile tools; this culture is known, after its typesite in the eastern Cape Province (South Africa), as the Wilton. Within this single culture, such distinctions as exist are now thought to reflect changing fashions and local adaptations rather than any clearcut cultural or ethnic distinction between the toolmakers. A larger, more mobile population with more varied settlement and subsistence patterns was able to exploit its environment more effectively than ever before.

Apart from the widespread Wilton culture, the Late Stone Age also saw the evolution of a specifically woodland tool industry in northeastern Zambia and northern Malawi, known as Nachikufan. This is characterized by

heavier tools, including bored stones for weighted digging sticks, and suggests greater emphasis on food-gathering than on hunting.

Though there is considerable controversy and terminological confusion over the nature of Middle and Late Stone Age skeletal remains, it seems clear that the Late Stone Age population of southern Africa encountered in historic times—Pygmies, San (commonly called Bushmen), and Khoikhoi (conventionally, but incorrectly, Hottentots)—were basically genetic developments of an ancient African stock that had evolved in the subcontinent. Khoisanoid (a composite term used to refer to the racially identical San hunter-gatherers and Khoikhoi herders) remains are associated with the Wilton culture over much of southern and eastern Africa. There is some slight evidence also of a small non-Khoisan element, which may be associated with the stone bowl cultures of East Africa and just possibly with the origins of sheep and cattle pastoralism in southern Africa.

During the Late Stone Age, hunter-gatherers, organized in loosely knit bands of which the family was the basic unit, achieved greater dominance over their environment by the use of bows and arrows, often tipped with poison, and a variety of snares and traps. With hooks, barbed spears, and wicker baskets, they were able to catch fish and to exploit rivers, lakeshores, and seacoasts in a new way. Each group of hunter-gatherers had its own territory in which special importance was attached to natural resources, such as staple plant foods, water holes, or wild bees' nests; in many instances, bands moved from summer to winter camping sites, following the game and vegetation, as they were to do later in historic times.

Stone Age art

Also associated with the Late Stone Age are magnificent cave paintings and engravings on rocky outcrops throughout southern Africa. Whereas in the northern woodlands the art is stylized and schematic, that of the savanna and coastlands is naturalistic—showing scenes of hunting and fishing, of ritual, and of celebration—and vividly conveys the life of the Late Stone Age hunter-gatherer. It is in the rock art, too, that there appears the first hint of the new groups of herders and farmers who were to revolutionize the way of life that had developed over the past hundreds of thousands of years.

At a number of coastal sites in the Cape Province, dating from about the beginning of the Christian era, there is clear evidence of the presence of sheep, associated with the distinctive pointed base, lugged pottery found in middens from Swakopmund in the north to at least as far as East London, as well as along the Orange River. Although as yet no cattle bones have been discovered so early, one of the outstanding problems for Late Stone Age research is to elucidate the spread of sheep and cattle pastoralism, which was later to be so closely associated with the Khoi herders encountered by the first Europeans to round the tip of the subcontinent. (For further information see AFRICANS, PREHISTORIC.)

THE COMING OF THE IRON AGE

Since there is no evidence in southern Africa of a Neolithic culture similar to that of the east and north, the coming of the Iron Age to the subcontinent appears to have brought with it the food-producing revolution—one of the most significant developments in its history. Agriculture combined with pastoralism supported far larger settled communities than had been possible with Stone Age techniques of hunting and gathering and enabled more complex social and political organizations to develop; and it is with the evolution of these Iron Age farmers' communities into present-day African societies that the precolonial history of the region is concerned.

As in the Stone Age, it is the archaeologist (with the help of linguists, anthropologists, and ethnobotanists) who provides the evidence—incomplete though it may be—of the coming of agriculture, new forms of pottery, and ironworking to southern Africa.

Early Iron Age

Though documentary evidence and oral tradition go back at most 500 years, archaeological finds make it increasingly clear that the complex of traits that is labelled the Iron Age first appeared in southern Africa in the first half of the 1st millennium AD. From the 4th to 7th century AD, pottery, ironwork, traces of agriculture, and the bones of domestic animals appear in many southern African sites, as far south as the Transvaal, Swaziland, and Natal. Virtually no Early Iron Age sites have been excavated in Mozambique, while the couple of Early Iron Age dates for Angola—one of which is for AD 700–800—are disappointingly meagre for an area that shows so many signs of having had a rich Iron Age past. The spread of these Iron Age sites, nevertheless, is generally consistent with what is known of the spread of the closely related Bantu languages; and it is reasonable to suggest that, on the whole, these Early Iron Age sites represent the first settlements of Bantu speakers, the precursors of the majority of southern Africa's present-day population—though in certain areas, especially in the south, ironworking techniques and food production may well have been adopted by the earlier Khoisan population.

Spread of Bantu languages

This movement into southern Africa was part of the wider expansion of Iron Age culture. Penetrating to the woodlands south of the moist forest zone of Central Africa, it seems to have spread rapidly over the subcontinent, to areas with an adequate water supply and soils suitable for cereal cultivation. While the Early Iron Age farmers probably had cattle and therefore had to avoid the dangerous tsetse-fly belts, they also tended at first to avoid the more arid western half of the subcontinent, which was suited only to a specialized pastoral way of life. Here the first Iron Age dates appear to be later than in the better watered easterly regions.

As they spread, the Iron Age minority probably absorbed the earlier Stone Age population in a variety of ways, leaving small remnant groups in the southern and western fringes of the continent. The result is that today the Bantu-speaking peoples of southern Africa do not constitute a racially or genetically uniform group and include a number of physical types; the term Bantu is, in fact, only linguistic.

Most of the Early Iron Age communities in the subcontinent appear to have shared a common culture, which from the earliest times showed considerable regional variation. The regional diversity within the common pottery tradition suggests a rapid differentiation and proliferation; either groups broke off from a central or parent chiefdom, or new trickles of related peoples arrived from neighbouring areas. In the later period, for which oral tradition exists, both processes were at work, and the movement described is extremely intricate and difficult to unravel. It seems likely that the patterns were similar earlier, when the availability of sparsely occupied land led people to found new settlements for various social, economic, and political reasons. The old picture of invading hordes of Bantu speakers moving into southern Africa, brandishing iron weapons, and wiping out the earlier inhabitants is not supported by the evidence, which suggests small trickles of migration and the cultural and genetic absorption of earlier peoples.

From early times, Iron Age farmers made use of metals, settled near mineral as well as salt deposits, and engaged in short-distance trade. Iron implements, such as hoes, and possibly pottery as well, passed from hand to hand, as some communities became known as specialist craftsmen or settled near exceptionally good iron or clay deposits. From potsherds and iron implements found in otherwise Late Stone Age sites, it is clear that a good deal of trade also passed between the Iron Age farmers and the earlier hunter-gatherers. From about the turn of the 1st millennium AD there is a sharp break in the earlier cultural traditions over most of what are now central Zambia, southwestern Rhodesia, and parts of Malawi and the Transvaal, as shown by the markedly new ceramic wares. The new culture, Leopard's Kopje II, was apparently strongly pastoral; for the first time, large numbers of cattle bones appear in the archaeological record, as well as cattle figurines. It is plausible to suggest that this break represents the entry of immigrants into the subcontinent. Again, however, it is not necessary to posit the movement of large numbers of people; the expansion of both human and cattle populations at this time may

Later Iron Age

well have been the result of the absorption, on a relatively large scale, of earlier Iron Age farmers and Khoisan hunter-gatherers and herders into the new culture complex. From this time on, there is a broad continuity in the archaeological record, until colonial times—an essential unity within which widely disseminated political, social, and technological ideas were developed and refined as a result of local circumstances. It was in this period that the foundations of the contemporary Bantu-speaking societies of southern Africa were laid.

From the 12th century on, a number of sites in southern Africa bear witness to the beginnings of more complex state structures, increased social stratification, new religious and ritual ideas, and more sophisticated techniques of agriculture and metalworking, and in particular gold mining, probably initiated by an increase in long-distance trade. From the late 1st millennium, Iron Age cultures included mining in copper and some gold. In Rhodesia alone there are more than 1,100 prehistoric gold mines and 150 copper mines, associated predominantly with Leopard's Kopje II, while in the Transvaal there are literally thousands of prehistoric iron, copper, and tin mines. While most of these date from the 2nd millennium and are associated with the stone buildings of the Uitkomst and Buispoort cultures of the southern and central Transvaal, at Phalaborwa in the eastern Transvaal iron was being mined by the 8th century, and a little later the area became one of the richest sources of copper. By the 10th century, alluvial gold and copper were not simply being traded internally but were also finding their way to outlets on the east coast, though still in small quantities.

Prehistoric mining

In the early centuries of the Christian era, the northeast African coast was well-known to the traders of the Greco-Roman world. These contacts diminished with the rise of Islām, and the east coast became part of the great Indian Ocean trading network; by the 8th century, Arab traders from settlements on the northeast coast began to visit the more southerly harbours, where gold and copper were the most valuable exports. By the 10th century, the gold trade of Sofala (now Nova Sofala) was well known in the Muslim world, and by the 14th century the trade was being controlled by Kilwa, off the coast of what is now southern Tanzania. Between the 11th and 15th centuries some 37 new towns were founded along the coast, and there were trade routes between the coast and the gold-producing area of Penhalonga and Inyanga in present-day Rhodesia, the copper-bearing regions of the interior, and possibly Lake Malawi (Nyasa), with its large herds of elephants.

Though these towns never united politically, they developed a common Afro-Arab or Swahili culture and a splendour that amazed the Portuguese when they arrived on the east coast at the end of the 15th century. On the whole the Arabs relied on African intermediaries to bring their valuable exports to the coast, and a good deal of the trade depended on the older regional networks, although by the 15th century Arab traders had also established overland routes into the interior of south central Africa.

The later Iron Age and its trading networks are best known from three sites in southern Africa—though they may not typify the later Iron Age as a whole. They are Zimbabwe in southwestern Rhodesia, Ingombe Ilede in Zambia, and Mapungubwe, just south of the Limpopo River in the northern Transvaal. These sites all show a broad cultural similarity and were linked to the coastal trade. The most northerly, Ingombe Ilede, on the north side of the middle Zambezi, dates to the 14th or 15th century, and is associated with several settlements of the same people in the extreme northwest of Mashonaland. Its heyday was probably contemporary with that of both Zimbabwe and Mapungubwe. In a strategic position astride the Middle Zambezi and close to mineral and salt deposits, it probably exported chiefly copper. By the middle of the 2nd millennium it was a flourishing trading centre. The archaeological evidence suggests that, as in the case of Zimbabwe and Mapungubwe, the trade was in the hands of a wealthy ruling elite who could afford to be buried in gold and copper ornaments, imported beads, and exotic cloth; their graves are in striking contrast to the stark burials of the common folk at the site. The agricultural economy of the latter closely resembles that of the recent Tonga inhabitants of the area.

Zimbabwe, Ingombe Ilede, and Mapungubwe

It is, however, Zimbabwe, with 60 acres (24 hectares) of imposing ruins, that has dominated the discussion of the Late Iron Age of southern Africa. Long the subject of mystery and romance, the Zimbabwe complex—Great Zimbabwe is simply the largest of more than 150 similar ruins extending from southwest Rhodesia into Mozambique—has now been clearly established as of African origin and of medieval date. A ritual centre until the 19th century, concerned with rain making and the cult of Mwari, the great god of the Shona, in the 14th and 15th centuries Great Zimbabwe was also the centre of a powerful and wealthy empire, which owed its rise to the east coast gold trade. By the latter half of the 15th century, however, Zimbabwe had experienced an abrupt end to its prosperity, and the focus of power had shifted elsewhere.

The first references to Zimbabwe and Mbire, its 14th-century semi-mythical founder, come from oral traditions: from the 15th century the traditions are clearer. Mbire's great-great grandson Nyatsimba, the first *mambo,* or ruler, to earn the title Mwene Matapa (or Mwene Mutapa; also Monomotapa), or Ravager of the Lands, launched ambitious military campaigns to create an empire, which, at its height, is reputed to have stretched from the Zambezi to the Limpopo and from the Kalahari to the Indian Ocean. Zimbabwe hegemony probably also extended over the hill site at Mapungubwe in the middle Limpopo Valley.

Fierce controversy has raged around the racial identity of the occupants of Mapungubwe Valley. Their culture, however, is both contemporary with and similar to the Leopard's Kopje II culture of Rhodesia, whereas the Mapungubwe hill site itself, which dates from the 14th and 15th centuries, belongs firmly to the Zimbabwe tradition, with its extensive stone walling, imported glass beads, spindle whorls, gold, copper, and bronze jewelry, gold sheeting, and solid mud and clay huts.

The reasons for the shift, during the reign of Nyatsimba, of the Zimbabwe empire's focus from southwest Rhodesia to Mt. Fura on the Zambezi bend may be explained by the poor agricultural or mineral resources of the Zimbabwe area; the overextension of its human and natural potential in an effort to maintain the opulence of its buildings and exotic imports; and the importance, by the 15th century, of the Zambezi as a waterway linking the east coast with the interior.

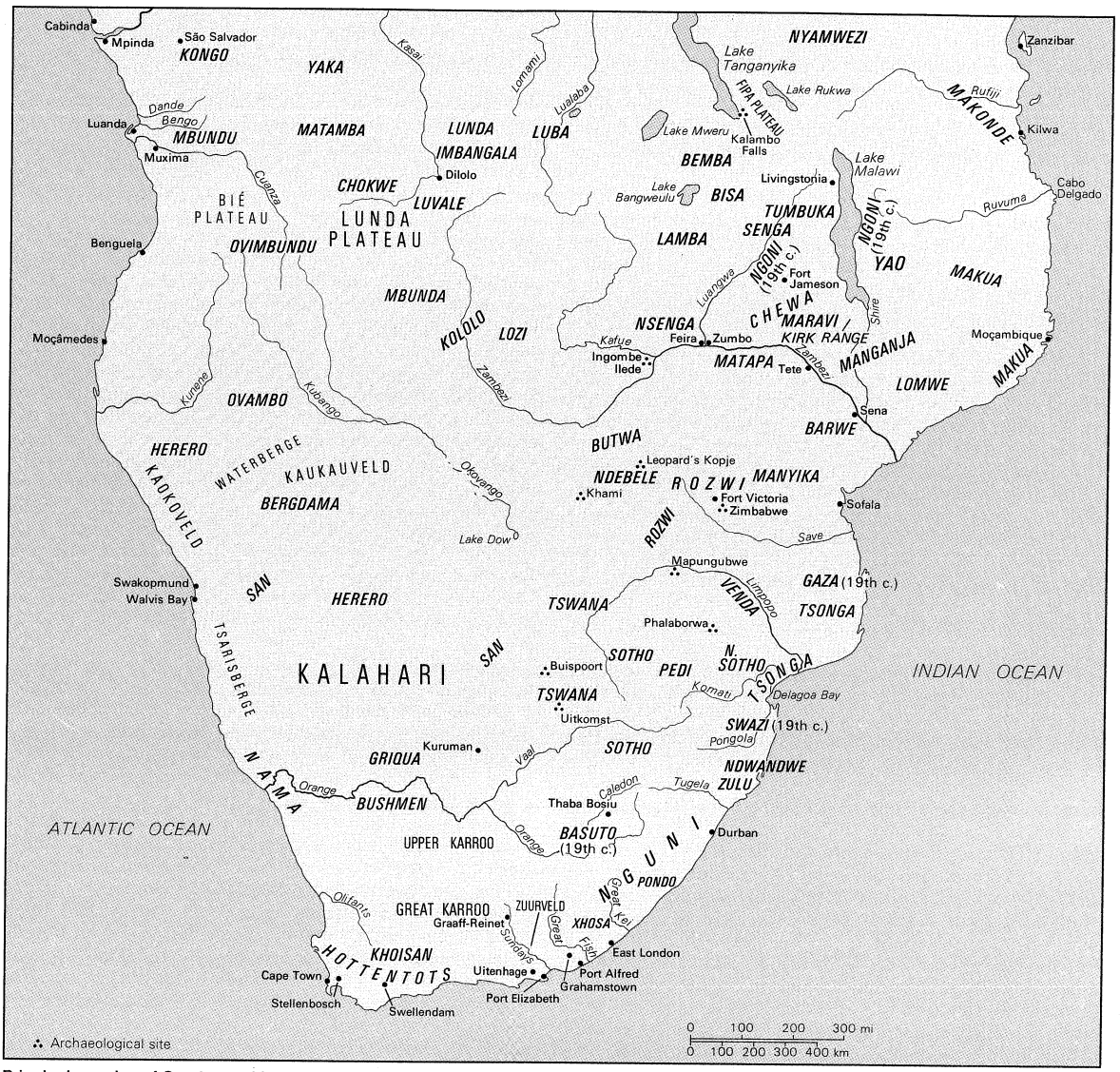
The decline of Great Zimbabwe

Though the empire of the Mwene Matapa owed its creation and consolidation to the prowess of Nyatsimba and to his very able son and successor, it was so large and unwieldy—given the nature of the communications and technology of pre-industrial Iron Age societies—that by the end of the century the Mwene Matapa was forced to relinquish his hold over Butua, the older centre of the empire in the southwest, and to recognize by the first half of the 16th century that power was already more restricted than it had been at the end of the 15th.

By this time, it appears, the Iron Age ancestors of South Africa's contemporary Bantu-speaking peoples to the south had stabilized as farming communities in roughly their present-day habitats, reaching their ecological frontier along the 40–50-centimetre rainfall isohyet on the Highveld and gradually clearing the forests along the southeast coast.

THE PORTUGUESE

In the hope of outflanking Islām, finding a sea route to the riches of India, and discovering additional sources of food, the Portuguese had edged their way around the African coast from the 15th century onward. They reached the Kingdom of the Kongo, to the north of present-day Angola, in 1482–83; early in 1488 Bartolomeu Dias rounded the tip of the continent; and just over a decade later, Vasco da Gama sailed along the east coast



Principal peoples of Southern Africa, 17th to mid-19th century.

of Africa, before striking out to India. Within eight years of their arrival, the Portuguese had seized all of the important centres along the east coast, including Sofala, Kilwa, and Mombasa, and by 1507 they had built a fortress and refreshment station on Mozambique Island, which soon became the key port of call for ships on the way to the Indies.

The Portuguese in southeast Africa. The immediate result of the Portuguese conquests was the economic and cultural decline of the east coast cities. The Portuguese soon discovered that they were unable to control the vast area they had conquered. Throughout the 16th century they faced resistance from the Swahili communities on the coast, and the profits that they had expected from the gold trade at Sofala failed to materialize.

Portuguese expansion

In an attempt to control the trade and to discover the precious minerals for themselves, the Portuguese began to expand into the Zambezi Valley. In the 1530s they established fortresses at Sena and at Tete, where Portuguese captains, elected by Portuguese residents, were delegated certain powers by the Mwene Matapa. Individual Portuguese and Goans, who followed in their wake, were also able to get land grants and jurisdiction rights from local African chiefs. Out of these early grants grew the system of landholding along the Zambezi known as the *prazo* system. Between the 17th and 19th centuries the *prazo* holders became immensely powerful and were able to resist attempts by the Portuguese crown to reform their landholdings and to interfere actively in local African politics. The Afro-Portuguese society of the lower Zambezi Valley that was produced by this process showed remarkable resilience well into the 19th century.

Despite the gradual spread of the Portuguese along the Zambezi Valley and their desire to control the gold trade, there was little impetus for exploration until King Sebastian gained control of Portugal in 1568. In the following year Francisco Barreto was sent as governor to Mozambique with instructions to explore the mineral resources of the Mwene Matapa's kingdom. Despite a military victory in 1572, Barreto's expedition made little impact in southeast Africa, but by the end of the 16th century the Portuguese presence was being felt increasingly, both in the Zambezi Valley and beyond. Though mainly interested in the gold trade south of the Zambezi, the Portuguese also disrupted the ivory trade north of the river.

This disruption and the effect it had on the earlier configuration of power and wealth north of the Zambezi may have contributed to the changes evident in the area of present-day Malawi toward the end of the century. These are chiefly known from the sudden irruption of the so-called Zimba, who appeared along the Zambezi in the 1580s, attacking Sena and Tete between 1592 and 1593 and then the Portuguese settlements along the coast. The Zimba—a term indiscriminately applied to any marauding group—were followers of the *lundu* (chief) living to the south of the Shire River, and their migrations were probably set off by rivalries between the two most powerful Maravi kingdoms in the Shire Valley.

Through the 17th century Maravi was a powerful and widely recognized confederacy, able to dominate the lesser peoples to the east and control the access of traders to Mozambique. By the end of the century the rise of the Yao traders on the east of Lake Malawi and of the Bisa to the west contributed to the disintegration of the Maravi confederacy into several more or less autonomous fragments.

Though the Maravi empire flourished in the 17th century despite, or even because of, the Portuguese presence, direct contact with the Portuguese proved hazardous to the Mwene Matapa. After an unsuccessful attempt by the reigning Mwene Matapa to drive all the Portuguese out of his domains in 1629, he was replaced by a puppet king from whom the Portuguese, taking advantage of internal dissensions, extracted trading and mining privileges. Yet this appearance of Portuguese power and prestige was deceptive, for even vassal kings could rebel, and the Portuguese never had the resources to really control the interior. The *prazo* holders were as much an African as a Portuguese force and acted with increasing independence from the Portuguese crown, while the gold trade probably never did more than balance the administrative costs of Mozambique.

Decline of Portugal in southeast Africa

There were, moreover, external factors weakening the hold of Portugal in southeast Africa. From the beginning of the 17th century the Portuguese faced increasingly severe competition from Dutch and British ships in their colonial waters, while north of Cabo Delgado the Arabs took advantage of Portuguese weakness. In 1631 the first of a series of revolts began, which, by the first decades of the 18th century, had driven the Portuguese from the coast north of the Ruvuma River and had turned African trading interests southward.

Even before this, Portuguese power in Mashonaland had crumbled completely. When, between 1693 and 1695, the Mwene Matapa had, in desperation, called for the help of Dombo, the *changamire* (or ruler) of the powerful Rozwi kingdom (which had been established in the 17th century in the old centre of the Mwene Matapa's empire in southwestern Rhodesia), Dombo responded with alacrity. Sweeping up to the Zambezi, he forced a Portuguese retreat. His intervention, however, was as disastrous for the Mwene Matapa as for the Portuguese. Completely shadowed by the rising power of the Rozwi empire, the Mwene Matapa was reduced to ruling a small chiefdom in the northeastern corner of Rhodesia. The hegemony of the *changamire* was probably recognized over much of present-day Rhodesia until the early 19th century, though the area of polychrome pottery and stone building in southwest Rhodesia probably represents the Rozwi heartlands. Heading a confederacy rather than a unified state, the *changamire* had both religious prestige and military power. His wealth was based on the agricultural and pastoral resources of his kingdom, internal tribute, and trade with his African neighbours as well as with the Arabs, Indians, and Portuguese.

In addition to gold, the Portuguese were also interested in ivory and in other mineral resources. A search for silver mines first led them into Malawi in the 17th century, and by the early 18th century they had penetrated into present-day Zambia. In 1714 they established a trading fair at Zumbo, and in 1732 Feira was founded. Though there were no highly organized broker kingdoms in the area, a number of different African peoples were trading with the Portuguese by this time.

Similarly, from the mid-16th century onward the Portuguese sent trading expeditions down the coast from Mozambique Island to Delagoa Bay, where they found the local Tsonga inhabitants willing and keen intermediaries in the local ivory trade. As the Portuguese position on the southeast coast weakened in the 18th century, their expeditions diminished, and they were supplanted in the dominant trading role by British and Dutch rivals until the end of the century. Nevertheless, the ivory trade and European competition in Delagoa Bay and its hinterland significantly affected political relationships and the rise of state systems in the area in the course of the second half of the 18th century.

The Portuguese in west central Africa. Though the Portuguese influence in west central Africa radiated over a far wider area and was far more permanent, dramatic, and destructive than that on the east coast, it was, paradoxically, initially more constructive. The Portuguese arrived in the north of present-day Angola in 1483 and there encountered the Kingdom of the Kongo, one of the most highly centralized and complex states in southern and Central Africa. The Kongo Kingdom seems to have originated in the 14th century. Like the kingdom of the Mwene Matapa, it was not the only centralized state of its kind in the area, but it is the best known example of similar state structures in west central Africa. At its head was the king, or *manikongo,* around whom everything revolved. The kingdom, which stretched from beyond the Congo River in the north to the Dande in the south, was divided into six provinces, each of which was headed by officials appointed by the king. Below the province was the village, organized around matrilineages. Though the economy was essentially agricultural, it appears that even before the arrival of the Portuguese there was a flourishing trade in shells, salt, and palm cloth. The *manikongo* also appears to have extracted tribute from neighbouring chiefdoms in recognition of Kongo paramountcy.

The Kingdom of the Kongo

Within a few years of their initial contact with the Kongo Kingdom, the Portuguese had conceived a vast scheme for technical aid and education; this was attributable as much to the desire of the Kongo ruling elite to adopt those aspects of European culture they saw as most attractive and useful as to the desire of the Portuguese to spread Christianity and Western education. This period of idealism did not last. The Portuguese had neither the material nor the psychological resources necessary to implement the plan of selective acculturation; and their other innovations, scarcely penetrating beyond the ruling elite, created tensions and schisms in Kongo society. Above all, however, the difference between precept and practice turned the experiment into disaster. Almost from the outset, slave traders followed in the wake of priests and teachers to tie west central Africa to the demands of the transatlantic slave trade. Though the kings of the Kongo and their subordinates all had their household slaves and vassals, Portuguese demands outstripped anything known to their traditional society, as the needs of the plantations established on the West Indian islands and, later, in Brazil grew. By the 1530s, between 4,000 and 5,000 slaves were being exported from Mpinda alone, and by the 17th century the Kongo Kingdom was producing up to 15,000 slaves a year. Attempts by the *manikongos* to stem the outflow of their able-bodied male population simply turned the attention of the raiders southward, toward the ascendant Mbundu (Kimbundu) state of Ndongo. The consequent tensions within the society and the hostility toward the Portuguese erupted into open warfare in 1622, in 1665, and in 1681. Each war left a trail of destruction and fomented social unrest and internal rebellion.

By the end of the 18th century, Kongo had been reduced from one of the most highly organized and sophisticated of west central African kingdoms to a series of fragmented and virtually autonomous villages, although Portuguese influences and Christianity lingered on among the elite at the capital, São Salvador do Congo. In the second half of the 19th century, the Kongo area, with its impoverished rival and warring local chieftains, was annexed as the northern province of Angola.

The end of Kongo

If Portuguese pressure resulted in the progressive disintegration of the Kongo Kingdom, it was ultimately equally disastrous for the Mbundu people. They had been united with the Ndongo kingdom, on the Lunda Plateau to the south, in the course of the 16th century, under a ruler called the *ngola.* After some 50 years of sporadic contact with them, the Portuguese, believing reports of great mineral wealth in the interior and anxious also to acquire the salt mines of Ndongo, expanded their political dominion to Ilha de Luanda. In 1575 Paulo Dias de Novais was appointed captain general, or governor, of an undefined area between the Dande and Cuanza rivers, which became known as Angola, and granted wide per-

sonal control over any areas he might conquer or colonize. He founded what became the city of Luanda in 1576. Portuguese plans to colonize the kingdom of the *ngola*s were greatly hampered, however, by tropical disease, the unsuitability of the region for agriculture, and strenuous African resistance to settler encroachments.

Though the *ngola*s were initially eager for contact with the Portuguese, they soon discovered that their immediate presence spelled destruction. Within a few years the first war of a century of almost constant warfare had begun. Originating as a means of territorial expansion and establishment of Portuguese control, the wars soon resolved themselves into efforts to acquire slaves on Portuguese terms. Almost from the outset, the Portuguese relied on African auxiliaries or allies in their campaigns. Though these "allies" were capable of considerable independent action, they had to pay tribute in slaves, ivory, and probably salt.

The effects of the slave trade

Although it is not possible to compile an exact balance sheet of the devastation caused to west central Africa by the slave trade, by the early 17th century some 5,000 to 10,000 slaves were being exported from Luanda annually —and this figure does not include those smuggled out of the bays and inlets along the coast in order to escape official taxation. The figure may indeed have represented a relatively small proportion of the total population of the area, but it also represented the economically active young adults. While it has been argued that there—as elsewhere in the slave-trading societies of Africa—their loss was balanced by new food crops, such as manioc and maize (corn), which the Portuguese imported from South America and which probably led to increased populations, in parts of Angola the very direct and damaging campaigns of the Portuguese and their African allies probably made slave trading more destructive than elsewhere. Accounts of Ndongo in the 16th century as rich and populous give way to lamentations about its desolation in the 17th. In addition to the direct depopulation wrought by the slave trade, by the 1630s large groups of people had moved out of the territory, beyond the raiders' reach.

Of these migrations, the two best documented, and perhaps the most significant, were the movements of a section of the Mbundu, under Queen Nzinga, and that of the Mbangala, under their ruler, the *Kasanje*. Nzinga, who has been described as one of the most remarkable women in Central African history, dominated the politics of the area for nearly half a century. When, in 1618, the Portuguese finally conquered the Ndongo kingdom, Nzinga attempted to use the Portuguese to secure her leadership in Mbundu politics. When this failed, she moved eastward with her followers. Conquering the state of Matamba, she developed it into a powerful slave-trading centre, with networks reaching north to the Kongo and Luanda and eastward to the Lunda of Katanga.

At the same time as Nzinga's migration, the Mbangala (Imbangala) established their new state of Kasanje in the Kwango Valley. Like the ferocious and cannibalistic Jaga raiders who had disrupted the Kongo Kingdom in 1569, the Mbangala probably originated in the Luba-Lunda kingdoms of Katanga, though their date of departure is thought to be the late 15th century. By the end of the 17th century, Kasanje had become the leading slave-trading kingdom in the area and the chief entrepôt between the coast and the Lunda in the east. Like the *manikongo* and the *ngola* of Ndongo, both Nzinga and Kasanje welcomed the Dutch conquest of Luanda in 1641, though their relief was short-lived. By 1648 the Portuguese had reconquered Luanda and were thus free to reconsolidate their hold on their African possessions. Nzinga herself ended her days as a Roman Catholic collaborating with the Portuguese in the slave trade; in 1681 her successor as *ngola* of Matamba was finally conquered, 10 years after the remnant Ndongo kingdom farther west had been completely dismantled.

At the beginning of the 17th century, as they were penetrating inland from Luanda, the Portuguese were also exploring southward in the hopes of expanding the slave trade and of finding a navigable waterway across the continent. In 1617 they established a colony at Benguela, which, like the Kongo Kingdom, was annexed as part of Angola in the 19th century. There was, however, little expansion inland from Benguela until the 18th century, when soldiers followed slave traders to the fringes of the Ovimbundu kingdoms. Though very little is known of the early history of the Ovimbundu kingdoms, they appear to have originated in the late 16th and 17th centuries with the arrival of Lunda, Mbangala, and Mbundu raiders and refugees fleeing from the Portuguese. By the end of the 17th century the slave trade from Benguela may have exceeded that from Luanda, while in the 18th century the Ovimbundu rivalled Kasanje as slave-trading kingdoms, with routes going from Ovimbundu not only to the official Portuguese outlets but to Dutch, French, and British interlopers at the coast as well.

The Ovimbundu kingdoms

These interlopers were a major concern of the Portuguese in the 18th century. Portuguese attempts to control the coastal trade and Ovimbundu resistance led to a war and to the Portuguese conquest of the Ovimbundu heartlands between 1774 and 1776. By this time the Ovimbundu were a formidable enemy, having adopted Portuguese military strategies and being well equipped with firearms. Despite their military victory on this occasion, the Portuguese were unable to exert political control over the Ovimbundu until more than a century later.

Through the 18th century and for the first third of the 19th, the slave trade remained at the centre of Angola's economic existence. Routes stretched well into the centre of Africa, with most of the trade in the hands of the African and half-caste intermediaries, known as *pombeiros*. The area of actual Portuguese colonization was smaller, consisting of a number of estates huddled round the inland forts and along the Bengo and Dande rivers. The estate holders were never to acquire the position and following of the *prazo* holders of Mozambique.

The late 18th century saw Portuguese fortunes in both Angola and Mozambique at a low ebb. In both, they found themselves unable to control the trade, either inland or on the coast. The number of settlers was small—about 1,000 in each colony—and the economies were poor and the administrations weak and corrupt. Though both these territories recovered somewhat during the Napoleonic Wars, it was not until the last quarter of the 19th century that the Portuguese regained anything of their initial colonizing energy.

Despite Portuguese mythology that the Portuguese did not differentiate according to race in their colonies, from quite early times whites had superior status, prestige, and power in Angola as in Mozambique; attitudes toward race in the 17th and 18th centuries were perhaps not so different as is generally asserted from those of the Dutch colonists. (The history of west central Africa during this period is also dealt with in the article CENTRAL AFRICA, HISTORY OF.)

THE DUTCH AT THE CAPE

In some ways, the end of the 18th century also found the only other enclave of European settlement in southern Africa threatened, though for very different reasons. This was the colony at the Cape of Good Hope, established as a stopping point on the way to India in 1652 by Jan van Riebeeck and some 125 men on behalf of the Dutch East India Company. In the late 16th and 17th centuries, as the Cape had become a regular port of call, the crews of European ships there found herders ready to barter some cattle in exchange for iron, copper, beads, tobacco, and brandy. By the time of the Dutch settlement in the mid-17th century, a class of Khoikhoi (Hottentot) intermediaries had grown up and a network of trade stretched far into the interior. Relationships established in this half century of trading profoundly affected the nature of contact between the Khoikhoi and the Dutch when the Dutch East India Company established its outpost there.

The Cape of Good Hope

Though initially intended as a refreshment station, almost imperceptibly and despite the company's reluctance, the outpost grew into a colony of settlement. In 1657

the company released a number of its servants to become free burghers in order to cultivate land and herd cattle on its behalf; and in the same year the first slaves were imported into the colony from Angola and the East Indies to perform the labour which the San (Bushman) and Khoikhoi (Hottentot) inhabitants were reluctant to undertake. In contrast to the Portuguese, the Dutch feared to imperil the cattle trade on which the colony depended, and from the first they prohibited the enslavement of the indigenous population. Moreover, at the Cape there was no indigenous slavery and local chiefs had neither the people nor the power to engage in the slave trade.

Dutch expansion and the trekboers. If the Dutch did not raid the local inhabitants for slaves, they were soon to undermine the loosely organized social structure of the Khoikhoi by their increasing demands for cattle and by encroaching on their grazing and hunting lands. As one Khoikhoi group became impoverished and reluctant to trade, another would take its place. Though the Khoikhoi did not crumble as rapidly after the arrival of the Dutch as is often suggested, and offered considerable resistance to Dutch expansion, their small numbers enabled the Dutch to move beyond their base at the coast far more easily than the Portuguese had been able to do in Angola. Moreover, the gentle Mediterranean climate of the Cape was well suited to European settlement. While the Portuguese in Angola and Mozambique were ravaged by tropical diseases, at the Cape it was the non-immune Khoisan who were affected by epidemics brought by the whites.

The greatest geographical barrier to Dutch expansion was the ranges of mountains inland from Cape Town. Once these had been crossed and settlers had ascended the flat inland plateau lands, expansion to both east and north was relatively rapid. There, in the arid scrubland, economic necessity and geography dictated a semi-nomadic, pastoral way of life for the white settler, as it had for the indigenous peoples before him. Poor soil and inadequate rainfall could be compensated for by the appropriation of vast farms, and 6,000 acres (2,400 hectares) rapidly became the norm.

Stellenbosch, Swellendam, and Graaff-Reinet

Despite one or two sporadic attempts, the Dutch East India Company made little effort to follow the cattle farmers, or trekboers, as they came to be known. Though the new districts of Stellenbosch (1685), Swellendam (1745), and Graaff-Reinet (1785) were formed, they were large and unwieldy, and their centres were far from the expanding trekboers. Governmental authority was remote, and the trekboers, though they never built up the same kind of power as the *prazo* holders of Mozambique, became accustomed to handling emergencies on their own and to ruling over their slaves, servants, and clients as they saw fit.

The numbers of white settlers at the Cape grew slowly in the 17th century, but in the 18th, largely through natural increase but also through immigration, they rapidly outstripped the thousand or so whites in Angola and Mozambique. From 200 inhabitants in the 1660s they increased to between 4,000 and 5,000 by the 1750s, and to about 15,000 in the last decades of the century. Despite the different origins of the inhabitants, an official policy of integration, the shared vicissitudes, and their common Calvinist religion molded the whites into a relatively homogeneous group that became the Afrikaners. By the 18th century their differences were socio-economic rather than ethnic. Tensions between the various groups of settlers and the company to some extent contributed to a sense of identity from the beginning of the 18th century.

As the white population increased, so did the number of slaves. Experiments in the use of white labour were unsuccessful; and, as in the Portuguese colonies, working for a master became the stamp of an inferior status, which was inevitably associated with a black skin.

As in the Portuguese colonies, too, there was no bar in the early days of the colony to intermarriage between the Dutch and either the Khoisan inhabitants or freed slaves, and religion may have been the determining factor in ethnic identification. By the late 17th century, intermarriage was officially frowned upon, but concubinage with nonwhite women continued. The mixed population that resulted became known as the Cape Coloureds—a new amalgam, sharing many of the cultural attributes of its white, slave, and Khoisan forebears.

Assimilation of the Khoisan peoples. The Khoisan peoples of the Cape were quick to assimilate new cultural traits—indeed, their virtual disappearance as an ethnically identifiable group by the 19th century can, in large measure, be attributed to their genius for acculturation. As they lost their cattle and grazing lands, they became virtual serfs of the white man, on the vineyards and wheat farms of the Cape and on the cattle farms of the interior, though some groups were able to escape by moving both north and east. In the east they continued the long process of mingling with the Xhosa peoples, while on the Orange River, where they became known as Griqua in the early 19th century, they established a number of polities that bore a strong resemblance to the republics later established by the Dutch settlers.

Khoisan response to the Dutch

Throughout this period the responses of the Khoisan peoples to the Dutch displayed many of the same complexities and ambiguities that the Bantu-speaking peoples of Angola displayed toward the Portuguese. On the one hand, they were eager to trade and therefore to collaborate with the Dutch. On the other, when the demands of the intruders became intolerable, they opposed them by trying to retain control over the terms of trade. This opposition erupted into guerrilla warfare in 1659 and the 1670s. Though many secondary sources suggest that thereafter Khoisan resistance was at an end, it is clear from the documentary evidence that the cattle raids that punctuated almost every decade of the 18th century were a continuation of this guerrilla warfare.

The raids and counter-raids mounted in violence until, in the last quarter of the 18th century, the entire northern frontier of the colony was under arms and some white farmers were being driven out of their lands. Those who were not ruined by the cattle raids were being worn down by constant military service, which was left to the settlers themselves. Though ultimately the Khoisan peoples were vanquished by pure bloodletting, they prevented further colonial expansion northward for some 30 years.

Trekboer–Xhosa conflict. This was the more serious in that white expansion came up against a major block on its eastern flank at the same time that the trekboers encountered the populous and settled agricultural Xhosa people in the area of the Fish River. Trading contact had existed between the trekboers and the Xhosa from quite early in the 18th century. In the last quarter of the century, friction between black and white increased, over grazing grounds, water, and the terms of the cattle trade. Demands for fresh meat at Cape Town increased during the Anglo-French wars and during the American Revolution; Khoikhoi reserves of cattle had been exhausted, and the trekboers had now expanded up to the very edge of Xhosa settlement. Various attempts to draw a frontier between the colonists and the Xhosa were equally useless; by the time the governor himself visited the eastern frontier in 1778 and decreed the upper Great Fish River as the boundary between black and white, there were Xhosa to the west of the Great Fish River and trekboers embedded in Xhosa society to the east of it.

The establishment of the new district, Graaff-Reinet, with a full-time government official, hardly improved matters. The area of his jurisdiction was vast, its frontiers ill-defined, and its inhabitants hardly disposed to listen to the commands of a government that had left them to their own devices for more than half a century. Before the 18th century was out, minor cattle raids had developed into two frontier wars. They were the prelude to a struggle that lasted almost 100 years, and ultimately the trekboers were able to re-expand only after outflanking the Xhosa in the 19th century.

Despite the firearms of the Dutch, the two combatants were by no means unevenly matched, for the Xhosa were superior in numbers and social organization. The trekboers suffered from internal divisions and were facing an even more severe threat from the Khoisan resisters on their northern frontier. The commando system, which

had evolved through the 18th century as a method of local defense, was ill adapted to concerted warfare against a well-organized foe. Nor was the Dutch East India Company prepared to aid the frontiersmen. As a result, the last quarter of the 18th century saw a stalemate, with the trekboers probably getting the worst of the encounters.

Between 1799 and 1802, the third and most disastrous war erupted, when conflict with the Xhosa coincided with an uprising of Khoisan servants, who deserted with their masters' guns and horses. By this time the Cape had passed temporarily into the hands of the British (1795–1803), who were anxious to bring the war to an end as swiftly as possible. They managed to divide the African forces by making land and labour concessions to the Khoisan peoples but were quite unable to dislodge the Xhosa from the area west of the Great Fish River. It was only with the appearance of regular British troops on the Cape frontier in 1811 that the balance was tilted more definitely in favour of the white man, though the Xhosa were not finally defeated until well into the 19th century.

II. Southern Africa in the 19th century

After 1800, new internal and external forces radically changed the demographic, social, political, and economic configurations established over the preceding two centuries in much of southern Africa. Though certain areas were not to be affected by change till the last decade, for the area as a whole the first three-quarters of the 19th century brought many new developments. Some, like the rise of the Zulu Empire, which sent its offshoots in every direction, and the expansion of white settlement from the Cape, reversed the north–south direction of change in Iron Age times. Others, like the intensification of the slave trade from the east coast and the parallel drive of ivory hunters into west central Africa, were more a shift of emphasis than totally new developments.

19th-century missionary activity

The 19th century, too, was the age of missionary expansion. In the footsteps of traders came Christian missionaries, who, by the end of the period, had reached most of the peoples of southern Africa, not only evangelizing and educating a new elite but frequently acting as the closest political advisers of the old, as well as preparing the ground for the age of European territorial acquisition.

THE MFECANE AND ITS EFFECTS

The first of these changes began at the end of the 18th century in the area of present-day Zululand, where the Zulu were simply one of numerous Bantu-speaking Nguni chiefdoms, closely related to the Xhosa on the Cape's eastern frontier. Yet their rise to power in the second decade of the 19th century profoundly altered the demography and political relationships of the entire subcontinent. Their expansion and its repercussions are known in the south as the Mfecane—the "Crushing"—a term that vividly conveys the destructiveness and social disorder caused by marauding tribes on the move. For many, the choice was to join the invader or to set off themselves. Yet out of the Mfecane new polities were formed; and the invaders, for those not crushed by their advent, brought a challenge and an example of military and social organization that the invaded were to turn to advantage in resisting later intruders.

Though it has been customary to attribute the Mfecane to Shaka (*q.v.*), who became king of the Zulu people in 1816, the entire pattern of the Mfecane, the way in which tribe was sent reeling against neighbouring tribe once he began his campaigns, suggests that Zululand was an area of high population and cattle density in relation to its ecology and technology; and it may be that the pressure of people and cattle on land at a time of reduced rainfall gave the rise of the Zulu its explosive force.

By the end of the 18th century there seem to have been five or six larger chiefdoms in Zululand that were beginning to absorb their neighbours, but these new corporate kingdoms had much in common with the smaller chiefdoms that had preceded them, and they employed diplomacy as much as warfare to achieve their ends.

All this was changed, however, when Shaka became chief of the Zulu people, at that time simply one of the tributary states of the wider Mtetwa (Mthethwa) confederacy, ruled by Dingiswayo, their paramount chief. Shaka, who was a military genius, revolutionized traditional methods of warfare. In the Mtetwa confederacy and, indeed, in many of its neighbouring chiefdoms, regiments composed of men of the same age and recruited from all the villages and clans of the kingdom were now organized for war. In addition, Shaka was a master of strategy and surprise tactics. Fighting became a matter of total war, not simply of limited raidings.

In 1817 Dingiswayo was killed, and Shaka took over the Mtetwa hegemony. In a short time he had brought under his control all of the peoples within striking distance of Delagoa Bay to the north and Pondoland (now part of the Transkei) to the south. Out of the numerous chiefdoms between the Tugela and Pongola rivers he welded a highly centralized military kingdom; to the south he created a buffer zone of havoc and devastation.

Zwide and Shaka

In the first years of his empire building, the Ndwandwe confederacy, under Zwide, was Shaka's most formidable rival. Indeed, even before his major campaigns, the Ndwandwe had pushed the Ngwane and Dhlamini under their chief Sobhuza to the north of the Pongola River, where Sobhuza and his successor, Mswati, conquered and absorbed the local Nguni and Sotho chiefdoms and created the Swazi nation. Though under frequent attack from the rising Zulu power, the Swazi, by withdrawing northward to rocky fortresses, escaped the fate of those who remained within the Zulu orbit.

By 1820 Shaka had decisively beaten Zwide's armies, which then broke into three groups. Zwide himself retired, but two of his commanders, Soshangane and Zwangendaba, fled northward, to be followed shortly thereafter by Nxaba, heading the Msane and Maseko peoples. Battles between the three on the middle Save River left Soshangane victorious, and he proceeded to build up the Zulu-type Gaza empire, which at its height stretched between the Zambezi and Komati rivers, reduced both the Portuguese and the Shona to tributary status, and incorporated many of the Tsonga and Shona peoples of south Mozambique.

After their defeats, both Zwangendaba and Nxaba swung westward, fighting one another and totally disrupting the old Matapa and Rozwi empires. In 1835 Zwangendaba crossed the Zambezi, while Nxaba's followers split in two, one section remaining in what is now Rhodesia, the other eventually settling in the highland country of the Kirk Range in the 1860s. From that location they raided a wide area west of the Shire and, after 1875, to the east of Lake Malawi (Nyasa) as well.

Zwangendaba and the Ngoni

Across the Zambezi, Zwangendaba's Ngoni—as these offshoots of the Nguni came to be known—raided the defenseless Chewa, Nsenga, and Tumbuka people, incorporating some into Zwangendaba's following, though the original Nguni nucleus retained its political and cultural dominance. In about 1845 Zwangendaba died, having established a settlement on the Fipa Plateau.

On his death, the tensions inherent in a constantly raiding society came to the fore, and a bitter and complicated succession dispute culminated in the fragmentation of his kingdom. Two groups now moved into present-day Tanzania, while Zwangendaba's sons, Mpezeni and Mbelwa (or Mombera), moved southwestward and southward, respectively. Having conquered the already enfeebled Bisa, Mpezeni clashed fiercely with the Bemba, by then becoming one of the most powerful polities in northeastern Zambia, and then moved to Fort Jameson (now Chipata) in about 1865, conquering more Nsenga and Chewa peoples en route. Mbelwa, with the largest fragment of Zwangendaba's people, moved back into Tumbuka country, where in 1855 he overthrew the Kamanga confederacy, already weakened by the activities of Swahili slave traders. From his vantage point in the healthful uplands, he also raided to the south and west. Although his composite following threatened to break up into its component parts in the 1870s, Mbelwa's Ngoni maintained their cohesion into colonial times.

While these fragments of the Ndwandwe confederacy were making their way northward into east central Africa, chaos was being spread southward and westward by Shaka's campaigns against the chiefdoms of Zululand and Natal. The ever-widening circle of disorder reached the westerly Sotho and Tswana groups across the Vaal, in Botswana, and in the northwestern Cape as raiders and refugees ricochetted in all directions.

Many moved south to join fugitives making their way from Shaka's campaigns in Natal to the territory east of the Cape Colony. All of these refugees, whose social structures had been disrupted if not shattered by the years of warfare, became known as Mfengu (or Fingos). Their presence added another complicating factor to the tense situation on the Cape eastern frontier. Some shattered peoples scrambled to safety in the mountains of modern Lesotho, where Mshweshwe, one of the most outstanding of the leaders produced in these years, built up a kingdom based on the rocky fortress of Thaba Bosiu.

Other dislodged Highveld peoples joined the Griqua polities along the Orange River or continued raiding along the Vaal and in the western Transvaal, where the disorders prepared the way for the coming of Mzilikazi, who, with 300 of his followers, who became known as the Ndebele (or Matabele), moved out of Zululand in 1823. Over the next 15 years or so he moved westward across the Transvaal in flight from the Zulu armies, until finally, in 1837, after a triple conflict with the Zulu, with Griqua raiders from the Orange River, and with the expanding white farmers, or Voortrekkers, from the Cape, he led his following (which had now increased to between 15,000 and 20,000) across the Limpopo River into what is now southwest Rhodesia.

The flight of Mzilikazi

Across the Limpopo, Mzilikazi found it easy to establish himself, for the once prosperous Rozwi empire and the various Shona paramountcies had been weakened by drought and the wars with the Ngoni. As in the Transvaal, so in Rhodesia, Mzilikazi rapidly absorbed large numbers of the local populace into his age-set regiments; a castelike society evolved, with the original Nguni nucleus at the top, the Sotho in the middle, and the Shona constituting the lowest rung in society. Beyond their immediate settlement, the Ndebele established a variety of relationships with the surrounding peoples, ranging from regular tributary status to occasional raiding and warfare. The first whites to enter the kingdom of Mzilikazi regarded all the Shona as totally subjected to the Ndebele, but in fact, beyond the range of Mzilikazi's raiding armies, many Shona chiefdoms continued in their old independent ways; in the second half of the 19th century, they were building up a stock of firearms to resist the Ndebele.

Yet another group dislodged by the Mfecane made its mark in Central Africa. This was the composite Sotho people known as the Kololo, under Sebetwane, who fled across the Zambezi; there, in about 1840, they conquered the Lozi (Barotse) kingdom, which had been built up in the 18th century and by the beginning of the 19th was one of the most important in present-day Zambia. Weakened, however, by an internal succession dispute and deserted by some of its subjects, the ruling Lozi dynasty was swiftly crushed, and Sebetwane established his own administration over the kingdom by appointing members of his family as headmen in every village. By 1864, however, the ravages of malaria, the accession of a weak and diseased Kololo king, and the revival of Lozi royal fortunes put an end to Kololo hegemony, though Sotho is still the language of the area.

THE SLAVE AND IVORY TRADE NORTH OF THE ZAMBEZI

At roughly the same time that east central Africa saw the incursions of the Ngoni warriors from the south, equally disruptive forces were making themselves felt from the east. Long before the beginning of the 19th century, Central Africa was crisscrossed by trade routes to both east and west coasts. By the second half of the 18th century there was probably a route all the way from Luanda in the west to Mozambique in the east.

Until the middle of the century the Luanda trade was mainly concerned with slaves. With the demand in Europe for ivory and the increasing movement toward abolition of slavery in the early 19th century, however, and with the attempted legal abolition of the slave trade in Angola and the lifting of the royal monopoly on ivory trading in 1836, ivory increasingly became the most important export, and Benguela replaced Luanda as the main outlet. As ivory replaced the slave trade, so the Mbangala of Kasanje and the Portuguese traders on the coast were displaced as intermediaries by the Ovimbundu of the Bié Plateau, whose Chokwe neighbours were already renowned as daring elephant hunters in the 18th century. Now they penetrated deep into Central Africa and with their firearms decimated the elephant population in one area after another. By 1850 they were in the Luvale and Lozi country and were reaching the south Congo forests, where, in the 1880s, they brought Mwata Yamvo's Lunda empire to an end.

The paramount importance of ivory

East of the Lunda the ivory trade was more significant in the 18th century, and so it remained in the 19th. To the south, however, the slave trade—which had gained momentum from the second half of the 18th century, spurred by the labour demands of the plantations of Réunion, Mauritius, Pemba, and Zanzibar, and later of Brazil and Cuba—became as important as ivory. British pressure to abolish slavery in east central Africa had little effect until the last quarter of the century. In Mozambique the slave trade was officially abolished in 1836 and slavery itself legally ended some 20 years later, but this merely led to new ways of extracting forced labour from the African population. Farther north, British pressures on the sultan of Zanzibar to ban the slave trade from Zanzibar to Oman were easily circumvented, and though an abolition treaty forced on the Zanzibaris in 1873 was more effective, the reduced coastal demand for slaves led to even more ruthless methods in the interior, as slaves, no longer needed as merchandise for export, were exploited locally.

The Arabs now began to play a much more dynamic role in the interior. Though initially they were prepared to work through local chiefs, they soon came to exercise a wide military and political jurisdiction over the northern routes from their strategically placed commercial centres, while establishing alliances with African raiders to the south.

The most important area of the slave raiders appears to have been in present-day Malawi and northeastern Zambia, where predatory overlords such as Msiri (leader of the Nyamwezi) and the Arab Tippu Tib devastated a wide area from their bases in the Congo. In addition, to the east of Lake Malawi, the Yao—keen traders since the 17 century—moved from ivory trading to slave raiding, obtaining firearms from the Arabs, subjugating the Nyanja agriculturalists, and building up powerful polities under successful new commercial and military leaders. Similarly, the Bemba, who raided the loosely organized Maravi to the west of the lake from their stockaded villages on the infertile plateau of Zambia, were also able to increase their power through the slave and ivory trade. It was in this area of east central Africa, too, that the Ngoni intruded in about 1850, and though they never became slave raiders, the turmoil in their wake swelled the number of people torn loose from their social setting. Yet the extent of the destruction should not be overestimated; as with the Mfecane, European observers may have exaggerated the degree of depopulation and devastation.

Ngoni intrusions

It is against this changing situation that the arrival of the first European missionaries and traders, inspired by crusades against the slave trade, needs to be seen. In 1875, after an abortive earlier attempt, the Free Church of Scotland established the Livingstonia Mission, while the Church of Scotland started to work among the Yao. From Lake Malawi (Nyasa) the Scottish missions spread inland to northeast Zambia, to be followed by a bewildering array of other denominations in the last two decades of the century. In general, the missionaries soon found that whereas weaker, raided peoples, like the lakeside Tonga, were prepared to accept Christianity, the

Arab, Ngoni, Yao, and Bemba raiders were more resistant; some were even ready to take up arms against the missionaries and European traders who threatened their entire way of life.

In part as a result of pressure from the missionaries and the African Lakes Company (a Glasgow-based firm working closely with them), in 1883 a British consul was appointed for Mozambique and Lake Malawi, with a watching brief on the slave trade. Increasingly, the British government became embroiled in struggles with the Arab traders—which resulted in open warfare in 1887–88—and with the Portuguese. It was not, however, until the appointment of Harry (later Sir Harry) Johnston (an associate of Cecil Rhodes) as consul in 1889 that the British came to play a more aggressive imperial role in the area.

BRITISH OCCUPATION OF THE CAPE

By the beginning of the 19th century, the small outpost of the Dutch East India Company at the Cape had grown into a sprawling multiracial settlement, in which 22,000 whites constituted the top stratum over a labouring class of some 25,000 slaves and an indeterminate number of Khoisan peoples and Xhosa in the eastern districts. In 1806, after a short period (1803–06) of rule by the Batavian Republic had been ended by the renewal of the Napoleonic Wars, the British, anxious to guard their vital sea routes to India, occupied the Cape again.

Most of the early British governors at the Cape were military men who saw the problem of the eastern frontier in military terms: how to defeat and drive back an external enemy. They were tied, moreover, by the reluctance of the British Treasury to spend more than the barest minimum on the colonies. At the same time, neither the African nor the white inhabitants of the frontier zone wanted the complete segregation that the government at the Cape—whether Dutch or British—was eager to impose, and every attempt to establish boundaries between them broke down in the face of their intractability.

The Cape Frontier Wars

In 1811, 1819, 1834–35, and 1846 the almost constant unrest on the frontier exploded into open conflict, and though each conflict pushed the Xhosa closer in on one another and the tensions in Xhosa society increased as the colony was extended farther into the eastern Cape, the problem remained unsolved. Each war, moreover, left the Xhosa more formidable and desperate. They were able to develop guerrilla strategy during the wars, and by 1850 they had also acquired and were able to use effectively a considerable number of firearms; thus, the eighth war, in 1851, was the most drawn-out and expensive of them all. Furthermore, a combined Khoisan and Coloured uprising broke out in the Kat River settlement (established in 1828 to solve the problems of Coloured landlessness), to exacerbate the situation.

In the end it was not so much British arms or settler prowess that defeated the Xhosa as the internal tensions that had built up as a result of the expansion of the colonial economy and ideology. The pressures on the Xhosa increased when the Mfengu refugees from Natal and the interior arrived in the eastern Cape. During the war of 1834–35, the Mfengu fought on the British side and were rewarded for their services with Xhosa lands and cattle —a pattern that was to be repeated in each subsequent war. In 1857 the Xhosa, exhausted by years of attrition and in the face of a new aggressive policy of assimilation inaugurated by the governor of the Cape, Sir George Grey, turned to millenarian prophecies and slaughtered all of their own cattle and destroyed their own crops in the hopes that their ancestors would rise from their graves and drive the white man into the sea. When the awaited salvation failed to materialize, an estimated 20,000 to 30,000 Xhosa streamed westward across the frontier to seek work in the colony, while an equal number died of starvation in the Transkei. Although there was war on the eastern frontier yet again, from 1877 to 1878, the back of Xhosa resistance had been broken. The annexation of the remaining African territories between the Cape Colony and the frontiers of Natal proceeded peacefully, if piecemeal, for the next three decades.

If the Cape eastern frontier saw some of the fiercest and most prolonged resistance to white expansion anywhere in 19th-century Africa, it was also the scene of the closest black and white interaction. Trading relations in the frontier zone existed from the early 18th century, and by the 19th, with the coming of the British settlers, there was a great expansion in trade. The advent of the missionaries at the beginning of the 19th century also had its effect, encouraging the growth of new peasant communities.

Mfengu collaboration

The first converts in the Cape were the Khoisan and mixed groups along the Orange River. In the eastern Cape, Mfengu refugees were soon the focus of mission activity. By 1850 the Mfengu were regarded as the most progressive element in the eastern Cape and were developing into prosperous peasant cultivators. In the 19th century their example seemed to illuminate the advantages of collaboration with the colonial power. After the cattle killing of 1857 and even more after the last wars on the eastern frontier in 1877–78, an increasing number of Xhosa also turned to Christianity and to new political techniques with which to oppose colonialism. By the last quarter of the century, missionaries of almost every denomination were working among the African peoples of southern Africa. Nevertheless, though most of the territories east of the Great Kei conquered in the Cape Frontier Wars remained as reserves in African hands, subject to white overrule, by the end of the century their productivity was already grossly undermined through a combination of ecological and political factors.

THE GREAT TREK

Though the colonial experience came first and in its most concentrated form on the Cape's eastern frontier, by the middle of the 19th century the frontiers of white settlement had been vastly extended. There were few African societies south of the Limpopo, if not of the Zambezi, that were not aware of its implications. This was largely the result of the "Great Trek," or migration, of Afrikaner Voortrekkers (Pioneers) from the Cape Colony in the 1830s, but in part also was a result of British missionary and trading activities in the interior.

The Great Trek is generally regarded as a seminal event in South African history, as in some ways it certainly was, though in the short term it was far less revolutionary than the Mfecane. In part it was simply a continuation of 18th-century trekboer expansion. Blocked from expanding into the high-rainfall area of the eastern Cape, by the 1820s many trekboers were moving northward to the Orange River, where Khoisan resistance had at last been overcome.

Unlike the earlier trekboers, however, the Voortrekkers of the 1830s left the colony with a deep sense of grievance against rulers with very different notions of social order and, in particular, property and labour relationships. Although early British attempts to reform the land law at the Cape were inept, the labour question was even more vexed. Apart from slaves, the basic labour force of the colony was the Khoisan and Coloured (partly of European ancestry) peoples. On the whole, their condition was serflike; they were extremely poorly paid, had little security or freedom of movement, and had no legal redress.

Though the first attempts of the British government to regulate master–servant relations were cautious, Ordinance 50, passed in 1828, was a serious attempt to improve the lot of Cape Coloured servants. In particular, it relieved them of the obligation to carry a pass—a document that previously any servant had had to have signed both by his master and by local officials if he wished to leave his registered place of abode. Ordinance 50 caused a furor among the whites, many of whom were anxious over impending changes in the status of their slaves. (In 1807 trade in slaves had become illegal in British possessions, and in 1833 the remaining slaves at the Cape were emancipated.) Though only 2 percent of slaves were owned in the eastern districts from which the bulk of the Voortrekkers came, Ordinance 50 and slave emancipation aimed to change the basis of labour relations.

For the frontiersmen, however, the last straw came when, after the 1834–35 war on the Cape eastern frontier, an attempt by the governor to annex Xhosa territory was rejected by the British colonial secretary. At the end of 1835, some 4,000 Boers, organized in small groups, left the colony, their movement gathering momentum through the 1830s and slowing in the 1840s. The Mfecane had cleared a great swathe of country for them, though, as the trekkers soon discovered, it had also created a number of powerful and well-organized chiefdoms, such as the Zulu, Ndebele, Sotho (Basuto), Pedi, and Swazi, with whom they would have to come to terms.

The most important of the first parties of Voortrekkers were those under Hendrik Potgieter, Gerrit Maritz, and Piet Retief. In the first instance, all groups made their way across the Orange River, where Potgieter led his supporters toward the Vaal, while Retief and Maritz, eager to secure an outlet to the sea, crossed the Drakensberg to Zululand, hoping to make arrangements with the king, Dingane, who had succeeded his half brother, Shaka, in 1828. Potgieter's party soon clashed with the Ndebele, and by 1837 Mzilikazi had decided to move northward across the Limpopo, leaving the Voortrekkers the dominant power on the western Highveld.

The Natal Voortrekkers were less immediately successful. In 1838 Dingane wiped out a negotiating body under Retief before falling on the encampments of the trekkers, who had followed him and had scattered over northwestern Natal. Dingane's triumph was short-lived. In the second half of 1838, reinforcements arrived from the Cape, and the Voortrekkers, forming an alliance with his brother Mpande, deposed Dingane (1840) and established their rule in Natal, south of the Tugela River.

The Voortrekkers soon found, however, that initial conquest was easier than overcoming their own divisions or establishing their rule permanently over large numbers of African people—especially as they had to contend also with the opposition of the British government. In 1838 a small British force occupied Natal, and five years later the British annexed it in a bid to protect their sea route to India and to prevent pressure on their eastern frontier through further upheavals among the Africans in Natal, which threatened to result from trekker land and labour policies. Even before this, the Afrikaner republic threatened to break down, as thousands of refugees streamed back to Natal from Zululand, once the hold of the despotic Shaka and Dingane had been broken.

British annexation of Natal

Most of the Voortrekkers now recrossed the Drakensberg to join their compatriots on the Highveld. In the 1820s and '30s the British relied for law and order in the interior on a series of treaties with African rulers such as Mshweshwe of the Sotho and the Griqua leaders Andries Waterboer and Adam Kok. With the dispersion of considerable numbers of whites in the interior, however, these treaties inevitably broke down. In 1848 the governor of the Cape annexed the territory across the Orange River in an attempt to control the colony's turbulent frontiers; but mid-Victorian Britain had little real desire to follow the trekkers into the vast interior. In 1852 and 1854 independence was granted to the trekkers across the Vaal and Orange rivers, respectively. At the same time, the alliances with African chiefs, except for that with Adam Kok, were unilaterally disclaimed. Trade in firearms was allowed to white Afrikaners but was forbidden to Africans—a factor that in the future effectively weakened the Africans' resistance against the Afrikaners.

THE TRANSVAALERS AND THEIR NEIGHBOURS

With the abdication of British authority over the interior, the trekkers had first to overcome their own divisive tendencies and then to come to terms with the surrounding African peoples. In the Transvaal in mid-century there were about 20,000 whites—a figure which had risen to about 30,000 by 1870—grouped in four main centres, each the headquarters of an independent "republic." Between these centres the remaining Boers were scattered in farms of from 6,000 to 9,000 acres. Between 1845 and 1864 political life in the ramshackle republics was racked by bitter personal, religious, and political feuds. Even after 1864, when the different centres merged to form the South African Republic, central authority was weak and there was little sense of national unity.

Like the Africans, the trekkers were fundamentally cattle farmers in search of land and water, though they had very different concepts of land ownership and usage. When they moved into the Transvaal, they claimed all the lands and peoples the Ndebele had ruled and dealt as overlords with the peoples living in the lands that the trekkers apportioned to themselves.

As the land between the Vaal and the Limpopo was appropriated—either by the Boers or by the British and colonial land speculators who followed—African tribes were often able to live under their chiefs who had, however, now to provide labour for the European farmers. Itinerant traders, mainly British, served the needs of both communities, bringing the essential firearms as well as such luxuries as tea, coffee, and sugar, in exchange for ivory, cattle, and a little wool. From the early 1860s considerable numbers of Africans migrated to the Cape Colony to earn money and obtain firearms.

Administration of the Transvaal

With their near-subsistence economy, the Boers' demands for labour could not have been great. Nevertheless, in addition to the labour tax, they had to use force to get sufficient labour. Each raid against an African chiefdom left numbers of so-called orphans, who were "apprenticed" to individual farmers. Virtual slaves, they constituted an important part of the labour force. This apart, Africans suffered not so much from systematic exploitation as from the sporadic plunder and occasional terrorism that served to remind tribesmen of the white man's firepower.

Encircling white settlements in the Transvaal to the east, northeast, and north were a number of relatively powerful chiefdoms with which the Boers made alliances, both in pursuit of their own internal politics and because of their lack of power. To the west, they were faced with the Tswana chiefdoms, regrouping after the Mfecane and powerfully supported by missionaries. Whatever the relationships with their labourers, the Voortrekkers in their dealings with African chiefs were circumspect and also inferior in position. As late as 1877 the Boers were threatened by the Zulu and unable to defeat the Pedi. The conquest of all these peoples (except the Venda) in the last quarter of the 19th century was left to the British with their superior armaments.

THE ORANGE FREE STATE AND LESOTHO

When the British withdrew from Transorangia in 1854, it too had a relatively small white population owning vast quantities of land in the midst of a large African population. By 1875 this white population had risen to about 26,000. Though there were only about 97,000 Africans in the Orange Free State at this time, on its eastern frontier was the well-organized state of Basutoland under Mshweshwe. Unlike the Transvaalers, the Free State tackled the problems of government effectively, establishing a Volksraad, or parliament, and drawing up a clearly defined constitution. With both the Cape government and Mshweshwe on their doorstep, the Orange Free State could not indulge in excessive internal strife.

Basutoland and the Free State

Even before the British withdrawal from Transorangia in 1854, there had been constant friction between the Sotho and settlers over land. As refugees streamed back to their old homes after the Mfecane, Mshweshwe's following increased to an estimated 80,000 in 1848. By the time they returned, however, they found that the Boers had begun to infiltrate into the fertile lands between the Caledon and Orange rivers. The Sotho, overcrowded, now spilled over onto Boer farms as well as onto lands occupied by other African groups, and conflict arose.

Despite Mshweshwe's attempts to keep the peace, cattle raiding from those of his subjects who saw no hope of recovering their land by negotiation, together with increased demands for land by white sheep farmers, led to war in 1858 and again in 1865. On the first occasion the Orange Free State was forced to sue for peace. On the second, however, Basutoland, starved of arms and torn by internal jealousies, was beaten. Some chiefs, es-

pecially in the north, offered their allegiance to the Boers and, with their followers, became labour tenants on white farms; others fled into the Transkei. In 1868, in response to repeated appeals from the Sotho, the governor of the Cape intervened to declare the Sotho British subjects and their land British territory—much to the chagrin of the Free Staters. Nevertheless, the Orange Free State gained control of the fertile land in the Caledon Valley and gained in security.

In 1869 the frontiers of Basutoland were delimited, and shortly thereafter it was handed over to the Cape. In 1881, however, when the Cape government tried to disarm the Sotho, war broke out, and the Cape returned the territory to Britain in 1884.

THE BRITISH COLONIES OF THE CAPE AND NATAL

The two British colonies of the Cape and Natal, though their social organizations differed considerably, were far more closely tied to the 19th-century world economy than were the interior republics.

Natal. White settlers in Natal were outnumbered by blacks in a ratio of 10 to 1 and surrounded on all sides by powerful African kingdoms. At the same time, their interests were divided, as land speculators, plantation owners, and small farmers vied for control over Natal's native policy. For much of the 19th century, black–white relations were dominated by Theophilus Shepstone, diplomatic agent to the African tribes and secretary for native affairs from 1846 to 1875.

Achievement of Shepstone

Shepstone's first task was to control the enormous reflux of Africans to their homelands after Dingane's defeat. It is estimated that the numbers of Africans south of the Tugela increased from about 20,000 in 1840 to 200,000 in 1845. Between 1846 and 1864, eight reserves, totalling 2,250,000 acres (900,000 hectares), were set aside for African occupation. Shepstone reconstituted the chiefdoms broken by the Mfecane and placed them on the reserves, which he governed through the hereditary chiefs. Customary law was recognized, and the governor held the position of supreme chief. Though provision was made for educated and Christian Africans to be exempted from customary law and to acquire the franchise, it was so complicated that by the turn of the century only 1,000 or so Africans had certificates of exemption.

Shepstone's achievement was to keep the peace in Natal for some 30 years. The only violent resistance to his rule came in 1873, when Chief Langalibalele of the Hlubi people, refusing to obey summonses from Shepstone, fled with his followers over the Drakensberg to Basutoland. The ruthlessness with which he was pursued and the summary way in which he was brought to trial roused a humanitarian outcry in Great Britain and led in part to Britain's paring down Natal's limited representative institutions. Natal achieved full self-government only in 1893.

To some extent Africans in Natal were prepared to accept white rule because of their memories of the Mfecane. To the north the Zulu state, though restricted, was still formidable, and in the middle years of the century the small white population of Natal was careful to maintain good relations with its kings. They, in turn, were eager to use the British connection as a counterweight to the encroachments of Transvaal farmers.

While they had adequate land, Africans saw little reason to work for the white man. Initially, too, land speculators were prepared to settle for African rents rather than labour. As the demands of the smaller farmers became more powerful, however, a hut tax of seven shillings a year was imposed; it was doubled in 1875. Initially the Africans were able to pay the tax by disposing of their surplus agricultural products on the new European market; but as the reserves became more crowded, increasing numbers of Africans were forced to work for the white man. This was too slow for the white sugar planters, and in 1860 the first indentured Indian labourers arrived in Natal, soon to be followed by Indian, mainly Muslim, traders.

Despite the undoubted prosperity Indian labour brought to the colonists, the Indians were bitterly discriminated against and by the end of the century had few social or political rights. On the expiration of their indentures, labourers were subject to a £3 poll tax, and every endeavour was made to repatriate them. By the 1890s they outnumbered the white population of Natal and had spread to the South African Republic, where they were equally discriminated against. Largely in response to this situation, Mahatma Gandhi founded the first Indian political organizations in the Transvaal and Natal.

Cape Colony. The Cape, as the oldest and most populous of the colonies in South Africa, had the most complex society. Its economy was based on agriculture and mercantile capital with the labour force drawn from the Coloured population with an increasing number of Bantu speakers as the Ciskeian and Transkeian territories became less and less productive.

Representative institutions in the Cape

In 1853 the colony was granted representative institutions. Elections were based on a colour-blind franchise, low enough to enfranchise sufficient Coloured voters to support British interests and counteract the enfranchisement of the Afrikaner population. As African territories were progressively incorporated, the franchise was also extended to qualified Africans. In 1872 the Cape received responsible government and the colour-blind franchise was retained, though in 1887 and 1892 the qualifications were raised in an effort to restrict the number of African voters. Even the Afrikaner Bond, the colony's first and most important political party, which consistently attacked the franchise, came to accept its basic colour-blind principle. Predicated on a loose affinity of interest between Cape merchants and African peasants, Cape liberalism proved fragile in the face of South Africa's industrial revolution.

THE ERA OF MINERAL DISCOVERIES AND CONFEDERATION

The pressures of Africans to seek work in the white economy were vastly accelerated by the mineral discoveries of the last third of the century. In 1868 diamonds were found by some children playing along the Orange River; and in 1886 the even more important gold reefs of the Witwatersrand (Rand) were discovered.

Until 1868 the South African economy had been poor, backward, and pastoral, with wool the most important export. There was little to attract the great stream of European immigration that was making its way to Australia, New Zealand, and the Americas. In 1820 some 5,000 British emigrants had been settled in the eastern Cape, in order to defend the frontier, and in the 1840s and '50s there were numerous schemes of immigration to Natal. Between 1858 and 1862 more settlers were brought to the eastern Cape, but many left during the prolonged economic depression of the 1860s.

With the diamond discoveries, immigration no longer needed official encouragement. As thousands of fortune seekers made their way to the diamond fields, a fierce ownership dispute raged between the Voortrekker republics, the Griqua (mixed European and Khoisan people) on the Orange River, and the Tswana; ultimately the entire diamondiferous area, known as Griqualand West, came under British, and then Cape, control. Afrikaners of the Orange Free State were thereby incensed, even though the British lieutenant governor of Griqualand West had difficulty in controlling the near-anarchic mining town of Kimberley, with its unruly white miners and its thousands of migrant black labourers.

As events in Griqualand West threatened the equilibrium of the whole region, it appeared to the colonial authorities that some unification or federation of the divided South African territories was essential. Their timing was so bad, however, and their handling of the situation so tactless that in the long run it served only to kindle pan-Afrikaner nationalist feeling.

British annexation of the Transvaal

In an attempt to push through the confederation of the South African territories, Britain annexed the Transvaal in 1877. The pretext was the inability of the bankrupt republic to subjugate the Pedi in the Zoutpansberg. Under their chiefs Sekwati (died 1861) and Sekhukhune (ruled 1861–81) the Pedi had successfully withstood the onslaughts of the Mfecane, had attracted many of the surrounding Africans into their orbit, and had built up a

state powerful enough to resist white demands for taxation and labour services. This war was one of the cluster with similar origins that broke out all over South Africa in the late 1870s.

At the time of the Transvaal–Pedi war, it seemed as if conflict were imminent also on the Transvaal–Zulu frontier. There Cetshwayo, who had succeeded his father, Mpande, in 1872, was disputing Transvaal encroachment on Zulu lands. In 1879 a war broke out that was very largely engineered by the newly appointed British high commissioner for South Africa, Sir Henry Bartle Frere, who believed the Zulu to be a threat to the security of whites in Natal and the Transvaal and a stumbling block in the way of confederation. Despite early disasters and tremendous logistic problems on the imperial side, Cetshwayo was captured after the Battle of Ulundi in July 1879, deposed, and deported.

Though the spirit of confederation was completely discredited by the aggressive and expansionist policies it had engendered, and Zululand was not annexed, the divisive settlement after the war led to bitter civil war in Zululand and to further encroachments by the Boers. Britain proclaimed a crown colony in Zululand in 1887, and in 1897 Zululand was finally handed over to Natal.

Ironically enough, with the recession of the danger to the Transvaal from Zululand, which had provided one of the excuses for British intervention, the Transvaalers were able to rise against an increasingly unpopular administration and defeated the British resoundingly at Majuba in 1881. In 1881 the Convention of Pretoria stipulated the continuance of an ill-defined British "suzerainty" over the Transvaal, while handing over internal self-government; though major concessions were made to the Boers in the Convention of London in 1884, the suzerainty clause was left ambiguous.

Battle of Majuba

THE SCRAMBLE FOR SOUTHERN AFRICA

Over the next 15 years the ambiguity of British suzerainty in the interior proved a constant source of friction between the South African Republic and Britain, which formed a pretext for the South African War of 1899–1902. Though in the early 1880s the Liberal government in Britain seemed to favour withdrawal from the interior, by the middle of the decade a new phase of intervention began, spurred on by the European scramble for African territories, a battle over borderlands caused by the continued drift into African territory of land-hungry Boers, the grandiose schemes of Cecil Rhodes, and the discovery of gold in the Transvaal.

The response of Africans to these events ranged from outright resistance to some form of accommodation, depending on their previous experience of contact with the white man, on their observation of what had happened to neighbours already subjected to colonial power, on the advice they received (especially from missionaries), on their relationships with neighbouring African chiefs, and on calculations of their internal political interests. As the first small numbers of concession hunters appeared, many chiefs tried to play the British, Boer, German, or Portuguese intruders off against one another. Though a few chiefs even emerged from the scramble with enlarged domains, most Africans in the south were to lose out to white-settler and capitalist expansion.

German intervention. The new assertion of British claims and prior rights in southern and Central Africa was caused by the sudden interest of Germany in the area. In 1883 the Germans landed at Lüderitz Bay, and in 1884 they annexed the whole of southwestern Africa, much to the dismay of colonists in the Cape.

German interests in southern Africa were not confined to the arid coastal lands of the southwest. German movements along the coast farther east led in 1885 to the declaration of a British protectorate over Pondoland (which became absorbed into Cape Colony 10 years later) and to the establishment of a crown colony in Zululand in 1887. Germany's apparent interest in the hinterland of South West Africa and, thus, in a possible linkup with the South African Republic probably gave the necessary urgency to the impassioned pleas for aid that Tswana chiefs, their missionary advisers, and Cape traders and politicians were directing to the British Colonial Office against the incursions into their territory of Transvaal Boers. When, in 1882, Boer adventurers, initially invited in to aid rival chiefs, had established the two small republics of Stellaland and Goshen, they not only undermined the African polities but also threatened to disrupt the missionaries' and traders' "Road to the North." Moreover, by this time, rumours of rich mineral wealth beyond the Limpopo had begun to interest Cecil John Rhodes, the man who was to dominate imperial expansion in southern Africa.

Rhodes and the era of expansion. The story of how Rhodes went to South Africa to repair his frail health and stayed to become a millionaire on the diamond fields before he was 30 is legendary (see RHODES, CECIL). At Kimberley, chaotic underground conditions—the result of the uncontrolled enterprise of individual miners and the unrestricted flow of diamonds to the world market—had, by the 1880s, made some form of centralized planning or amalgamation necessary. By the late '80s Rhodes had bought up all his rivals and had created the vast consortium of De Beers. Having secured the "Suez Canal" to the north, as he called it, by the British action in Bechuanaland, Rhodes next planned to take control of the area lying to the north of the Limpopo, to acquire its rumoured mineral wealth and to encircle the Transvaal.

De Beers consortium

His schemes were given added potency by the discovery in 1886 of the deep seams of gold of the Witwatersrand; this had led to a vast strengthening of the South African Republic and to an apparent threat to British dominance in the region, especially since at this time the president of the Transvaal, Paul Kruger (*q.v.*), was bent on expanding both eastward through Swaziland to the sea and northward, across the Limpopo, where one of his agents had secured an alliance with the Ndebele in 1887. Both of these projects were foiled by the British government, though in 1894 Britain granted the Transvaal sovereignty over Swaziland, despite the bitter protests of its rulers.

At the beginning of 1888, Lobengula, king of the Ndebele, surrounded by concession seekers and swayed by the apparently official British support for Rhodes, granted the magnate's representatives complete and exclusive mining rights in his territory, Matabeleland. Lobengula was soon to regret his decision, but by then it was too late: in 1889 Rhodes immeasurably strengthened his hand by forming his British South Africa Company, which was authorized by the British government to extend administrative control over a vast area of southern and Central Africa.

In 1890 Rhodes sent a 200-strong "Pioneer" Column into Ndebele raiding grounds in Mashonaland. The resentment of the Ndebele mounted as they saw land parcelled out to the first Pioneers; the King's power was further undermined by European demands for Shona labour. In 1893 a deliberately provoked war between the two broke out. The kingdom was destroyed to solve the financial problems of the British South Africa Company, open Matabeleland's reputed gold mines for exploitation, and secure the region for white settlement. Ndebele cattle were confiscated and land was alienated.

The Ndebele War

In neighbouring Barotseland, Lewanika, head of a complex centralized state, had deliberately invited in missionaries and then requested British protection. He had not, however, reckoned on the British South Africa Company, whose agents misconstrued the nature of their connection with the British government and made wide promises of technical and financial assistance in order to secure major mining concessions. The promises remained largely unfulfilled; and though Lewanika undoubtedly welcomed the defeat of the Ndebele, he must have watched events to the south with trepidation. Though Lewanika took care to concede mineral and, later, land rights in the territories of his subject peoples, and the Lozi were able to expand territorially by playing the company off against Portuguese claims on their western and northeastern borders, ultimately Lewanika's fears were warranted. Although subject to the same pressures from whites as



Colonial South Africa 1884–1905.
From J. Fage, *An Atlas of African History;* Edward Arnold (Publishers) Ltd.

the Ndebele, the Lozi chose accommodation and manoeuvre rather than resistance. Nevertheless, by the end of company rule in Barotseland, the King's position had been seriously undermined, and his realm was as underdeveloped as that of his neighbours, despite a certain favoured legal position, which fostered a local Lozi nationalism even after independence had been granted to Zambia in 1964.

In northeastern Zambia and Malawi, too, Rhodes's company had been busy through the mediation of the British consul, Harry Johnston, who had accepted from Rhodes £2,000 toward the administration of east central Africa and the commission to obtain treaties from chiefs as far as copper-rich Katanga. Although Rhodes was beaten to Katanga by Leopold II of Belgium, treaties were obtained from a number of chiefs over a wide area between Lakes Malawi and Tanganyika.

In 1889–90 a British protectorate was declared over the Shire Highlands against Portuguese claims, and the following year this was extended to the area around Lake Malawi. British administration was gradually extended into what is now northeastern Zambia over the next five to six years. In the area that became known successively as the British Central Africa Protectorate (1893–1907), Nyasaland (1907–63), and Malawi, Johnston became embroiled in campaigns against the most prominent chiefs, many of whom feared the loss of their livelihood as slavery was stamped out and as Johnston encouraged British and Indian immigration. Here Yao resistance was the most prolonged, though it never threatened British control.

In northeast Zambia, where British settlers and concessionaires had begun to infiltrate and conflict with chiefs by 1896, Mpezeni's Ngoni put up the fiercest resistance but were forced to submit to British rule in 1898. Kazembe, king of the eastern Lunda, faced with a punitive force, fled from his capital, while the Bemba, despite their warlike reputation, remained discreetly quiescent.

The Jameson Raid

While Johnston was engaged in these punitive expeditions in the northeast, however, Rhodes had far more serious problems on his hands south of the Zambezi. By 1895, despite the progress of his plans all over southern Africa, Rhodes's hopes of rich gold deposits in Matabeleland were disappointed, and his deep-level mines on the Witwatersrand were an extremely heavy drain on capital —a situation that was exacerbated by the policies of the Transvaal regime. Making use of the largely fomented clamour of British immigrants in the Transvaal, known as Uitlanders (aliens; *i.e.*, non-Afrikaners), over their lack of voting rights, Rhodes, with the full knowledge of the British colonial secretary, Joseph Chamberlain, plotted the invasion of the republic by his lieutenant, Leander Starr Jameson, and a number of armed followers.

The affair was a complete fiasco. The hoped-for uprising of Uitlanders failed to materialize, and Jameson was soon arrested by the Boers. Nevertheless, the Jameson Raid had profound repercussions, both in South Africa and to

its north, where the Ndebele, still smarting from their harsh treatment after the 1893 war, took the opportunity of Jameson's absence to rise against their white rulers. Within a week of the raid, the Ndebele had killed more than 100 whites, and they were soon joined by the Shona in what was probably the most devastating resistance to colonialism in central Africa in this period. British troops had to be called in, and as a result greater imperial control came to be exercised over the activities of the company, both in Southern Rhodesia, as it now came to be called, and north of the Zambezi.

The South African War. Though Rhodes retired from active politics in the Cape after the Jameson Raid, the series of events he had set in train gathered momentum in the next five years. Whereas in the early 1890s Rhodes successfully united English and Afrikaners of the Cape in his imperial schemes, now many Cape Afrikaners, formerly antagonized by the restrictive policies of the Kruger regime in the Transvaal, united behind the republics in a Pan-Afrikaner nationalist opposition to British imperialism, while most of the English-speaking community, both in the Cape and elsewhere in South Africa, became infected with the spirit of jingoism that marked the last years of the century. The dispatch of Sir Alfred Milner as British high commissioner to South Africa exacerbated the situation; and in 1899 the war, which Milner had regarded as inevitable from the outset, broke out.

Growth of the Pan-Afrikaner opposition

The clash between Kruger and the representatives of British imperialism in southern Africa—Rhodes and Milner—has often been portrayed in personal terms. Yet, essentially, the clashes of these years were economic and political. The South African (or Anglo-Boer) War was a struggle for supremacy over an area that contained the richest gold mines in the world; and though it has become fashionable to discount economic explanations for the war and to stress the strategic calculations of the British, the nature of the state controlling gold production and the new markets of the southern African interior was a crucial British interest. It gave rise to the most bitter and closely fought colonial war in Africa, which left its mark on South African politics for more than half a century.

ANGOLA AND MOZAMBIQUE

The European bid for African territories in the last quarter of the 19th century was accompanied by a reawakening of Portuguese colonial ambitions. Throughout much of the 19th century the Portuguese in Angola and Mozambique had been confined to a few towns and forts on the coast, where Portuguese traders eked out a petty existence.

Tensions among the Portuguese government in Lisbon, local officials, white settlers, *assimilados* (as the assimilated Africans were called), and the Africans in the interior existed in the 19th century, as they had in earlier times, with conflict flaring up when local groups found their interests ignored by the metropolitan power or in moments of Portuguese vulnerability. Thus, revolts in Luanda and Benguela coincided with the declaration of Brazilian independence, while Portuguese attempts to abolish the slave trade in the 1830s led to widespread resistance by local whites and African intermediaries. Between 1835 and 1860, the energetic but largely abortive policies inspired by the liberal Portuguese statesman the Marqués da Sá de Bandeira provided fresh occasion for local uprisings; and his attempts to extend Portuguese control in the interior led to a series of wars in Angola with the Ndembu (Dembo), Kongo (Bakongo), Mbangala (Imbangala), and Mbundu (Ovimbundu) peoples, who, nonetheless, were able to remain independent until the 20th century.

In Mozambique, the Portuguese experienced even less success. Despite official abolition of the slave trade and of slavery by 1858, the practice still continued until at least the 1880s. Several attempts, including major military campaigns, to outlaw the *prazo* system in Zambezia were ineffectual; and even, in 1880, the declaration that they were crown property brought little change to their essentially feudal nature. Until 1895 the Gaza kingdom controlled a wide area between the Komati and Limpopo rivers, while the Portuguese had little control over the Africans in the rest of the territory.

The only bright spot in Portuguese fortunes in southeastern Africa was the growing prosperity of Delagoa Bay as trade with the Boers in the Transvaal increased. In 1875 Portuguese rights to Delagoa Bay were formally recognized, and in the 1880s, with the discovery of gold on the Witwatersrand, Delagoa Bay acquired a new importance as the closest outlet for the rapidly growing metropolis of Johannesburg. In 1888 Delagoa Bay became capital of Mozambique, and the focus of Portuguese colonization in eastern Africa shifted southward.

The rise in importance of Delagoa Bay

Despite the failures of the expansionist plans in the mid-century, the continued criticisms of foreign observers and the activities of explorers brought to Portuguese statesmen a new awareness of the need to repair their fortunes in Africa actively—especially when the late 19th-century scramble for African territories threatened even those areas the Portuguese considered theirs by historical right. The dream of linking Angola and Mozambique was revived, although with the weakness of Portugal in Europe, with its slaving record and its monopolistic trade policies, it received short shrift from the other imperial powers. At the Berlin West Africa Conference of 1884–85, Portugal received the Cabinda exclave and a portion of the left bank of the Congo—considerably less than it had claimed. In the hinterland of northwestern Mozambique it fared even worse: in 1890 Portugal was pushed out of the Shire Highlands, and the expansion of the British South Africa Company in Southern Rhodesia led to further retractions in western Mozambique. Skirmishes with the British South Africa Company in Rhodesia and Barotseland continued through the 1890s, and competition with the Germans on the southern frontier of Angola and the northern frontier of Mozambique dragged on into the 20th century. Though Portugal's more ambitious plans were doomed by the end of the century, it had nevertheless acquired about 800,000 square miles (2,000,000 square kilometres) of African territory, of which it controlled less than one-tenth.

Portuguese plans to develop the colonies hinged on attracting immigration and foreign capital to provide the necessary economic infrastructure. Even before the scramble proper, the arrival of 300 Afrikaner pastoralists from the northern Cape was a welcome addition to the white population on the Huíla Plateau in southern Angola, where they helped to subdue the Ovimbundu. In Mozambique, the *prazos* were adapted to the needs of three foreign chartered companies, which were granted wide mining, agricultural, and commercial concessions. For the Africans on the estates, however, the system of forced labour, taxation, and tribute continued to impoverish the local economy.

As Portugal began to extend its control and policies in the interior of both Angola and Mozambique, it was faced with fresh waves of resistance. In Mozambique in 1894 an uprising by peoples around Lourenço Marques was forcibly suppressed; and in 1895 the Portuguese opened the attack on Gungunhana's already weakened Gaza kingdom. With the conquest of the kingdom, which took three years, Portugal removed the most formidable threat to her position in southern Mozambique. Nevertheless, from 1897 to 1903 the Shona peoples of Barwe also rose against Portuguese rule, and the revolt spread among Shona speakers in Rhodesia in a continuation of the 1896–97 rebellion. As late as 1917, in response to the Portuguese demands for carriers in the war, a major Shona rebellion broke out in the Zambezi Valley, which totally disrupted the Portuguese administration for a period of several months.

Resistance to the Portuguese

Resistance to taxation, forced labour and crop production, unfair trading and administrative practices, and military recruitment lay behind African opposition to Portuguese rule in Angola also. The Ovimbundu, who were extremely successful in the wild rubber trade from eastern Angola between 1870 and 1911, were conquered in 1902–03, the Ndembu between 1907 and 1909, and

the Kongo in 1913–14. Portuguese attempts to control the Kwanyama, an Ovambo (Ambo) group who lived on the Kunene River, led to further resistance, made the more formidable by German intervention and Kwanyama stockpiling of trade guns. Conquered in 1915, the last Kwanyama king, Mandume, fled into South West Africa, where he was killed in 1917.

Though sporadic outbreaks of resistance to Portuguese rule continued till 1930, Portugal's military control of its colonies was greatly strengthened after World War I as the result of foreign, largely British, economic investment. This brought little benefit to the African population, and for many the only escape from the combination of oppressive rule and absence of local economic opportunity was to migrate temporarily or permanently to neighbouring territories.

THE GERMANS IN SOUTH WEST AFRICA

The Germans were the last imperial power to arrive in Africa; they annexed South West Africa in 1884. At that time the region was inhabited by groups of San (Bushman) and Bergdama (Damara) hunter-gatherers, as well as by the purely pastoral Nama and Herero in the south and centre and by Ovambo (Ambo) agriculturalists in the north.

By the third decade of the 19th century, the sharpening competition between the Nama and Herero over the grazing lands of central South West Africa led the leading Nama group to ask for aid from Jonker Afrikaaner, the head of a well-armed, mixed Khoikhoi (Hottentot) and half-caste chiefdom on the Orange River. The Herero were subjugated by Jonker and the Nama between 1835 and 1860. Jonker Afrikaaner established himself at Windhoek, where he controlled the trade from both Walvis Bay and the Cape.

Throughout the 19th century, displaced Khoikhoi and half-Khoikhoi communities from the Cape Colony, known as Oorlams, had entered South West Africa, adding to the competition for water and grazing lands. The last and most important of these were the followers of Hendrik Witbooi, who, arriving in the 1850s, soon dominated the Nama and the so-called Baster community who settled around present-day Rehoboth in the 1870s.

After Jonker Afrikaaner's death in 1861, the Herero shook off the Nama yoke under the leadership of their new chief, Maharero. As both sides acquired increasing numbers of guns, the conflict between them became more destructive and, apart from a breathing space between 1870 and 1880, continued from 1863 to 1892, opening the way to German annexation in 1884.

Initially, the Germans secured their hold through treaties with the Herero and Nama chiefs, though that with Witbooi was extracted only by force. There was further fighting when the Germans extended their landholdings beyond the environs of Windhoek into the grazing grounds of eastern Herero and Mbandieru people and confiscated their cattle.

Anti-German rebellions

With the growth of the white-settler population, the demands for African land, labour, and cattle, the contemptuous attitudes of whites toward blacks, unfair German trading practices, and the absence of legal redress led to a series of revolts between 1903 and 1907, which began and ended with the Bondelswarts (a Coloured group) in the south of the country but which also involved the Herero and Witbooi's followers. The rebellions were put down at tremendous cost to the African population. These wars, among the most destructive waged against Africans in colonial times, opened up South West Africa to German settlement and exploitation, though in the event their dominion only lasted until World War I, when the territory was conquered by South African troops. After the war, South West Africa was handed over to the Union of South Africa (now the Republic of South Africa) as a mandate of the League of Nations.

III. Southern Africa, 1900–45

For black and white southern Africa, the 20th century has wrought vast changes, though north of the Zambezi and in the Portuguese territories, where colonial development was hampered by the two world wars and a depression, they were initially less dramatic. Nevertheless, by 1945 in most areas Western administrative, medical, and educational practices, however inadequate, were established, roads and railways built, minerals exploited, and Africans drawn into the world economy more intensively than ever before. New frontiers had been created, which have become the boundaries of modern states; and everywhere the demands of new rulers were undermining traditional authority and small-scale precolonial societies and creating new class and political alignments.

Farther south, where this process had first begun, whites increased their political, economic, and military hold through the unification of the South African colonies, creating, by 1945, the most highly industrialized state in Africa. The Union (later Republic) of South Africa has dominated its neighbours throughout the 20th century. Capital expansion northward was initiated by the south and was paralleled by a vast labour migration to the mines and farms of South Africa. At the same time, class and colour relations shaped by the early days on the diamond and gold fields and given characteristic form by the first decades of the century in South Africa were repeated both in the development of secondary industries in the union and in the mines and the industries of the Rhodesias.

In all the territories, the economic exploitation of blacks was facilitated by the development of a racist ideology. Africans were seen essentially as a source of cheap unskilled labour, not, as elsewhere in colonial Africa, as cash-crop producers. Discriminatory policies, whether described as segregation or parallel development, were prompted by settlers' fears of African political and economic competition and lent respectability by anthropologists and paternalist administrators anxious to preserve African societies from rapid social change. Though in the Portuguese colonies "native policy" was theoretically to assimilate Africans to a Portuguese way of life, obstacles to the *assimilado* progress were in practice greater in the 20th century than they had been in the 19th; and the *indigenato* policy, in force from 1926 to 1961, while granting equal rights to a very small handful of assimilated Africans, established separate administration, lawcourts, institutions, and status for the vast majority.

WHITE SETTLERS SOUTH OF THE LIMPOPO

If the South African War was the high-water mark of direct British intervention in South Africa, the tide rapidly receded in the next five years. In 1906–07 the British granted self-government to the defeated republics, and by 1910 the four self-governing colonies had been unified. This unification was dictated largely by the demands of the industrial revolution consequent on the mineral discoveries. With the South Africa Act, the 1,250,000 enfranchised whites achieved complete political control over approximately 4,250,000 unrepresented Africans, 500,000 Coloureds, and 165,000 Indians.

The unification of South Africa

Though there was talk of incorporating the landlocked British protectorates, or High Commission Territories, of Basutoland (now Lesotho), Bechuanaland (now Botswana), and Swaziland, the opposition of the Africans of the protectorates, as well as that of the radical lobby in Britain, led to their exclusion from the union. Until the mid-20th century, however, it was thought in Britain that the protectorates would ultimately be incorporated, and on a number of occasions—especially during the 1930s—South African politicians demanded this. Strong African opposition, the effective lobbying of radical groups in Britain, and, finally, the outbreak of World War II prevented this from happening; nevertheless, the assumption meant that the territories were fully locked into South Africa's migrant labour system and became increasingly impoverished. Administrators were often South African; and local chiefs, subject to a virtually unassailable paramount chief, were left to run their tribes as they saw fit. Though this dual administration, as well as the stagnant economy of the territories, was severely casti-

gated by the so-called Pim Commission of 1934–35, and moderate reforms were instituted, the High Commission Territories remained backward and neglected.

Though South Africa failed to incorporate the High Commission Territories, after World War I it acquired the vast German colony of South West Africa. Despite a League of Nations mandate that the territory be administered as a "sacred trust," white settlement was encouraged and subsidized from the outset. Though almost half of the German population was repatriated at the end of the war and the rest were naturalized in 1924, Nazi demands in the 1930s for the return of German colonial possessions had a counterpart in South West Africa; by that time there were about 30,000 whites in the territory, largely from South Africa, and considerable vested interests had grown up, especially in diamond mining and fishing. There were almost annual criticisms from the Permanent Mandates Commission of the League about the lack of economic and educational development for Africans, and the League was much disquieted by the violent confrontations between the South African government and the Bondelswarts (1922), the Rehoboth community (1924–25), and the Ovambo (1932). Nonetheless, the major preoccupation of the mandatory power continued to be to foster white prosperity, generally at the expense of African tribesmen.

White politics in the union

Within the union, between 1910 and 1948, whites were divided by class and ethnicity. Until 1924 the South African Party (SAP) was in power under the leadership of two successive prime ministers, Louis Botha (1910–19) and Jan Smuts (*q.v.*; 1919–24), representative of the more successful Afrikaner farmers and in close alliance from 1914 with the Unionists, the party of mining capital, which the SAP absorbed in 1920. The SAP's essential aims were to reconcile the English- and Afrikaans-speaking communities, to maintain the imperial connection, and to promote commercial farming and mining interests. This led to an increasingly bitter conflict with the white working class on the Rand as well as with the Afrikaner nationalists, who broke away from the SAP in 1913–14 under the leadership of Gen. J.B.M. Hertzog.

In 1924 a coalition of Hertzog's Nationalists and the predominantly British Labour Party removed Smuts from power. Between 1924 and 1929 legislation protected white workers from African competition, while the state sponsorship of industrial and agricultural development was accelerated in an attempt to help solve the "poor white" problem. Largely through Hertzog's efforts, the nature of the British Commonwealth connection was redefined in 1929–30, and the concept of dominion independence was fully enunciated.

In 1933, in the face of South Africa's worsening economic position during the Great Depression and the crisis of confidence surrounding its abandonment of the gold standard, the Afrikaner generals Smuts and Hertzog formed a coalition government, and in 1934 their parties fused as the United Party. The outbreak of World War II destroyed their somewhat precarious consensus. As in World War I, when fierce civil war erupted between the Afrikaners on the issue of South Africa's participation on the British side, Afrikaners again were divided over the country's contribution to the Allied forces. When South Africa voted by a narrow margin to fight the Germans, Hertzog resigned.

Purified Nationalist Party

Even before this, in 1934, a group of nationalists—largely members of the extreme Broederbond, a secret society that aimed at Afrikaner economic advance and ultimately Afrikaner dominance in a republic—had left the United Party to form the Purified Nationalist Party. Though their electoral support remained relatively small in the '30s, the Purified Nationalists built up their power through the establishment of extensive welfare and economic subsidiary organizations, directed on the one hand at promoting the interests of Afrikaner petty bourgeoisie and on the other to gaining the support of the Afrikaner rural poor and workers in the towns through the provision of cultural, social, and trade union organizations outside the predominantly English-speaking Labour and Communist parties.

Though Hertzog now rejoined the Purified Nationalist Party, the alliance was uneasy; the history of the Nationalists during the war is a complicated series of manoeuvres by disunited factions. Through the war years, General Smuts, by this time a prominent world statesman and for the second time a member of the British War Cabinet, headed the government and the United Party, which was now supported by virtually all English-speaking whites and contained within its ranks both extreme right-wingers and liberals. The Afrikaner divisions, however, masked the fact that Smuts had lost the support of the group that by this time made up the clear majority of the South African white population.

WHITE SETTLERS NORTH OF THE LIMPOPO

Though the Union of South Africa had by far the most powerful and complex settler polity, in the first half of the 20th century settlers acquired firm control over Southern Rhodesia (now Rhodesia), too, while even in Nyasaland and Northern Rhodesia (now Malawi and Zambia), which came under the jurisdiction of the British Colonial Office in 1892 and in 1924, respectively, it was accepted that white settlers would provide the necessary economic development. Until 1923, Rhodesia was administered by the British South Africa Company. In 1898 a Legislative Council was established, and by World War I the 25,000 settlers had an elective majority in it. The colour-blind franchise qualifications, modelled on those of the Cape, were sufficiently high to exclude all but a handful of Africans. In 1922, when the company handed over formal responsibility for the territory to the British crown, the settlers rejected a proposal to join the Union of South Africa in favour of self-government. Again there was no attempt to include Africans in government; though the constitution provided for an imperial veto over discriminatory legislation, the veto was never exercised, and complete control of the police and armed forces was handed over to the minority government.

Though a number of white political parties flourished under a variety of names between the 1920s and 1950s, there were basically two parties of importance: (1) the "government" party, closely allied to big business interests, and (2) the "opposition" party, more eager to secure the position of white farmers and workers in the face of African competition. Outside formal politics, farmers, mining companies, and trade unions formed crucial pressure groups. Apart from socio-economic issues, the major issue in the late 1920s and '30s, as indeed later on, was over relations with Northern Rhodesia, where the wealth of the Copperbelt proved a considerable attraction.

Colonial rule in British Central Africa

Across the Zambezi the handful of white settlers were in a far weaker position than those farther south. In Northern Rhodesia, white traders and farmers, largely from South Africa, were granted an advisory council in 1918, but it had very little real power. When, in 1924, the company withdrew from its administration, the territory came under crown colony rule. Settler numbers jumped to 11,000 by 1930, with the discovery and exploitation of the rich deposits of copper by British, South African, and American capital. Various attempts by the whites of Northern Rhodesia in the 1920s and '30s to amalgamate with Southern Rhodesia in order to bolster their political position and to secure independence from imperial control were defeated by the Colonial Office, which enunciated a doctrine of trusteeship. Nevertheless, in crucial areas, such as the stabilization of African labour, equal pay for white and black, and the job colour bar, the Colonial Office was not prepared to counter the opposition of either the copper companies or white workers.

In Nyasaland, also, though the principle of African "paramountcy" was frequently proclaimed, it has been shown that British imperial interests generally prevailed and led to the territory's depressed economy. The Shire Highlands were considered highly suited to white settlement, and the demands of the planters there for labour and communications were acceded to. Moreover, in railway construction more attention was paid to the needs of British capitalist strategy than to those of the territory, which was saddled with huge public indebtedness.

Though the number of settlers in the Portuguese colonies was roughly comparable with those in Southern Rhodesia, they never achieved comparable political power. Despite the transfer to civilian government in the colonies during the republican period in Portugal (1910–26) and supposedly greater local autonomy, real power was in the hands of the governor general, or, as he became in 1921, the high commissioner, the highest colonial representative of the Portuguese government.

Republican Portugal

The republican period saw a flurry of activity among settler political groups, some of them in alliance with African *assimilados* and *mestiços* angry at Portuguese neglect and maladministration. This culminated in Angola in 1930 when the governor was forced temporarily to withdraw from Luanda. With the consolidation of António Salazar's regime in Portugal, however, strict government vigilance was exerted over all economic and political activity in the interests of stability, though in 1942 municipal politics were revived among the settlers. Though the greater control in the hands of the administration could, and at times did, restrain the exploitation of Africans by foreign interests, the final conquest of Africans, the greater power in the hands of the government, and increased demands for African labour and crop production led to their everyday lives being ever more strictly controlled.

LAND, LABOUR, AND COLOUR BARS

At the root of relationships between black and white in the 20th century in southern Africa are the closely connected issues of land and labour.

Alienation of land. At the beginning of the century, the vast majority of Africans in the subcontinent lived on the land, which in many areas is now claimed by white settlers. South of the Limpopo, the century-long battle over land and water was largely won by whites, though increasingly crowded pockets of land were reserved for Africans in the Transkei, Zululand, and the three High Commission Territories. In the latter, further land alienation to whites could not take place without the express agreement of the high commissioner, although in Swaziland land concessions granted to whites by chiefs in the 19th century had already alienated one-third of the land resources of the kingdom and constituted a continuing source of unrest and friction.

In the reserves and protectorates of southern Africa, governments retained African tribal organization and governed through sufficiently compliant chiefs and headmen. Economically, these areas were largely based on subsistence agriculture.

In the union about 40 percent of the African population lived on reserves, and about the same number lived on white-owned farms either as labour- or rent-paying tenants or sharecrop farmers (squatters). The poorly paid labourers were immobilized by master-and-servant laws and Pass Laws, which made strikes and desertion criminal offenses. Small numbers of Africans, from the second half of the 19th century, were successfully producing cash crops for the new markets; increasingly, however, because of government subsidizing of white farmers and preventing further land purchase by Africans, and because of African population pressure and soil erosion, these initiatives were crushed.

Natives Land Act of 1913

Probably the crucial blow to the Africans in South Africa was the Natives Land Act of 1913, a complex measure aimed at securing a supply of cheap labour by limiting African land ownership to the already existing reserves and by transforming African sharecroppers and tenants into wage labourers. Under this act and the complementary Native Land and Trust Act of 1936, about 12–13 percent of the land was scheduled for more than 4,000,000 Africans, while 87 percent of the land was reserved for 1,250,000 whites. Though the 1913 act prevented the total dispossession of Africans and ensured a rural base for migrant labourers where their welfare costs could be borne, it also led to the dislocation of thousands of Africans, who were to swell the numbers already working for whites.

In South West Africa, the white farming population increased at the expense of both the Herero and Nama. In 1921 the administration set aside apparently large reserves for the two indigenous groups, pursuing the policy of segregation established by the South African Natives Land Act, but these were generally in the dry hinterland of the Kalahari or on sandveld unsuited to man or beast. By the 1930s settlers were in control of more than a third of the land.

In Southern Rhodesia the pattern was not dissimilar, though Africans retained a higher proportion of land. After the Shona and Ndebele rebellions, the administration sold large tracts of land to settlers and speculators, with little concern for African rights. As hopes of vast mineral wealth were disappointed, the British South Africa Company turned to the development of white commerical agriculture—maize (corn), cattle ranching, and tobacco. By the 1920s, of 96,000,000 acres of land in Southern Rhodesia, 31,000,000 were in the hands of about 11,000 whites; 22,500,000 had been reserved for the 676,000 Africans; and 45,500,000 acres remained open to purchase and occupation by people of any race. In 1930 the Land Apportionment Act, modelled in part on South African legislation, attempted to define areas of ownership by barring African land purchase outside the reserves and by a special native purchase area of 8,000,000 acres, which soon proved inadequate.

As in South Africa, a considerable number of Africans "squatted" on white-owned and crown lands in Southern Rhodesia. By the 1920s, with the increased utilization of land by commercial farmers and their demands for labour, there was little room for such practices, while the initiatives of the Shona at cash-crop production were also crushed by the artificially created land shortage, discriminatory marketing practices, and the distance of the reserves from roads and railways. The rapid growth of the African population and its stock in the limited reserves, combined with its lack of security in urban areas, made the land question increasingly acute through the 1930s and '40s. Though in 1951 the Native Land Husbandry Act was adopted, allegedly to grapple with the problems of overstocking and erosion in the reserves and to implement the necessary agrarian reform, for a variety of reasons it remained unimplemented.

The Native Land Husbandry Act of 1951

In Northern Rhodesia, land never became as crucial an issue, except along the railway line in Tonga country and at Fort Jameson, where white settlers, mainly from South Africa, established themselves; in the reserves drawn up in 1928–29, however, there were problems of overcrowding and overstocking.

In Nyasaland, despite Johnston's attempts to control land alienation after 1896, about 15 percent had passed to whites by the turn of the century, particularly in the Shire Highlands, where tobacco, cotton, coffee, and tea plantations were established. On their lands, planters charged rent or demanded labour of Africans who wished to remain. Though by the end of World War II the amount of land in settler hands had been reduced to 5 percent, about 10 percent of the African population lived on white-owned plantations as tenants-at-will, squatters, or labourers. In the first three decades of the century, the administration favoured white plantation agriculture over African cash-crop production, though this gradually changed. In 1936 the bulk of the land was vested in a Native Trust. By the end of the war there was serious overcrowding in the southern provinces of Nyasaland, especially in Cholo and Mlanje districts, where, in 1953, the population density was estimated at 250 to the square mile. This congestion led to very serious rural discontent, which erupted in riots in Blantyre in 1943, in Cholo in 1945, and in Cholo and Mlanje in 1953.

In Angola and Mozambique at the beginning of the century, despite the then sparse settler population, a land policy was initiated by which all land not privately owned was in the hands of the state. Though Africans were to be left on their communally owned lands, the way was open to their expropriation in response to the demands of the concession companies and, later, of white settlers. In the Portuguese territories, however, land alienation never became the key method of forcing out

African labour, which was more directly provided for in the labour codes of 1878, 1899, and 1928. The codes institutionalized a variety of forced labour practices.

In addition, from the 1930s until 1961, with economic development directed solely to Portugal's benefit, Africans in certain areas had to pick stipulated cotton or coffee quotas, which had to be sold to the companies or government at a fixed price below the world market level.

African labour. In the British-controlled territories, as in South Africa in the 19th century, before Africans felt the effects of land shortage, the most important method of forcing out labour was taxation, which had the double advantage of producing revenue for administration. In Nyasaland, tax collection started in 1892 and extended over the entire territory by 1904. In Southern Rhodesia, the British South Africa Company imposed a hut tax in 1898, which was doubled in 1904, while in Northern Rhodesia the whole country became subject to tax in the first decade of the 20th century. Taxation increased steadily in all of the territories until World War II, though in many areas Africans working for whites received rebates.

Taxation

At the beginning of the century by far the most dramatic demand for labour resulted from the mineral discoveries in South Africa, and it was here that much of the subsequent pattern of southern Africa's industrialization was established. Initially, Africans went to the diamond fields and later to the gold mines for short periods, leaving their families behind on the reserves. The mineowners soon realized the advantages of employing unorganized migrant workers, who could be paid low wages on the grounds that they were single men who worked only for specific targets. On the diamond fields, Africans were housed in closed compounds to maximize control and prevent desertion. Both the compound system and migrant labour became features of the gold mines of South Africa and Southern Rhodesia, as well as of the copper mines of Northern Rhodesia in the 1920s and '30s.

With the opening up of the gold mines, demands for African labour expanded far beyond the frontiers of the Union of South Africa. In 1889 the Chamber of Mines was formed and soon came to control African wages by monopolizing the labour supply. By the early years of the century the chamber's recruiting agents had spread as far afield as Barotseland and Nyasaland.

From the first decade of the 20th century, by far the largest number of workers for the gold mines came from Mozambique. Under conventions signed between Portugal and the South African government in 1903, 1909, and 1928, Mozambique received a grant for every worker sent to the mines, while the bulk of the Africans' wages was handed over to Portuguese officials for repayment in the territory. The grant constituted a major source of the colony's income. In Angola this system had its counterpart in the contract labour sent to the cocoa plantations of São Tomé. In South Africa, where the effect of this ready supply of powerless labour enabled the Chamber of Mines to prevent any rise in real wages for Africans between 1911 and 1971, the system was relatively unremarked. In 1911 African wages in the mines were lower than they had been in 1897.

The justification for low wages

Part of the justification for the very low wages paid to black migrants in the white economy was, and still is, the assumption that their families were able to subsist on agriculture in the reserves. Nevertheless, the continual absence of the majority of adult men broke up families and contributed to the impoverishment, underdevelopment, and demoralization of the rural areas, while hampering the growth of working class consciousness in the towns and thus the improvement of urban wages and living conditions.

The job colour bar. While it was possible for the mining industry to acquire cheap unskilled labour through a variety of administrative techniques, skilled labour had to be attracted from overseas to the interior of South Africa by the promise of extremely high wages. Though by the turn of the century many Africans had acquired the requisite skills, and it was in the economic interests of the mineowners to substitute Africans for the increasingly expensive white labour, the political power of white workers and, later, trade union organization, together with the powerlessness of the Africans, enabled white workers to establish a job colour bar, which was legally entrenched in South Africa in the Mines and Works acts of 1911 and 1926.

In 1922 the Chamber of Mines in South Africa attempted to relax the job colour bar and cut white wages, but this led to a wave of strikes on the Witwatersrand that culminated in a five-day battle between white workers and troops and resulted in the firmer entrenchment of the job colour bar.

Though the job and wage colour bars had their counterparts in the Rhodesias and the Portuguese territories, on the Copperbelt of Northern Rhodesia white unionists were ultimately too weak to protect their privileged position. The major copper companies were the Anglo American Corporation of South Africa and the Rhodesian (later Roan) Selection Trust, with mainly American capital.

Though production slumped during the Great Depression, by 1935 the copper mines were already recovering. Ten years later Northern Rhodesia was one of the world's greatest copper producers, and the job colour bar, the enormous differential between black and white wages, and discriminatory practices had become essential features of the Copperbelt. By the mid-century, however, the growth of trade unionism and skills among Africans and the political weakness of the white working class enabled the mining magnates to replace white miners with trained Africans. Today the well-organized African miners on the Copperbelt constitute a highly paid elite.

Urbanization of white and black Africans. Although the first major battleground over the colour bar appeared in the mining industry, the presence of an increasing number of Africans in the urban areas outside the mining compounds brought new problems for government in southern Africa from the 1920s onward. By the turn of the 20th century, indeed, there was already a substantial number of Africans in the four major cities of South Africa, and by 1921 there were more than 500,000 Africans in towns. These numbers increased sharply to 1,250,000 in the late 1930s, as South Africa's economy recovered from the world depression and the growth of secondary industry led to greatly increased demands for African labour. By the end of World War II there were probably 2,000,000 Africans in the towns and by 1970 more than 3,500,000.

African and Afrikaner migration to the towns

From the beginning of the 20th century the stream of black migrants making their way to the towns was paralleled by a stream of Afrikaners displaced from the land. By the early 1930s the drift had become a flood, and one in five of the white population was classified as "poor white." Poor, unskilled, semi-literate, they were in direct competition with the equally poor and semi-literate Africans, some of whom had more industrial skills.

Although commission after commission attributed the plight of the "poor whites" to their refusal to do what was designated Kaffir work (*i.e.*, manual labour, Kaffir having been a derogatory name for the Xhosa), the Afrikaners were at a double disadvantage in the towns. Whereas the black Africans left their families to subsist, at however low a level, with their kin in the reserves, the Afrikaner was totally dependent on his urban wage. At the same time, at the high wages demanded by organized labour, he was unemployable. Not surprisingly, therefore, Afrikaner nationalism in the first half of the 20th century had a distinct anti-capitalist edge.

In the territories of Central Africa, also, by the third decade of the 20th century, Africans were increasingly making their way to the urban centres, pushed out of the rural areas by economic necessity and attracted by the higher wages and wider experience of the towns. Until the end of World War II, however, white administrators, capitalists, and settlers alike were determined that the towns should be the white man's preserve. Nowhere were authorities prepared to see the establishment of a perma-

nent African urban population. The results were low wages, appalling living conditions, poverty, and broken family life. It was only after World War II, when change was speeded up all over the continent, that more serious consideration was given to the problems of African urbanization, although even then, administrations, whether colonial or later independent African, could never keep pace with the move to the towns.

THE AFRICAN RESPONSES

The reactions of the African peoples who had been drawn painfully into the settler-dominated capitalist society of southern Africa and subjected to ever-increasing administrative, economic, and political control were varied. Once their traditional economic and social structures had been defeated and undermined, it became necessary to find new ways of expressing their opposition to white rule. In certain areas, dreams of resistance organized along traditional lines continued well into the 20th century, especially in the Portuguese territories but also in Mashonaland in the 1930s and in the High Commission Territories in the 1950s and '60s. Indeed, in isolated rural regions, there may still be small groups of traditionalists who have been relatively unaffected by the 20th-century nationalist struggles against colonialism.

Royal family politics

In some regions—pre-eminently those in which large centralized states had existed at the time of the colonial takeover—royal family politics continued to be of greatest significance. In Barotseland, Swaziland, and Basutoland, where the paramount chiefs were given special recognition by colonial regimes, the traditional aristocracy combined with the educated to protect their position and to demand redress of their grievances. In Matabeleland, the first decades of the 20th century were marked by an attempt to see a restoration of the kingship, while in Nyasaland in the 1930s the elite combined with traditionalists to seek a Tonga paramountcy and to restore the Ngoni king. In general, these attempts were most successful during the 1930s, when theories of indirect rule generally prevailed, and in areas where the settlers were weakest. In all territories, tribal and regional divisions and rivalries, despite African attempts to overcome them, continued to shape African responses to colonial rule and have remained easily exploitable in some areas.

The experience of white rule for Africans has been longest and most intense south of the Limpopo and in the Portuguese territories, although the existence of a substantial but unprivileged Coloured and Indian minority in the former and of *assimilados* and *mestiços* and settlers lacking political rights in the latter gave an extra dimension to anti-colonialism.

In South Africa, between 1906 and 1913, Gandhi led the first large-scale nonwhite, nonviolent resistance campaigns against anti-Indian legislation and gained very limited successes, though restrictions on Indian movement and immigration to South Africa remained in force. After his departure (1914), however (despite the formation of the South African Indian Congress in 1919), the initial militancy of Indian opposition was lost until after World War II, when younger, more radical groups won power from their moderate middle-class precursors who had dominated the party.

Compared to the Indians, the Coloureds of the Cape and Transvaal were never as successfully mobilized, though the African Political (later People's) Organization, founded as a South Africa-wide body in 1902, was the first large nonwhite political organization. It was, however, soon eclipsed by more vigorous African nationalist movements, and for most of the century the marginal Coloured population has been divided and relatively voiceless, though the predominantly Coloured Non-European Unity Movement, founded in the 1940s, has influenced African political thought.

The African response to settler rule in South Africa has in many ways been paralleled in the territories farther north. As early as the time of the unification of South Africa, educated Africans—teachers, preachers, clerks, small traders, and more prosperous farmers, as well as a sprinkling of professional men—had begun to feel the need for some organization larger in scale than the local vigilance and welfare associations that had grown up in the Cape in the last decades of the 19th century.

These were paralleled elsewhere in South Africa immediately after the South African War and in the Central African territories in the second and third decades of the 20th century. In 1912 the South African Native National Congress, which became the African National Congress (ANC) in 1917, was founded to represent African grievances, to overcome African tribal divisions, and to attempt to gain acceptance within the white polity through self-help, education, and the accumulation of property. Though demands for industrial education, individual land tenure, and some form of representation in Parliament later broadened to include attacks on the Pass Laws, the industrial colour bar, and the Natives Land Act of 1913, their methods remained strictly constitutional and they appealed mainly to the black middle class. Again, the territory-wide ANC has had its counterparts in Central Africa: in Southern Rhodesia in 1934, Nyasaland in 1944, and Northern Rhodesia in 1949–51. Though there have been local differences, the methods were similar, at least until after World War II, with South Africa leading the way.

African National Congress

Despite the increased numbers of Africans in industry in South Africa by the end of World War I, African trade unions were hampered by the Pass Laws, nonrecognition, and police harassment; spontaneously organized strikes were regarded as criminal offenses and often suppressed with violence. Nevertheless, in 1920 the Industrial and Commercial Workers' Union (ICU) was founded in Cape Town by the fiery Clements Kadalie, a migrant worker from Nyasaland. Based initially on Coloured workers in Cape Town, the ICU spread rapidly among Africans, both in towns and on the reserves and white farms of the countryside, claiming at its height some 100,000 members and establishing branches in Rhodesia and South West Africa by the mid-1920s. ICU, however, like the ANC, soon divided between those who wished to proceed constitutionally and those who demanded greater militancy and mass participation; by 1929 it had virtually disintegrated.

From the early 1920s in South Africa and from the 1930s in Central Africa, governments attempted to provide channels for the expression of African grievances through a variety of local consultative councils, often with a strong component of chiefs. Despite their powerlessness, educated Africans participated in them in the hope of gaining greater concessions, though by the end of World War II they were increasingly disillusioned with the prospects for reform.

Assimilado goals

In Angola and Mozambique during the republican period in Portugal (1910–26) the Congress movements and ICU had their counterpart in *assimilado* political organizations, trade unions, and a local press. For a brief period in Angola it seemed as if the *assimilados* and Europeans would fight for common reformist goals; but the Africans soon broke away from the less radical European organizations to form their own Liga Angolana and Centro (later Grêmio) Africano, which had their equivalents in the Grêmio (later Associação) Africano in Mozambique. These associations engaged in petitions to demand limited educational and welfare benefits and to make African grievances more widely known. In this they were partly inspired by the formation of political associations among colonial Africans in Lisbon from 1910 onward; these, by the 1920s, were influenced by broader Pan-African diaspora movements. In general, however, they were dominated by middle-class moderates and had little influence beyond the urban areas. Even before the advent of the Salazar regime, these nationalist stirrings were crushed by the autocratic governor of Angola Norton de Matos (1921–26). Though the associations revived as ostensibly social and educational organizations, officially recognized by the government in the 1930s and '40s, it was not until the 1950s that they returned to militantly political expression.

Political participation and organized political movements, even at the height of their popularity, never reached more than a fraction of the African population.

For the aspirations of many, however, independent churches, which broke away from the orthodox white-controlled missions, provided another outlet. In South Africa the first independent church was founded in the 1880s; by the turn of the century European administrators noted with alarm the growth of Ethiopianism, as it was called, with the existence of some 15 to 20 independent sects; by the 1960s there were well over 2,000, ranging in size and organization from little more than a family group to vast churches with thousands of members and links with the overseas world.

In the early years of the century, Europeans in South Africa interpreted the separatist church movement as politically dangerous and in some areas sought to crush it administratively. By and large, however, the independent churches in South Africa probably diffused African energies. The only violent episode involving the separatists in South Africa took place in 1921, when nearly 200 members of a sect known as the Israelites were killed by police gunfire when they refused to move from an African location near Queenstown in the eastern Cape, where they were living peacefully, if illegally, awaiting the millennium.

In northern Nyasaland in the first decades of colonial rule and in Northern Rhodesia in the 1930s, the millennial doctrines of the Watch Tower movement (Jehovah's Witnesses) were far more politically explosive. Within a year of its establishment in northern Nyasaland in 1909, Watch Tower had more than 10,000 adherents, who were partly attracted by its prophecies that British rule would end in 1914. In 1915 John Chilembwe, greatly influenced by its millenarianism as well as by his own colonial experiences, led an unsuccessful rebellion against British rule.

For the rural Shona, also faced by severe economic crises in the late 1920s, the millennial dreams of Watch Tower provided the most adequate substitute for traditional religious and political supports. As in Nyasaland, however, the outburst of Watch Tower was replaced relatively quickly by more subdued Christian independency, which, by the mid-century, had tens of thousands of followers and offered solutions to contemporary problems by drawing on fundamental ideas of Shona religion as well as by providing form and discipline in times of social upheaval.

Prophet movements

In Angola, too, and especially among the Kongo (living in northern Angola and the Belgian and French Congos), prophet movements, influenced by the Protestant missionaries and Watch Tower, drew on traditional religious and cultural beliefs, imparting as well a specifically anti-European message, which partly prepared the ground for the anti-colonial war of the 1960s.

Though the initial break with the mission church betokened a desire for independence from the white man, there were many motives, including the religious, for separatism. The prophet movements and independent churches expressed the millennial dreams of an insecure and poverty-stricken population torn from its traditional moorings.

By 1945, despite the fact that Africans had not established mass nationalist movements anywhere in the subcontinent, a remarkable range of protest movements existed, which became merged in many areas in the next decade to produce modern nationalist parties and give them their characteristic shape.

IV. Southern Africa since 1945

If, in the first half of the 20th century, policies in the southern African territories were remarkably similar, after World War II their political destinies diverged. During World War II, politics all over southern Africa appeared to stand still. Nevertheless, when the colonizers turned their attention once again to the African colonies, it was in the context of a very different world. The colonizing nations emerged from the war as secondary powers, under attack from the United States and the Soviet Union, which, for their own reasons, demanded rapid decolonization. The Bandung Conference of Asian and African nations (1955), the founding of the United Nations, and the liberation of South Asia were significant milestones also for the colonized peoples of Africa. There was increased awareness in Britain that the African territories, too, had to be prepared for independence, and the tentative beginnings of colonial development were accelerated. The indirect rule policies of the 1930s were abandoned, and in all territories steps were taken toward granting representative institutions.

The economic effects of the war had been profound. In South Africa and Southern Rhodesia it had brought about a tremendous increase in urbanization and secondary industry. The boom on the Copperbelt shifted the economic focus in Central Africa northward and further radicalized African politics; and in the Portuguese territories, greatly accelerated white immigration and the coffee boom in Angola had a similar effect in the 1950s.

In the south these developments led to the victory of the (Afrikaner) National Party, determined that whites should maintain total social, political, and economic control in both South and South West Africa. In the British Central African territories, after a short-lived experiment in multiracial federation, they resulted in 1964 in the emergence of independent African states in Malawi (formerly Nyasaland) and Zambia (Northern Rhodesia) and in 1965 in the unilateral and unconstitutional declaration of independence by white settlers in Rhodesia.

Independent states in the 1960s and '70s

For the High Commission Territories, the "winds of change" in Africa also brought political independence: Lesotho (formerly Basutoland) and Botswana (formerly Bechuanaland) in 1966 and Swaziland in 1968. In Angola and Mozambique violent anti-colonial struggles that lasted more than a decade led to the downfall of the Salazar regime in Portugal and to the ultimate independence (1975) of the territories themselves.

Despite the political differences, in socio-economic terms the subcontinent has been welded together into areas of high capital development, supported by rural areas of increasing impoverishment that provide a large reservoir of cheap labour. Most striking has been the very rapid industrial growth in South Africa, which accounts for 74 percent of production in the area.

As in the first half of the century, white South Africa, with its immense wealth, dominates its neighbours, black and white. The Portuguese withdrawal, however, together with increasingly explosive internal pressures, has opened a new era of uncertainty for white rule in the subcontinent.

SOUTH AND SOUTH WEST AFRICA

In 1948 the tentative and ambiguous moves in South Africa toward less rigid social policies were swept aside by the electoral victory of the National Party, which had reunited after the dissensions of the war years under the leadership of Daniel F. Malan (prime minister 1948–54). Though the Nationalists fought the election on the "race issue" under the new slogan of apartheid, the essence of the change was over which group of whites was to control the South African economy. Whereas up to 1948 every Cabinet had drawn its members from both language groups, the Nationalists were able, unlike any previous Afrikaner party, to unite the majority of their people behind them and to consolidate the victory in the following 20 years. Initially, the Nationalists secured only 40 percent of the vote, but by 1966 this had increased to 60 percent. By this time an increasing number of English-speaking voters had accepted the government claim that it alone was capable of securing white dominance—this despite the fact that in 1960–61 South Africa was forced out of the British Commonwealth and declared a republic.

Behind the Nationalists' electoral success lay South Africa's accelerated industrialization. Although this process was already under way by the end of World War II, South Africa's economic growth between 1948 and 1970 was surpassed only by that of Japan, undoubtedly partly the result of government intervention in the control of labour, the direction of investment into industry, and the provision of essential infrastructure through the increased number of para-state corporations designed to increase

Industrialization in South Africa

the Afrikaner hold on the economy. Apart from 1948 and 1960–61, when there was a flight of foreign capital from South Africa in response to the regime's apparent political instability, capital investment from the United Kingdom, the United States, and western Europe, attracted by the high rate of profit, contributed greatly to South Africa's economic growth. Since 1945 secondary industry has replaced mining and agriculture as the major economic sector.

Thus, by the 1960s, if not before, white South Africans enjoyed one of the highest standards of living in the world. It was achieved very largely at the expense of Africans, through the rigorous control of their labour power. The reverse of the high per capita income of the small white population is the extremely low per capita income of the blacks. Thus, in 1969, the average monthly wage of a white man was $133, that of a black man $10. Despite wage increases, in real terms wages were little different in the mid-1970s, while the absolute white–black wage gap had widened.

Extending the philosophy that only those Africans who were of service to the whites were to be allowed in towns, remaining African rights in the urban areas were removed, and thousands of Africans returned to their so-called tribal homelands. The basis of the black work force, both in the towns and on white-owned farms, reverted to the system of migrant labour, which had been breaking down since the 1930s. As a result, in the mid-1970s some 2,500,000 Africans were arrested each year for statutory offenses—750,000 for infringing the Pass Laws alone.

In the 1950s Hendrik Verwoerd, the minister for native affairs and from 1958 to 1966 prime minister, developed the second aspect of the National Party's apartheid policies. Verwoerd argued that the only solution to racial antagonisms in South Africa was to balance African rightlessness in the white areas with the establishment of ethnically defined "Bantu Homelands," where Africans could develop semi-autonomous states, or "Bantustans." The territorial bases for the Bantustans were the African reserves set aside in 1913 and 1936. In 1954–55, the government-appointed Tomlinson Commission, which produced a plan for territorial segregation, admitted that the reserves were incapable of carrying their current population, let alone the increase necessary if Africans were to be removed from the white areas. Nor was it ever intended that this should happen; though more than 3,000,000 Africans have been forcibly moved out of white rural areas and so-called surplus Africans have been moved from towns under a variety of apartheid laws to the poverty-stricken reserves, the number of rightless Africans returning regularly to so-called white areas has nonetheless steadily increased to keep pace with labour demands.

Authority of the chiefs

Legislation in 1951 and 1959 attempted to resurrect the authority of the chiefs in the reserves and establish tribal, regional, and territorial advisory councils. While many hereditary chiefs still retained some popular support, the new policy created profound discontent in the rural areas as chiefs walked a tightrope between popular hostility if they implemented government policy and government anger if they failed to do so. In several rural areas there was violent resistance and apartheid chiefs were assassinated, while in the Transkei, which held its first Bantustan elections in 1963, a state of emergency had to be declared and was still in force in the 1970s. Under the constitution granted to the Transkei and later to the other Bantustans, the South African government still controls crucial aspects of internal and external affairs. "Full independence" was demanded by the Transkei. This was opposed by other African leaders, who feared the fragmentation of their cause, but the Republic of Transkei was nonetheless proclaimed on October 26, 1976. Despite the militant tone of some Bantustan leaders, notably Chief Gatsha Buthelezi, whose Zulu constituency in KwaZulu is the largest of any "homeland," their room for manoeuvre is limited by their total economic dependence on South Africa. In the 1960s the Bantustan policies were also extended to South West Africa.

The Coloured and Indian minorities, who cannot be repatriated to "tribal homelands," have also suffered from such segregatory legislation as the Group Areas Act and the Population Registration Act, which led to widespread removals of people.

During the 1950s the practices and policies of apartheid in South Africa were under heavy fire both internally and abroad. India's complaints in the United Nations in 1946 about South Africa's treatment of the Indian population initiated a series of UN resolutions condemning South Africa. Its treatment of Africans in South West Africa became a serious point of contention at the United Nations from the 1950s. The issue was taken to the International Court of Justice at The Hague, which finally gave judgment against South Africa in 1970. With the admission of new African nations to the UN and the formation of the Organization of African Unity in 1963, the international attitude became increasingly hostile.

Internally, too, the 1950s was a decade of unprecedented opposition to government policies. The growth of African and Indian militancy after the war may have contributed to the National Party victory. The measures of the Nationalist government gave the deliberations of the African National Congress new urgency. With the suppression of the Communist Party in 1951, many white radicals joined the liberation struggle, while the increasing awareness of the need for a united opposition among Indian, African, and Coloured leaders found expression in a passive resistance campaign in 1952 and in the founding of the Congress Alliance by African, Indian, and white opponents of the government in 1955. Despite the arrest in 1956 of more than 150 Congress leaders on charges of treason (all of them were subsequently released), the later 1950s saw increased militancy among the African nationalists. In 1958 the Pan-Africanist Congress (PAC) was formed, and in 1960 it launched a campaign against the Pass Laws. It was during this campaign that police fired on a peaceful crowd of unarmed demonstrators at Sharpeville, killing about 70 persons and wounding hundreds of others; this led to an even greater outcry overseas against South Africa's racial policies and to the South African government's banning of both the ANC and the PAC.

Growth of Indian and African militancy

Though underground sabotage movements were founded in 1961–64, the wide use of spies and informers —a regular feature of the regime—and draconian security legislation led to their swift collapse. Many Africans fled to Zambia and Tanzania, where guerrilla training was undertaken and outside support sought.

Since 1945 a liberal tradition has continued in South Africa, with the emergence of a Liberal and later a Progressive Party, and there have been constant protests against government policies from students at English-language universities, from prominent church leaders, and from such organizations as the Springbok Legion, the War Veterans' Torch Commando (which faded away after the election of 1953) and the Women's Defence of the Constitution League (called the Black Sash). Essentially, however, the opposition movements were weak and divided. White liberals were never wholeheartedly behind the destruction of white law and order, while the ANC, though increasingly militant, was easily overwhelmed by a government with vast military and security resources.

Opposition movements

After 1973, however, black opposition again began to mount within the country. Initially, most serious for the government was the industrial unrest, particularly in Durban and in the gold mines, where much violence occurred, at a time when the future of the external migrant labour supply on which the mines are dependent was diminishing. A variety of black worker organizations was formed, while black students, partly influenced by the black power movement in the United States, formed the militant South African Students Organization, with African, Indian, and Coloured members; this had its counterpart in the Black Peoples' Convention. All these groups were encouraged to more militant aspirations by the Frelimo victory in Mozambique (see below).

Nevertheless, despite a number of ambiguous statements by the prime minister, B.J. Vorster, and at the

United Nations, there was no lessening in the essentials of apartheid. From 1975, increasing inflation and black unemployment (officially estimated at more than 17 percent in 1976) exacerbated tensions in the African townships. By April 1976, Chief Buthelezi was warning of serious trouble ahead if the government did not change its course, and this erupted in June, ostensibly over the enforced use of Afrikaans as the medium of instruction in black schools. Although the government retracted this policy, it was only after police had fired on thousands of demonstrating schoolchildren and the disturbances had spread over most of the republic, especially to the black townships around Cape Town. Unrest continued through the second half of the year, with thousands dead, injured, and imprisoned. Key symbols of apartheid were destroyed—beer halls (the proceeds of which are used to finance "Bantu" administration) and administrative buildings and schools—and the protest extended far beyond the original language issue. Thus increasing internal strains within South African society were paralleling the changing external pressures, and the widely held assumption that economic growth would improve the lot of Africans had been proved false both by events since June 1976 and by the increasing evidence of black unemployment and rural impoverishment.

South Africa and its black neighbours

The weak and divided black African states—despite their protests of the early 1960s—posed little threat to South Africa, which until the Portuguese coup seemed a corner of stability in the continent. In improving its diplomatic and economic relations with its black African neighbours, South Africa established close trade and diplomatic relations with Malawi and improved its customs agreements with the former High Commission Territories. The Cabora Bassa hydroelectric scheme aimed at uniting South Africa even more closely with the territories of Central Africa, while roads and railways connected South Africa with South West Africa and Angola. With the growth of secondary industry, South Africa found it increasingly essential to establish markets in Africa. The Anglo American Corporation of South Africa, with its vast interests in the subcontinent, played an important role in fostering the policy of "détente" and in maintaining overseas confidence in the regime.

With the withdrawal of the Portuguese from the region, détente became more urgent for security reasons. Realizing Rhodesia's vulnerability once Mozambique and Angola achieved independence, South Africa joined with black African leaders to try to achieve a negotiated settlement of this issue.

In August 1976 a constitutional committee (the Turnhalle Committee) in South West Africa (Namibia), with the tacit approval of the South African government, announced plans for a form of political independence in the territory by the end of 1978. The South West Africa People's Organisation, however, which the UN recognizes as representing the territory's inhabitants, was excluded from the committee's discussions, and the guerrilla struggle intensified on the northern frontier.

THE HIGH COMMISSION TERRITORIES

Their proximity to the economically dominant South Africa has been most acutely felt by the three High Commission Territories. Despite British abandonment of the idea of their political incorporation into South Africa after 1948, the absence of a large proportion of the working population as migrant labour in South Africa—as high as 43 percent in Basutoland and 28 percent in Swaziland in the 1960s—stultified economic and constitutional advance. From the mid-1930s, British aid, though inadequate, was made available, agricultural improvement schemes were instituted, and a Native Treasuries Finance Committee and advisory councils were established; attempts were also made to reform the courts and, most importantly, the system of chieftainship, a necessity in Basutoland in particular, where petty chiefs had proliferated. Most of these reforms became more effective after World War II, and by 1960–61 legislative councils had been established in all three territories. Basutoland and Bechuanaland became the independent states of Lesotho and Botswana in 1966; Swaziland, where the existence of a number of white settlers and considerable South African economic interest complicated the transfer of power, became independent in 1968.

In all three territories, but especially in Swaziland, where traditional chiefs have remained extremely powerful, modern nationalist movements emerged relatively late. In all three, conservative governments anxious to avoid provoking their powerful neighbour emerged in the first elections after independence, though Swaziland, with greater economic potential and a shrewd, if autocratic, king, and Botswana, with its outlet to the north and sophisticated leadership, have had greater scope for manoeuvre than small Lesotho, which is entirely surrounded by South Africa.

BRITISH CENTRAL AFRICA

In the colonies of British Central Africa—Northern and Southern Rhodesia and Nyasaland—the immediate postwar years also brought significant economic and political changes. In Southern Rhodesia as in South Africa, secondary industry overtook mining and agriculture as the main sector of the economy; and from 1946 the influx of emigrants from Britain averaged 10,000 a year, increasing the white population to 125,000 by 1953. At the same time, in Northern Rhodesia the copper boom set off new conflicts as the colour bar intensified, while as early as 1935 and 1940 widespread strikes by African miners showed the potential of the African working class. Settlers, led by the white labour leader and former railway worker Roy Welensky, began agitating for some form of amalgamation with Southern Rhodesia.

Powerful economic arguments were used; and big business, especially the powerful Anglo American Corporation of South Africa, with its interests in both territories, gave the movement its backing. The National Party victory in South Africa in 1948 gave it further impetus. The British Colonial Office, anxious to strengthen Nyasaland's weak economy, insisted on its inclusion in any form of closer union. In 1953 the Federation of Rhodesia and Nyasaland (Central African Federation) was formed.

Federation of Rhodesia and Nyasaland

In all three territories, where economic developments and the general trend of world opinion had strengthened the African political organizations of the interwar period, politically articulate Africans were almost unanimously opposed to federation. As talk of closer union mounted among whites between 1949 and 1953, the Congress movements became more widespread and representative. In all three, Africans feared that federation would lead to further land alienation, undermine the position of chiefs, prevent political advance, and extend farther north the colour bar and discrimination of Southern Rhodesia.

With the establishment of the federation, however, the Africans, especially in Southern Rhodesia, were initially cooperative. But by the mid-1950s, the economic advantages of federation seemed so exclusively to benefit the whites of Southern Rhodesia that African hostility was re-aroused. Though federal leaders preached racial partnership, and in Southern Rhodesia urban African housing was improved and a multiracial university built, greatly increased white immigration from South Africa and Britain intensified black–white conflict. The continued demands by federal leaders for dominion independence status confirmed the Africans' worst fears.

In Nyasaland by 1955 a group of young radicals had emerged to demand universal suffrage and self-government and for the first time turned to the masses to achieve their ends. By 1957 they claimed 60,000 members and were developing a modern mass nationalist party. Feeling the need for a leader of standing they invited Hastings (later H. Kamuzu) Banda, at that time practicing medicine in Ghana, to return and take over the leadership of the Congress movement. Banda returned in 1958 and vigorously denounced federation.

In Northern Rhodesia, African political opposition revived in 1957 after a period of quiescence, when the Congress organization had been eclipsed by the rise of the millenarian Lumpa Church of Alice Lenshina. In

1958–59 the removal of the more conservative president of Congress and the influx of new militants under the leadership of Kenneth Kaunda transformed the organization. Like its counterpart in Nyasaland, it became far more radical and created branches all over the country.

Southern Rhodesian opposition became more radical in the later 1950s also, as urban and rural discontents were united for the first time, and a group of young leaders formed the Southern Rhodesian African National Congress under Joshua Nkomo in 1957. This was later to become the Zimbabwe African People's Union (ZAPU).

The rise of mass nationalist movements brought direct confrontation between the Africans and the colonial and federal authorities. In 1959, after sporadic disturbances in Nyasaland, a state of emergency was declared, nationalist leaders were arrested and their parties banned, to be replaced in all the territories by new parties under new names by 1960. The banning of the nationalist parties was followed by disturbances and disorders all over the federation, but particularly in Southern Rhodesia, where the worst rioting occurred in the impoverished African townships. While in Nyasaland and in Northern Rhodesia the rioting was sufficient to persuade a distant Colonial Office to grant majority rule, in Southern Rhodesia it increased white determination to hold more tightly onto their position. In 1961 Nyasaland and Northern Rhodesian leaders were released, and new constitutions were granted to the territories; and in 1963, after further disorder in Northern Rhodesia, the federation was finally dissolved. In the following year Northern Rhodesia and Nyasaland became the independent states of Zambia and Malawi, governed by the nationalist leaders who had fought for independence.

Despite independence, however, many problems remained in both territories. In Malawi the unity of the nationalists was short-lived, and after an abortive uprising in 1965 Banda stifled all internal opposition. Regional differences—the result of Malawi's uneven economic and educational development—continued, while Banda's policy of close ties with settler regimes was resented by many black Africans, large numbers of whom were still dependent on the white south for wage employment.

Independent Zambia

When, in 1964, Zambia achieved its independence, it was still a poverty-stricken and backward country, despite the potential of the Copperbelt. Indeed it was not until 1964 that the British South Africa Company ceded its concession rights to copper royalties, which had, over the years, gone to enrich the company and the British treasury rather than the territory itself. The result was that very little capital had been made available for investment in the increasingly impoverished rural areas or for welfare and educational services. Despite attempts, since independence, at reform, the countryside continues to be drained of manpower as thousands of hopeful young people make their way to the towns. In the towns, however, neither housing nor welfare is adequate, and unemployment grows. Fluctuations in the world price of copper and the oil crisis in the 1970s greatly aggravated Zambia's weakness; regional, ethnic, and rural–urban divisions persisted also. During the period of the transference of power, Lozi separatism was a serious problem, while the Lumpa Church became increasingly opposed to the government. In 1964 open warfare broke out between the church and the state; more than 700 people were killed and the church was banned.

In both Malawi and Zambia, but especially in the latter, the problems were compounded by developments in Southern Rhodesia. There, the violence in the early 1960s as well as events in the rest of Africa roused white fears and led to an increasing number of emergency laws against political opponents. Though limited concessions were made to Africans by the government in 1961, these were rejected both by the African nationalists, who refused to participate in the new elections, and by white farmers and artisans, whose support for the new right-wing Rhodesian Front Party led to its electoral victory in 1962.

With the end of federation, the Rhodesian Front, after crushing African opposition and increasing its internal support, renewed its demands for independence. Britain was unwilling to concede to these demands without evidence of progress toward majority rule and of the end of racial discrimination. In 1965, in a highly emotional atmosphere, the Rhodesian Front, under the leadership of Ian Smith, unilaterally declared Rhodesian independence. Total opposition to the prospect of black rule became policy.

Rhodesian independence

Despite international pressure, Britain refused to use force against the illegal regime, though it did support the economic sanctions imposed by the UN. These, however, were never fully effective, largely because of continued support of Rhodesia by South Africa and Portugal. Sanctions did, however, put considerable strain on Zambia, whose main routes of communication with the outside world lay through Rhodesia and which was dependent on Rhodesia for coal and for manufactured goods. Nevertheless, attempts were made by President Kaunda to carry out sanctions and, more important, to find alternative outlets by constructing roads and a railway to Tanzania; the latter, the Tan-Zam Railway, was completed in 1974.

With Rhodesia's declaration of independence, many African nationalists fled across the Zambezi to mount guerrilla resistance. Crippled, however, by internal rivalries, this proved ineffective until Mozambique's independence transformed the situation. Until then, Britain's attempts to achieve a negotiated settlement had failed. With the realignment of forces after the Portuguese withdrawal, however, Ian Smith came under unrelenting pressure from South Africa to release and negotiate with imprisoned black nationalist leaders. They in turn were under pressure from Zambia and Tanzania to engage in talks with Smith, in an attempt to avoid war. At the end of 1975, despite divisions within African nationalist ranks, the veteran leader Joshua Nkomo and Ian Smith undertook to try to establish a basis for talks.

In the spring of 1976, after South Africa's invasion force had to withdraw from Angola (see below), the pressures were intensified. Mozambique opened its frontiers to training and base camps for Rhodesian guerrilla fighters, who intensified their war in Rhodesia (now being called Zimbabwe). Fears of open war and Communist intervention led the Western powers to the view that it was necessary to get rid of minority rule, at least in its extreme form as represented by the Smith regime, before the last chance of a gradualist compromise vanished. The U.S. secretary of state, Henry Kissinger, had two meetings in Europe with the South African prime minister, B.J. Vorster, who realized that he might have to sacrifice the Smith regime in the interests of his own. After meeting Kissinger in Pretoria on September 19, 1976, Smith told his electorate, though in somewhat ambiguous terms, that majority rule would have to come in two years' time. A conference under the chairmanship of a British diplomat assembled in Geneva at the end of October to plan the transfer of power; whether the conflicting demands of the nationalists and of the Smith regime were capable of peaceful resolution was doubtful.

Rhodesian guerrilla pressures

ANGOLA AND MOZAMBIQUE

In the Portuguese territories, developments since World War II have in some ways been more dramatic than elsewhere in the subcontinent, with more than a decade of warfare leading to an army coup in Portugal itself and the ultimate transfer of power in the African territories. In both Angola and Mozambique, economic progress and development lagged behind the rest of the subcontinent. The Portuguese dictator António de Oliveira Salazar blocked economic development in the colonies for years in the interests of "stability" and Portuguese advantage, and it was not until 1961 that the territories were opened to overseas investment. Nevertheless the coffee boom in Angola in the 1950s and the discovery there of minerals and petroleum largely stimulated immigration and outside interest in the territory. White immigration increased Angola's settler population by 400 percent between 1940 and 1960 and doubled that in Mozambique between 1950 and 1960. Though this im-

migration was mainly to the towns, the government also instituted large agricultural settlements for whites, especially in the coffee-growing areas of Angola and in the fertile river valleys of Mozambique. Despite Portuguese claims that their policies were based on racial harmony, the increase in white settlement led to greater demands for African land and labour, while economic development brought little benefit to the African population.

The general movement toward colonial independence also affected the Portuguese territories. With the Indian attack on Goa in 1961, world attention was drawn to Portuguese colonial policies. Salazar responded by making the territories overseas provinces of Portugal and by granting limited local autonomy and increased participation in elections to the Portuguese National Assembly to whites and the handful of *assimilados*.

In the face of these new developments, African nationalist movements revived among *assimilados*, secondary school students, and trade union members in the towns. In neither of the territories, however, were the nationalists equipped for a political or military struggle. The divisions between the *assimilado* and *mestiço* elite in the towns and the rural masses, especially in Angola, were probably greater than elsewhere in southern Africa, while the meagre industrial and urban development and educational facilities before the 1950s militated against any broad national consciousness. Apart from a few areas where Protestant missionaries provided the organizational example, there were few nationalist leaders outside the towns.

In Luanda, however, from the early 1950s political associations proliferated. From among these the Movimento Popular de Libertação de Angola (MPLA), with a strong Marxist strain, emerged as the most important group by 1956, though an abortive uprising in Luanda in 1961 led to its temporary suppression by the Portuguese secret police. From the mid-1960s, the fortunes of the MPLA revived under its poet–president Agostinho Neto; moving its headquarters from Brazzaville (in the Congo) to Lusaka (in Zambia), it extended guerrilla operations in eastern Angola and resuscitated urban political activity. Although it failed to build up its base in eastern Angola, where it was supplanted by Jonas Savimbi's União Nacional pela Independência Total de Angola (UNITA) in the late 1960s and early '70s, the MPLA retained its hold on the capital and among Mbundu speakers.

The MPLA in Angola

In March 1961 the Kongo people of northern Angola—an area suffering most acutely from land alienation and forced labour practices, and with a long anti-colonial tradition—revolted. With their favourable terrain and the support of their newly independent kinsmen in Zaire, the most militant of the Kongo political organizations, the União das Populacões de Angola (UPA), later the Frente Nacional de Libertação de Angola (FNLA), threatened the Portuguese hold in the north. Hampered by ethnic, ideological, religious, and personal rivalries, and the constant proliferation of Kongo exile political organizations in Zaire and the Central African Republic, the movement failed to spread beyond the Kongo area. Despite Portuguese military reprisals, however, a war of attrition continued, with neither side able to press home its advantage for more than a dozen years.

In Mozambique the nationalist organizations, though also initially plagued by divisions, were more successful in overcoming their disunity. In 1962 three exiled organizations based in Tanzania came together to form the Frente de Libertação de Moçambique (Frelimo) and began organizing a political and military underground movement. By building on pre-existing political networks, such as the cooperative movement in northern Mozambique, and by educating selected leaders, Frelimo carefully prepared the ground before moving its guerrilla fighters into the northern provinces of Mozambique in 1964. Within three years they claimed to have an army 8,000 strong and to have established their own administrative, educational, and marketing networks in Cabo Delgado and Niassa districts.

Though Portugal's initial response to the outbreak of revolt was savage military repression, and some 70,000 to 100,000 Portuguese troops were kept in each territory, the uprisings produced a number of reforms. It was, however, always a case of too little and too late, and ultimately the sheer cost of the wars led to the revolt of Portuguese army officers and the downfall of the regime in Europe. After a short transitional period, Mozambique became independent in June 1975 under a Frelimo government headed by Samora Machel. It faced a number of urgent political and economic problems, largely connected with the poverty of its resources, the need to create a new state, and the nature of its relationship with Rhodesia and especially with South Africa. During this difficult period, it found it necessary, at least in the short term, to establish a relationship with neighbouring South Africa.

Independent Mozambique

In Angola, unlike Mozambique, the transitional period found the three main liberation movements—MPLA in the centre, FNLA in the north, and UNITA in the south—vying for control of any future government. Their rivalries were greatly exacerbated by the involvement of the great powers, Portuguese settlers, Zaire, and South Africa in their struggles. When on November 11, 1975, the governor general withdrew from Luanda, the country was in the throes of a civil war, which threatened to expand far wider as the Soviet Union flew in arms in support of MPLA and the U.S. indirectly aided the FNLA–UNITA alliance through Zaire. Meanwhile South Africa, equally anxious to avoid an MPLA government, played a hazardous interventionist role which drew its troops from the Angola–Namibia frontier almost to the outskirts of Luanda in support of UNITA. Although the South Africans were forced to withdraw by MPLA, now assisted by Cuban troops, and eventually MPLA under the leadership of Agostinho Neto established its control, sporadic fighting, especially in the UNITA-dominated region in the south, continued.

BIBLIOGRAPHY

General Works: J.D. FAGE and R.A. OLIVER, *A Short History of Africa* (1975), helpful for placing southern Africa in the context of African history; *The Cambridge History of the British Empire*, vol. 8, *South Africa, Rhodesia and the High Commission Territories*, 2nd ed. (1963); P.L. VAN DEN BERGHE, *South Africa: A Study in Conflict* (1965), E.A. WALKER, *A History of Southern Africa*, 3rd ed. rev. (1962), are all useful for basic factual information; MONICA WILSON and L.M. THOMPSON (eds.), *The Oxford History of South Africa*, 2 vol. (1969–71), the only general reference work to make a serious attempt to record the history of all the peoples of South Africa; *The Cambridge History of Africa* (in process) contains long chapters on southern and Central Africa; HEINRICH VEDDER, *Das alte Südwestafrika* (1934; Eng. trans., *South West Africa in Early Times*, 1938, reprinted 1966), mainly useful for the 19th century; J.H. WELLINGTON, *South West Africa and Its Human Issues* (1967), a general reference work on the area; R.P. STEVENS, *Lesotho, Botswana and Swaziland: The Former High Commission Territories in Southern Africa* (1967); A.J. WILLS, *An Introduction to the History of Central Africa* (1964), informative though a little dated in parts; LEWIS H. GANN, *A History of Southern Rhodesia: Early Days to 1934* (1965) and *A History of Northern Rhodesia: Early Days to 1953* (1963), informative general studies; A. ROBERTS, *A History of Zambia* (1976), the most authoritative outline; R.S. HALL, *Zambia* (1965), a readable and reliable account; T.O. RANGER, *Aspects of Central African History* (1968), and ERIC STOKES and RICHARD BROWN (eds.), *The Zambesian Past: Studies in Central African History* (1966), both contain essays by leading authorities on the history of Rhodesia, Zambia, and Malawi, embodying new interpretations; see also B. PACHAI (ed.), *The Early History of Malawi* (1972); R.H. CHILCOTE, *Portuguese Africa* (1967); and DOUGLAS L. WHEELER and RENE PELISSIER, *Angola* (1971), general accounts.

Southern Africa to 1800 (*Early man and the Stone Age*): J.D. CLARK, *The Prehistory of Africa* (1970), "Africa in Prehistory: Peripheral or Paramount," *Man*, vol. 10, no. 2 (1975), and *Kalambo Falls Prehistoric Site*, vol. 2 (1974); H.J. DEACON, "A Review of the Post-Pleistocene in South Africa," *South African Archaeological Society*, Goodwin Series I (1972). (*The coming of the Iron Age*): This is a rapidly developing field of study, and older interpretations are being constantly challenged and general works need to be constantly supplemented by reference to specialist articles (*e.g.*, in the *Journal of African History*). See also B.M. FAGAN, *Southern Africa During the Iron Age* (1965); P.L. SHINNIE (ed.), *The African Iron Age* (1971); R.A. OLIVER and B.M. FAGAN, *Africa in the Iron Age, c. 500 B.C. to A.D. 1400* (1975); P. GARLAKE,

Great Zimbabwe (1973); and T.M.O'C. MAGGS, *Iron Age Communities of the Southern Highveld* (1976). (*The Portuguese in west central and east central Africa*): E.A. ALPERS, *Ivory and Slaves in East Central Africa* (1975); M.D.D. NEWITT, *Portuguese Settlement on the Zambesi* (1973); A.F. ISAACMAN, *Mozambique: The Africanization of a European Institution: The Zambesi Prazos, 1750–1902* (1972); DAVID BIRMINGHAM, *Trade and Conflict in Angola: The Mbundu and Their Neighbours Under the Influence of the Portuguese, 1483–1790* (1966) and *The Portuguese Conquest of Angola* (1965), a shorter, more popular account of the material covered in the preceding scholarly monograph; J.C. MILLER, *Kings and Kinsmen: Early Mbundu States in Angola* (1976); C.R. BOXER, *Salvador da Sá and the Struggle for Brazil and Angola, 1602–1686* (1952) and *Race Relations in the Portuguese Colonial Empire, 1415–1825* (1963). (*The Dutch at the Cape in the 17th–18th century*): C.R. BOXER, *The Dutch Seaborne Empire, 1600–1800* (1965); M.W. SPILHAUS, *South Africa in the Making, 1652–1806* (1966); S.D. NEUMARK, *Economic Influences on the South African Frontier* (1957).

Southern Africa, c. 1800–1900: (*The Mfecane and its effects*): J.D. OMER-COOPER, *The Zulu Aftermath* (1966), the standard account covering the entire process in southern, Central, and East Africa; E.A. RITTER, *Shaka Zulu: The Rise of the Zulu Empire* (1955), a popular but useful account. D. BEACH, "Ndebele Raiders and Shona Power," *J. of Af. Hist.*, vol. 14, no. 4 (1974); L.M. THOMPSON, *Survival in Two Worlds: Moshoeshoe of Lesotho, 1786–1870* (1975); P. SANDERS, *Moshoeshoe, Chief of the Sotho* (1975). (*The slave and ivory trade north of the Zambezi*): A.D. ROBERTS, *A History of the Bemba: Political Growth and Change in North-Eastern Zambia Before 1900* (1973); H.A.C. CAIRNS, *Prelude to Imperialism: British Reactions to Central African Society, 1840–1890* (1965); M.V. JACKSON HAIGHT, *European Powers and South-East Africa* (1967); A.J. HANNA, *The Beginnings of Nyasaland and North-Eastern Rhodesia, 1859–95* (1956). (*The expansion of white settlement south of the Limpopo*): J.S. GALBRAITH, *Reluctant Empire: British Policy on the South African Frontier, 1834–1854* (1963); W.M. MACMILLAN, *Bantu, Boer, and Briton*, 2nd rev. ed. (1963), and *The Cape Colour Question: A Historical Survey* (1927); I.E. EDWARDS, *Towards Emancipation: A Study in South African Slavery* (1942); E.A. WALKER, *The Great Trek* (1934), a readable and generally reliable account; J.A.I. AGAR-HAMILTON, *The Native Policy of the Voortrekkers: An Essay on the History of the Interior of South Africa, 1836–1858* (1928), and *The Road to the North: South Africa, 1852–1886* (1937); C.W. DE KIEWIET, *British Colonial Policy and the South African Republics, 1848–1872* (1929); EDGAR H. BROOKES and COLIN DE B. WEBB, *A History of Natal* (1965); D.R. MORRIS, *The Washing of the Spears: A History of the Rise of the Zulu Nation Under Shaka and Its Fall in the Zulu War of 1879* (1965), most valuable for its account of the Anglo-Zulu War but also contains a great deal of valuable information on Zululand and Natal in the 19th century; SHULA MARKS, *Reluctant Rebellion: The 1906–08 Disturbances in Natal* (1970). (*The era of mineral discoveries and confederation in South Africa*): C.W. DE KIEWIET, *The Imperial Factor in South Africa* (1937, reprinted 1965); C.F. GOODFELLOW, *Great Britain and South African Confederation, 1870–1881* (1966); F.A. VAN JAARSFELD, *Die ontwaking van die Afrikaanse nasionale bewussyn, 1868–1881*, 2nd ed. (1959; Eng. trans., *The Awakening of Afrikaner Nationalism 1868–1881*, 1961). (*The scramble for southern Africa and the African response*): D.M. SCHREUDER, *Gladstone and Kruger: Liberal Government and Colonial "Home Rule," 1880–85* (1969); J. FLINT, *Cecil John Rhodes* (1976); J.S. GALBRAITH, *Crown and Charter: The Early Years of the British South Africa Company* (1974); J.S. MARAIS, *The Fall of Kruger's Republic* (1961); D.W. KRUGER, *Paul Kruger*, 2 vol. (1961–63); CECIL HEADLAM (ed.), *The Milner Papers*, 2 vol. (1931–33); J.A. HOBSON, *The War in South Africa: Its Causes and Effects* (1900); L.S. AMERY (ed.), *The Times History of the War in South Africa, 1899–1902*, 7 vol. (1900–09); G.H.L. LE MAY, *British Supremacy in South Africa, 1899–1907* (1965). (*The protectorates and Central Africa*): ANTHONY SILLERY, *Founding a Protectorate: History of Bechuanaland, 1885–1895* (1965); STAFFORD GLASS, *The Matabele War* (1968); PHILIP MASON, *The Birth of a Dilemma: The Conquest and Settlement of Rhodesia* (1958); T.O. RANGER, *Revolt in Southern Rhodesia, 1896–97* (1967); R.I. ROTBERG, *Christian Missionaries and the Creation of Northern Rhodesia, 1880–1924* (1965); G.L. CAPLAN, *The Elites of Barotseland, 1878–1969: A Political History of Zambia's Western Province* (1970). (*Angola and Mozambique in the 19th century*): ERIC V. AXELSON, *Portugal and the Scramble for Africa, 1875–1891* (1967); R.J. HAMMOND, *Portugal and Africa 1815–1910* (1966); HENRY W. NEVINSON, *A Modern Slavery* (1906); P.R. WARHURST, *Anglo-Portuguese Relations in South-Central Africa, 1890–1900* (1962). (*The Germans in South West Africa*): W.O. AYDELOTTE, *Bismarck and British Colonial Policy: The Problem of South West Africa, 1883–1885* (1937); HELMUT BLEY, *Kolonialherrschaft und Sozialstruktur in Deutsch-Südwestafrika* (1968; Eng. trans., *South-West Africa Under German Rule, 1894–1914*, 1971); H. DRECHSLER, *Südwestafrika unter deutscher Kolonialherrschaft: Der Kampf der Herero und Nama gegen den deutschen Imperialismus, 1884–1915* (1966).

Southern Africa, c. 1900 to the present: (*South Africa and South West Africa*): GWENDOLYN M. CARTER, *The Politics of Inequality: South Africa Since 1948*, 2nd rev. ed. (1959); L.M. THOMPSON, *Politics in the Republic of South Africa* (1966); D.W. KRUGER, *The Making of a Nation: A History of the Union of South Africa, 1910–1961* (1969), a short, general political account from the white point of view; D.J.N. DENOON, *The Grand Illusion: The Failure of Milner's Reconstruction Policy in the Transvaal, 1900–1905* (1973); NICHOLAS MANSERGH, *South Africa 1906–1961: The Price of Magnanimity* (1962); L.M. THOMPSON, *The Unification of South Africa, 1902–1910* (1960); F.A. JOHNSTONE, *Class, Race and Gold* (1976); T.D. MOODIE, *The Rise of Afrikanerdom* (1975); F.A. JOHNSTONE, "White Prosperity and White Supremacy in South Africa Today," *African Affairs*, 69:124–140 (1970), an important article, challenging the assumption that increased prosperity will end the colour bar; W.H. VATCHER, *White Laager: The Rise of Afrikaner Nationalism* (1965); PETER CALVOCORESSI, *South Africa and World Opinion* (1961); DENNIS AUSTIN, *Britain and South Africa* (1966); EDWARD ROUX, *Time Longer Than Rope: A History of the Black Man's Struggle for Freedom in South Africa*, 2nd ed. (1964); MARY BENSON, *South Africa: The Struggle for a Birthright* (1966); PETER WALSHE, *The Rise of African Nationalism in South Africa: The African National Congress, 1912–1952* (1970). M. LEGASSICK, "Legislation, Ideology and Economy in Post-1948 South Africa," *J. of Southern African Studies*, vol. 1, no. 1 (1974); C. LEGUM, *The Secret Diplomacy of Detente and South Africa at the Crossroads* (1975); K.W. GRUNDY, *Confrontation and Accommodation in Southern Africa: The Limits of Independence* (1973); H. ADAM, *Modernizing Racial Domination: South Africa's Political Dynamics* (1971). (*The High Commission Territories: Lesotho, Botswana, and Swaziland*): LORD M. HAILEY, *Native Administration in the British African Territories*, pt. 5, *The High Commission Territories* (1953) and *The Republic of South Africa and the High Commission Territories* (1963); *Basutoland, Bechuanaland Protectorate and Swaziland: Report of an Economic Survey Mission* (1960); J.F. HOLLEMAN (ed.), *Experiment in Swaziland* (1964). (*Central Africa: Rhodesia, Zambia, and Malawi*): RICHARD GRAY, *The Two Nations* (1960), still one of the best analyses of Southern Rhodesia in the 20th century; PATRICK KEATLEY, *The Politics of Partnership* (1963); COLIN LEYS, *European Politics in Southern Rhodesia* (1959); C. VAN ONSELEN, *Chibaro: African Mine Labour in Southern Rhodesia, 1900–1933* (1976); G. ARRIGHI, *The Political Economy of Rhodesia* (1967); GODFREY WILSON, *An Essay on the Economics of Detribalization in Northern Rhodesia*, 2 pt. (1941–42); E. BERGER, *Labour, Race and Colonial Rule: The Copperbelt from 1924 to Independence* (1974); MONTAGUE YUDELMAN, *Africans on the Land: Economic Problems of African Agricultural Development in Southern, Central, and East Africa, with Special Reference to Southern Rhodesia* (1964); A.I. RICHARDS, *Land, Labour and Diet in Northern Rhodesia* (1939); WILLIAM WATSON, *Tribal Cohesion in a Money Economy* (1958); T.O. RANGER, *The African Voice in Southern Rhodesia, 1898–1930* (1970); R.I. ROTBERG, *The Rise of Nationalism in Central Africa: The Making of Malawi and Zambia, 1873–1964* (1965); GEORGE SHEPPERSON and THOMAS PRICE, *Independent African: John Chilembwe and the Origins, Setting and Significance of the Nyasaland Native Rising of 1915* (1958); COLIN LEYS and CRANFORD PRATT, *A New Deal in Central Africa* (1960); PHILIP MASON, *Year of Decision: Rhodesia and Nyasaland in 1960* (1960); JAMES P. BARBER, *Rhodesia: The Road to Rebellion* (1967); DAVID C. MULFORD, *Zambia: The Politics of Independence, 1957–1964* (1967); RICHARD S. HALL, *The High Price of Principles: Kaunda and the White South* (1969). (*Angola and Mozambique*): DAVID M. ABSHIRE and MICHAEL A. SAMUELS (eds.), *Portuguese Africa* (1969); ANDERS EHNMARK and PER WASTBERG, *Angola and Mozambique* (1963); R.J. HAMMOND, *Portugal's African Problem: Some Economic Facets* (1962); INSTITUTE OF RACE RELATIONS, *Angola: A Symposium* (1962); JOHN MARCUM, *The Angolan Revolution*, vol. 1, *The Anatomy of an Explosion, 1950–62* (1969); T.O. RANGER, "Revolt in Portuguese East Africa," in KENNETH KIRKWOOD (ed.), *St. Anthony's Papers No. 15* (1963); EDUARDO MONDLANE, *The Struggle for Mozambique* (1969); J.C. MILLER, "The Politics of Decolonization in Portuguese Africa," *African Affairs*, vol. 74 (1975).

(Sh.M.)

South West Africa

South West Africa (Namibia) lies on the Atlantic coast of Africa between latitudes 17° and 29° south. In the north it is bounded by the Kunene and Okavango rivers, which separate it from Angola. Here a narrow strip of land—the Caprivi Strip (Caprivi Zipfel)—extends eastward for nearly 300 miles, bringing the territory to the borders of Zambia. In the south the Orange River marks the boundary with South Africa. In the east the territory adjoins both Botswana and South Africa. Extending inland some 600 miles (965 kilometres) in the north, without the Caprivi Strip, and about 220 miles in the south, its total area is estimated at 318,000 square miles (824,000 square kilometres).

For articles on associated physical features, see KALAHARI; NAMIB DESERT; OKAVANGO RIVER; and ORANGE RIVER. For history, in addition to that given below, see SOUTHERN AFRICA, HISTORY OF.

HISTORY

As a geographically defined area, South West Africa came into being during the last two decades of the 19th century. It was the result of agreements and boundary settlements between Britain, Portugal, and Germany following Germany's decision in 1884 to acquire a colony in southwestern Africa. One important enclave, however, did not become part of the German acquisition. The largest port, Walvis Bay, and a coastal strip comprising in all an enclave of 434 square miles (1,124 square kilometres) had already been annexed by Britain in 1878 and became a part of Cape Colony in 1884. At the time of the formation of the Union of South Africa in 1910, it became South African property.

Apart from this enclave, South West Africa remained under German rule until shortly after the outbreak of World War I, when it was successfully invaded by troops from South Africa. The colony was then placed under South African military administration until January 1, 1921, when it became a Class C mandated territory, administered by South Africa on behalf of the League of Nations.

The Class C mandates explicitly authorized the administering governments to exercise full powers over the territories as integral parts of their own states, subject to the international supervisory authority of the Council of the League of Nations. Upon the formal dissolution of the League in April 1946, South Africa, after unsuccessfully seeking United Nations approval for the incorporation of the territory, refused to place it under the UN trusteeship system. Since that time the status and future of South West Africa have been in dispute between South Africa and the United Nations. The position of successive South African governments has essentially been that, with the demise of the League and its supervisory machinery, all their international obligations in respect of the mandate save one have lapsed. The acknowledged exception is the clause providing that the territory should be administered as a "sacred trust of civilization," and establishing an obligation to rule in accordance with the spirit of the mandate. The initial position of the United Nations was that the mandate was still in force, that the supervisory authority of the League was to be exercised by the United Nations, and that, both by its refusal to sign a trusteeship agreement for South West Africa and by the conduct of its administration, South Africa had violated the mandate.

Dispute between South Africa and the UN

On July 18, 1966, the International Court of Justice handed down a judgment in the South West Africa Cases (1960–66) but without ruling on these issues. On October 27, 1966, the UN General Assembly adopted a resolution declaring that the mandate was terminated and that the UN assumed direct responsibility for the territory until it attained independence. In May 1967, to implement this decision, the assembly created an 11-member Council for South West Africa; on June 12, 1968, it renamed the territory Namibia.

The South African government condemned these actions as illegal and impractical. Thereafter the UN Security Council repeatedly called for South Africa's compliance with the assembly's decisions, but without result. Demands by South West African and other African nationalists for the independence of the territory led South Africa to convene the South West Africa Constitutional Convention, the first session of which was held in Windhoek in September 1975. By early 1977 it had produced a commitment that full independence should be achieved by December 31, 1978, and in January 1977 a draft constitution providing for an interim government until independence was published.

THE LANDSCAPE

Physiography. South West Africa may be described as a highland between two desert areas. Three main topographical regions may be distinguished—the Namib coastal belt, the Central Highland of the Central Plateau, and the northern and eastern parts of the plateau known as the Kalahari Belt. The Namib Desert extends from the Atlantic coast in the west until, at an altitude of 3,000–3,500 feet, it meets the foot of the plateau edge, or Great Escarpment, in the east. A desolate strip of sand desert, high shifting dunes, salt pans, and bare rock outcrops, it is bounded by the Kunene River in the north and by the Orange River in the south; for much of its length it is less than 100 miles wide. To the north it becomes more mountainous, with spectacular granite masses rising abruptly from the plain. The Brandberg, with an altitude of 8,461 feet (2,579 metres), is the highest point in the territory. The Namib constitutes about 15 percent of the total area of the territory.

East of the Great Escarpment lies the Central Plateau, which has two major surface divisions—the plateau hardveld (or rocky plain), known as the Central Highland, and the Kalahari sandveld (desert). The Central Highland extends from the southern to the northern frontier and varies in altitude from 3,200 to 6,500 feet. Its diverse surface characteristics include broken veld, rugged mountains, sand-filled valleys, gentle plains, and isolated peaks. A mountainous central part in the vicinity of Windhoek divides the Central Highland naturally into Damaraland to the north and Great Namaland to the south. Comprising such ranges as the Khomas Highland, the Auasberge (where the Moltkeblick at 8,153 feet is the second highest peak), and the Eros and Onyati to the east, it forms the territory's highest mountain tract, constituting a central watershed from which the various headwaters drain north, south, east, and west.

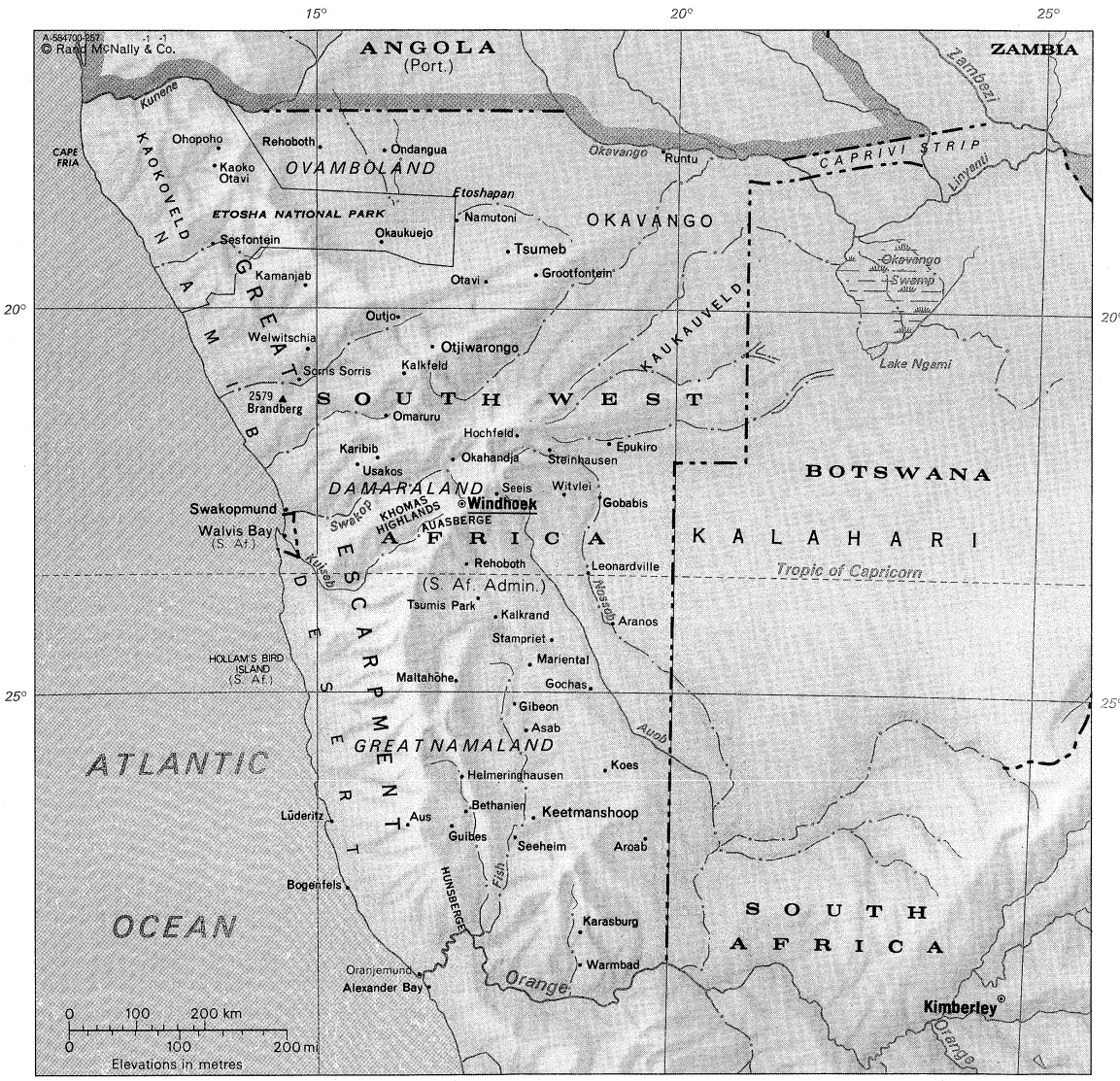
The Kalahari sandveld extends along the whole of the eastern frontier to the Okavango River. Its thick layers of terrestrial sands and limestone vary in depth from more than 1,000 feet to a mere covering at the hardveld edge. It varies in width from 50 miles in the south, where it covers most of the Nossob-Auob Basin, to 250 miles in the north, where it comprises part of the Etosha Basin and the middle Okavango Basin. It has a level, sandy surface, which may rise to as much as 300 feet. In the south it is crossed by long sand dunes and grass and is covered with thorn trees.

The scarcity of water

The water shortage, or the uncertainty of obtaining water, touches all aspects of life and of development. Of the larger rivers only the Orange, Kunene, and Okavango are perennial, and they constitute boundaries of the territory. All rise outside South West Africa and—as with the Okavango—the Cuando and the Zambezi in the north flow only through small sections of the territory. The potential of these rivers for irrigation, however vital, is thus limited to the southern and northern margins. The remaining rivers, whether the large westward-flowing deep-channel streams such as the Swakop and the Kuiseb or the eastward-flowing shallow watercourses (*omurambas*) such as the Epukiro and the Omuramba Omatako, usually run only intermittently. They rarely contain surface water, except after rains or floods, although they constitute storage places for groundwater. Surface water collects mainly in shallow clay or limestone pans, especially in the sandveld region—the most famous of which is the giant Etosha Pan measuring some 68 miles by 31—as well as in small lakes. There are also a number of springs.

The general scarcity of water thus makes the well and the borehole the basis for land utilization. The conservation of floodwater through the construction of dams is, however, also of importance. It is estimated that there are some 10,000 dams, with a total storage capacity of 670,000,000 cubic metres. The largest of these is the Hardap Dam on the southward flowing Fish River, just north of Mariental.

Climate. The climate is generally hot and dry. The territory is situated in the tropical high pressure belt of the Southern Hemisphere, located between latitudes 25° and 35° south, which governs the wind system of the plateau. The coast region, however, experiences the air circulation that occurs as a result of the interaction between the eastern sector of the South Atlantic High pressure system, and the upwelling of the cold Benguela Cur-



SOUTH WEST AFRICA

MAP INDEX

Cities and towns

Aranos.....24·09s 19·09e
Aroab.....26·47s 19·40e
Asab.....25·29s 17·59e
Aus.....26·40s 16·15e
Bethanien.....26·32s 17·11e
Bogenfels.....27·23s 15·22e
Epukiro.....21·41s 19·08e
Gibeon.....25·09s 17·43e
Gobabis.....22·30s 18·58e
Gochas.....24·55s 18·55e
Grootfontein.....19·32s 18·05e
Guibes.....26·41s 16·42e
Helmeringhausen.....25·54s 16·57e
Hochfeld.....21·28s 17·58e
Kalkfeld.....20·53s 16·11e
Kalkrand.....24·03s 17·33e
Kamanjab.....19·35s 14·51e
Kaoko Otavi.....18·14s 13·40e
Karasburg.....28·00s 18·43e
Karibib.....21·58s 15·51e
Keetmanshoop..26·36s 18·08e
Koes.....25·59s 19·08e
Leonardville.....23·29s 18·49e
Lüderitz.....26·38s 15·10e
Maltahöhe.....24·50s 17·00e
Mariental.....24·36s 17·59e
Namutoni.....18·49s 16·55e
Ohopoho.....18·03s 13·45e
Okahandja.....21·59s 16·58e
Okaukuejo.....19·10s 15·54e
Omaruru.....21·28s 15·56e
Ondangua.....17·55s 16·00e
Oranjemund.....28·38s 16·24e
Otavi.....19·39s 17·20e
Otjiwarongo.....20·29s 16·36e
Outjo.....20·08s 16·08e
Rehoboth.....17·52s 15·05e
Rehoboth.....23·18s 17·03e
Runtu.....17·52s 19·43e
Seeheim.....26·50s 17·45e
Seeis.....22·29s 17·39e
Sesfontein.....19·07s 13·39e
Sorris Sorris.....20·57s 14·50e
Stampriet.....24·20s 18·28e
Steinhausen.....21·49s 18·20e
Swakopmund.....22·41s 14·34e
Tsumeb.....19·13s 17·42e
Tsumis Park.....23·43s 17·28e
Usakos.....22·01s 15·32e
Walvis Bay (S. Af.).....22·59s 14·31e
Warmbad.....28·29s 18·41e
Welwitschia.....20·21s 14·57e
Windhoek.....22·34s 17·06e
Witvlei.....22·23s 18·32e

Physical features and points of interest

Atlantic Ocean..25·00s 13·00e
Auasberge, *mountains*.....22·45s 17·22e
Auob, *river*.....25·43s 20·00e
Brandberg, *mountain*.....21·09s 14·33e
Caprivi Strip, *physical region*.17·60s 23·00e
Damaraland, *historic region*..22·34s 17·06e
Etosha National Park.....18·45s 15·00e
Etoshapan, *salt flat*.....18·45s 16·15e
Fish, *river*.....28·07s 17·45e
Fria, *cape*.....18·30s 12·01e
Great Escarpment...23·00s 15·00e
Great Namaland, *historic region*..25·00s 17·00e
Hollam's Bird Island.....24·45s 14·34e
Hunsberge, *mountains*.....27·45s 17·12e
Kaokoveld, *plateau*.....21·00s 14·20e
Kaukauveld, *plateau*.....20·00s 20·30e
Khomas Highlands, *mountains*.....22·30s 16·30e
Kuiseb, *river*....22·59s 14·31e
Kunene, *river*....17·20s 11·50e
Linyanti, *river*....17·50s 25·05e
Namib Desert....23·00s 15·00e
Nossob, *river*.....24·52s 20·00e
Okavango, *physical region*.19·00s 19·00e
Okavango, *river*..18·17s 21·45e
Orange, *river*.....28·41s 16·28e
Ovamboland, *historic region*..17·45s 16·30e
Swakop, *river*...22·38s 14·36e

rent offshore. The mean temperature is, as a result, lower than the latitude might suggest. The temperature varies from 66° to 72° F in the Central Highland to about 60° F on the coast. During the summer months it may reach 100° F in the north and 120° F in the Namib and the south. In July, which is usually the coldest month of the year, frost may occur, although only lightly in the north. In the south, however, temperatures may drop to 20° F or lower.

Rainfall in the territory increases from south to north. In the arid Namib and the Orange River valley, the mean annual rainfall is less than four inches; in the north, and along the Caprivi Strip, the rainfall increases to between 24 and 26 inches. The mean annual rainfall for the territory as a whole is slightly above ten inches. Most rainfall occurs in the summer months (December to February). Only the extreme southwest receives as much as half an inch of rain in winter. The territory is occasionally subject not only to periods of flood, when the precipitation may be more than double the normal expectation, but also to periods of drought. The rains are the most reliable in the wet northeast, and the most variable in the dry south.

Vegetation and animal life. There are two categories of vegetation. In the west is found the desert and semi-desert type, with its desert grasses and shrub; in the east the woodland savanna type, with its grasses, bushes, and trees. In the Namib Desert region there is little vegetation except for succulents and the narra, the nut of a gourd-like plant. Further inland, and in the south, there are varieties of grasses, of succulents, such as the large aloe, of euphorbias, and of dwarf trees such as those belonging to the *Acacia* and *Boscia* genera. From south to north the veld changes to grass country, increasingly studded with thorn trees, of which most are acacia. Many others are also found, however, such as raffirboom, witgat, bush willow, and wild olive. Further north the annual grasses become more abundant, the "bush" vegetation and acacia scrub grows thicker, and the variety of thorn trees increases. This pattern is not, however, uniform. The areas to the east and west of the palm-tree belt in southern Ovamboland are, for example, sparsely wooded. In the north the growing density of the trees, the appearance of the palm-tree belt, and, to the west, the baobab tree, all are signs marking the proximity of the well-wooded northern forest region, which stretches from the Kaokoveld across Ovamboland in the west to the greater part of the Okavango and Caprivi Strip areas in the east. It is here that most of the indigenous timber trees of the territory are found—Rhodesian teak and mahogany, mopane, marula, manketti, and kiaat among others. In the west, however, as one approaches the dry steppe of the Kaokoveld, the trees and the bushes thin out, the grass spreads more abundantly, and the mopane trees and bushes diminish in size.

The northern forest region

Big game is relatively abundant and is found in many parts of the territory. The wildlife includes lion, elephant, rhinoceros, giraffe, leopard, buffalo, eland, oryx, hartebeest, kudu, and zebra, as well as many smaller species of game and of predatory animals. The most renowned of the game reserves is the Etosha National Park, which covers an area of 15,004 square miles and is one of the largest game parks in the world.

The landscape under human settlement. The ecological relationship between man and environment is a complex one. Among the elements affecting it are frequent drought, the low protein and low phosphorous content of grasses, and the high salt content and humus deficiency of the soil. The best grazing and the most groundwater are found in the plateau hardveld region, most of which—some 140,000 square miles—is now owned by European farmers. This region supports a white rural population of 19,000. The white-owned farm area amounts to 96,000,000 acres. Africans living in the Reserves occupy 54,000,000 (1960). The Reserves are now officially termed Home Areas or Homelands—of which there are 17 in the Southern Sector of the territory and four (in which there is no European settlement) in the Northern Sector. In 1960 the total estimated Home Areas population was 324,235 (of which only 38,648 persons were in the Southern Sector), with an average land area per capita of 168 acres.

The factors affecting the pattern of land settlement have primarily been the following: firstly, the uninhabitable character of the large desert regions; secondly, the original areas chosen for settlement; thirdly, changes caused by the German wars against the Herero and Nama; fourthly, the encouragement of white settlement, mostly Afrikaner, by South Africa since the mandate; fifthly, the setting aside of the Northern Sector for the Africans living there, and the allocation of additional scattered Reserves for Africans in the Southern Sector, when more are planned. The Southern Sector may be divided into urban areas, rural areas (comprising mostly surveyed white farms), and the Home Areas. According to the 1960 census, there were 59,100 Africans in the urban areas and 68,300 in the rural areas. The remaining 53,700 whites were in the urban areas, making a total population of approximately 526,000.

According to the *Report of the Commission of Enquiry into South West African Affairs* (1962–63)—the Odendaal Commission—there are 37 towns and township areas in the territory, all of which are situated in the Southern Sector. The largest is Windhoek, the administrative centre and principal town, with a population of 66,000 (1967 estimate). No other town has a population of half this size. The chief port for the territory is Walvis Bay, which has the only natural deep harbour. The other port is Lüderitz, but it is shallow and is used mainly by small coastal craft or lighters. Tsumeb is the principal copper- and lead-mining centre, some 250 miles to the north of Windhoek.

PEOPLE AND POPULATION

The total population of South West Africa was 746,000 at the 1970 census, giving the territory as a whole an average density of under 2.5 persons per square mile. Its composition was as follows: Europeans 91,000; Damara (Bergdama) 65,000; Herero 49,000; Nama 33,000; Coloureds 28,000; Rehoboth Basters 16,500; Bushmen 22,000; Tswana and others 18,500; Ovambo 342,500; Okavango 50,000; East Caprivians 25,000; Kaokovelders 6,500. The last four groups are settled in the Northern Sector, as also are 20 percent of the Bushmen. The remainder of the population is settled in the Southern Sector.

Of the white population approximately 67 percent are Afrikaans speaking, 23 percent German speaking, and 10 percent English speaking. The nonwhite population may be divided into two main types, the Southwestern Bantu and the Khoisan types. The Bushmen and the Nama belong to the Khoisan peoples and their language contains the characteristic Khoisan clicks. The Damara, a black African people of unknown origin, have adopted the Khoisan language. The Hereros, who speak their own language, belong to the Bantu language family, as do the Ovambo, the Okavango, and the Kaokovelders. The latter are closely related to the Hereros in origin, language, and culture. The East Caprivians also belong to the Bantu-speaking peoples but are related to the Lozi and Makololo groups of Barotseland (western Zambia), rather than to the Bantu of South West Africa. The Tswana and other groups include some 4,000 people from Angola. They are mostly Bantu speaking but are not related to the other ethnic groups in the territory. Apart from the established group of Tswana, who are related to the Tswana of Botswana, most of these peoples are dispersed throughout the rural and urban areas of the territory. The Basters of Rehoboth (or Rehobothers) are of mixed origin, being descended in part from unions between Afrikaner trekkers and Nama women. They moved from the Cape Colony in the 1860s and settled at Rehoboth. Their language, like that of the Coloured people, is predominantly Afrikaans, and their way of life is Western. The Coloured people, more recent immigrants, were originally a part of the Cape Coloured population of South Africa. According to official statistics all population groups have shown sustained growth. Figures for the Europeans and Col-

Linguistic groups

South West Africa, Area and Population

	area		population	
	sq mi	sq km	1960 census	1969 estimate
Magisterial districts				
Bethanien	7,125	18,454	4,000	...
Gibeon	18,285	47,358	16,000	...
Gobabis	36,966	95,742	26,000	...
Grootfontein	26,358	68,267	19,000	...
Karibib	7,252	18,783	9,000	...
Keetmanshoop	19,651	50,896	20,000	...
Lüderitz	28,734	74,421	12,000	...
Maltahohe	8,067	20,894	4,000	...
Okahandja	6,236	16,151	10,000	...
Omaruru	6,448	16,700	9,000	...
Otjiwarongo	9,377	24,286	18,000	...
Outjo	29,297	75,879	14,000	...
Rehoboth	12,973	33,600	18,000	...
Swakopmund	10,907	28,249	6,000	...
Tsumeb	7,388	19,135	13,000	...
Walvis Bay*	...	...	...	...
Warmbad	14,455	37,438	10,000	...
Windhoek	8,678	22,476	47,000	...
Other				
Eastern Caprivi Strip†	4,453	11,533	16,000	...
Tribal magisterial districts†				
Kaokoveld	22,650	58,664	10,000	...
Okavango	16,385	42,437	28,000	...
Ovamboland	16,202	41,963	204,000	...
Walvis Bay‡	374	969	13,000	...
Total South West Africa§	318,261	824,296‖	526,000	615,000

*In addition to the Walvis Bay enclave there is a magisterial district of Walvis Bay for which no separate area and population figures are available. †Administered by South Africa. ‡An integral part of South Africa, but administered as if it were a part of South West Africa. §Including Walvis Bay enclave. ‖Converted area figures do not add to total given because of rounding.
Source: Official government figures; UN.

oureds have, however, been increased by recent immigration and are not, therefore, comparable to those for the other groups in terms of growth.

The distribution of population in urban, rural, and Home Areas (1960) for the Southern Sector was as follows: towns 47%; rural areas about 37%; Home Areas 16%. The most highly urbanized groups were the whites (73%), Coloureds (64%), and Damara (42%). Of the Bushmen, 86% are found in the white rural areas, as are 56% of the Nama, 46% of the Damara, 30% of the Herero, and 29% of the Coloureds. Of the Basters and Hereros, 79% and 44%, respectively, live in their Home Areas.

THE ECONOMY

South West Africa's economy is determined, in part, by the harsh and arid environment. Its three key industries —farming, mining, and fishing—together account for more than 60 percent of the gross domestic product. Its internal market is small; traditional subsistence farming persists over large areas, and its economic dependence upon South Africa is virtually complete. Apart from loans and subsidies, South Africa supplies a little less than half the capital investment, as well as 90 percent of the total imports, and receives over half of the exports. The estimated contribution of agricultural production to the gross domestic product (GDP) is approximately 17 percent. Cattle account for about 60 percent and sheep (mostly Karakul) for about 35 percent of the total gross output of commercial agriculture. The territory obtains its fruits and vegetables, as well as the bulk of its grains for the Southern Sector, from South Africa. The South African rand is the unit of currency. In bad seasons, the Northern Sector, too, must rely on South Africa for these commodities. Mining contributes about 45 percent of the GDP, with diamonds accounting for about 60 percent of mineral sales, and base metals for about 40 percent. Manufacturing is severely restricted by high costs, and by a low aggregate demand from a small, poor, and widely scattered population. For the foreseeable future, despite progress already made, and despite new water and power developments—such as the Ovamboland irrigation and the Kunene hydroelectric schemes—planned by the South African government, the territory is likely to continue to remain dependent upon a few primary products. Mineral products account for over half the total value of exports, fishery products for approximately a quarter, and agricultural products for most of the remainder.

TRANSPORTATION

The transportation network primarily serves the Southern Sector. Transport development in recent years has been aided by capital and loan funds from the South African government. In 1953 there was not a single mile of tarred road. By 1966, however, over 800 miles out of the 20,000-odd miles of principal roads had been tarred. The tarred roads link Windhoek to Tsumeb in the north, to Keetmanshoop in the south, and to Swakopmund and Walvis Bay in the west—all of which are in the Southern Sector. All but 200 or 300 of the 40,000 vehicles in the territory were registered in the Southern Sector. A road construction program for the Northern Sector has been announced but is still in a rudimentary phase.

The railroad system

The territory's railway system has been amalgamated with that of the South African Railways and Harbours Administration, which subsidizes it. The track and sidings are all in the Southern Sector. The railroad runs from South Africa to Windhoek, and then north to Tsumeb. This north–south central spine is linked, via east–west lines, with the two ports of Walvis Bay and Lüderitz. A line to the east links Windhoek with Gobabis.

The high cost of road building and the distances involved add to the importance of air services. Windhoek has an international airport, and there are a number of other airfields and landing strips. There are five return flights from Johannesburg to Windhoek and four from Cape Town each week—service that since 1948 has been subsidized by the South African Railways and Harbours Administration. South West Airways, a private airline, maintains internal services and is also subsidized.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of the government. If the juridical status of the territory remains in dispute, the structure of its de facto administration is quite clear. The territory has not been incorporated into the Republic of South Africa but is so closely connected with it as virtually to constitute a fifth province. According to South African law, an administrator is appointed by the South African government. He is assisted by a white Legislative Assembly

chosen by the white electorate. The Assembly has executive and legislative powers in all matters not expressly reserved to the South African government, and its ordinances have the effect of law, so long as they are not "repugnant" (in the legal sense) to acts of the South African Parliament. Subject to the approval of that Parliament, the state president of the republic may issue proclamations in regard to the reserved matters—which include defense, police, foreign affairs, native affairs, transport, information, immigration, customs and excise, and banking and audit. The South African minister of Bantu Administration and Development of Bantu Education is responsible for "Native Affairs"—*i.e.*, for the administration of the majority of the nonwhite peoples living in the territory. He is not, however, responsible for the administration of the Basters or of the Coloured population, who fall under the jurisdiction of South Africa's minister of Labour and Coloured Affairs.

The South West African whites, for their part, elect six members to the white South African Parliament and have four seats in its Senate, two chosen by the combined vote of the 6 members of Parliament and the 18 members of the Legislative Assembly and 2 nominated by the South African State President. The majority of the white voters supports the South African National Party and a minority supports the United National South West Party, which is allied to the official opposition in South Africa, the United Party. Government policy and administration is worked out, with the respective differences having been considered, within the framework of "separate development" or apartheid.

The military. The responsibility for military defense is assumed by the South African government and is undertaken by the South African Defense Force. The territory's police force has been amalgamated with that of South Africa since 1939, and some three-quarters of the costs are borne by the Republic. It musters about 400 whites, about 200 Africans, and a few Coloureds. It does not operate in the Northern Sector except in unusual circumstances.

Justice. For the administration of justice the territory is divided into 21 magisterial districts. The law applied is either customary law (where it does not conflict with principles of public policy or natural justice) or Roman-Dutch law. The High Court is a division of the Supreme Court of South Africa, and sits in Windhoek; appeals are heard by the Appellate Division of the South African Supreme Court at Bloemfontein. In the Northern Sector, except with regard to the most serious crimes, full civil and criminal jurisdiction is given to the tribal chiefs and to councils of headmen, to be exercised in accordance with customary law. The system includes provision for District Courts, for the Tribal Court, and for the Court of Appeal, which is attended by the government's Bantu Affairs Commissioner in an advisory capacity. In the Southern Sector the Bantu Affairs commissioner's courts have concurrent jurisdiction with the magistrates courts to hear disputes between members of the African communities in accordance with African law and custom. In such cases assessors may be appointed from those communities. The chiefs and headmen are responsible to the Bantu Affairs commissioner, or to the district magistrate, for the maintenance of law and order in their areas. They are paid by the Department of Bantu Administration.

The "pass laws"

Racial policies. The system of "pass laws," so important a feature of the Republic's administration of its racial policies, is in force in the Southern Sector. With the exception of those included in certain exempted categories of office or profession, all African males over 18 years require permits or passes to move from one magisterial district to another, to visit urban areas, to leave or re-enter a Home Area, to leave the Northern Sector, or to purchase a railway ticket. They are also required to carry such other identification documents as the law may demand. Permits enabling Africans to seek work in towns are not issued to those who are not already qualified—by birth, residence, or long employment service—unless there is a temporary shortage of labour.

Africans are not required to pay general—or poll—taxes and are not subject to income, personal, or land taxes. The Coloured people and the Rehobothers, if they fall within the specified income brackets, pay the same income and personal taxes as the whites.

Education. Education is segregated and is compulsory only for white pupils, who number about 20,000. Most of the 66,000 African pupils enrolled are in the substandard or primary grades. There are about 10,000 Coloured children enrolled in school. There is no university in the territory; South West Africans who obtain university education usually do so in South Africa.

PROBLEMS AND PROSPECTS

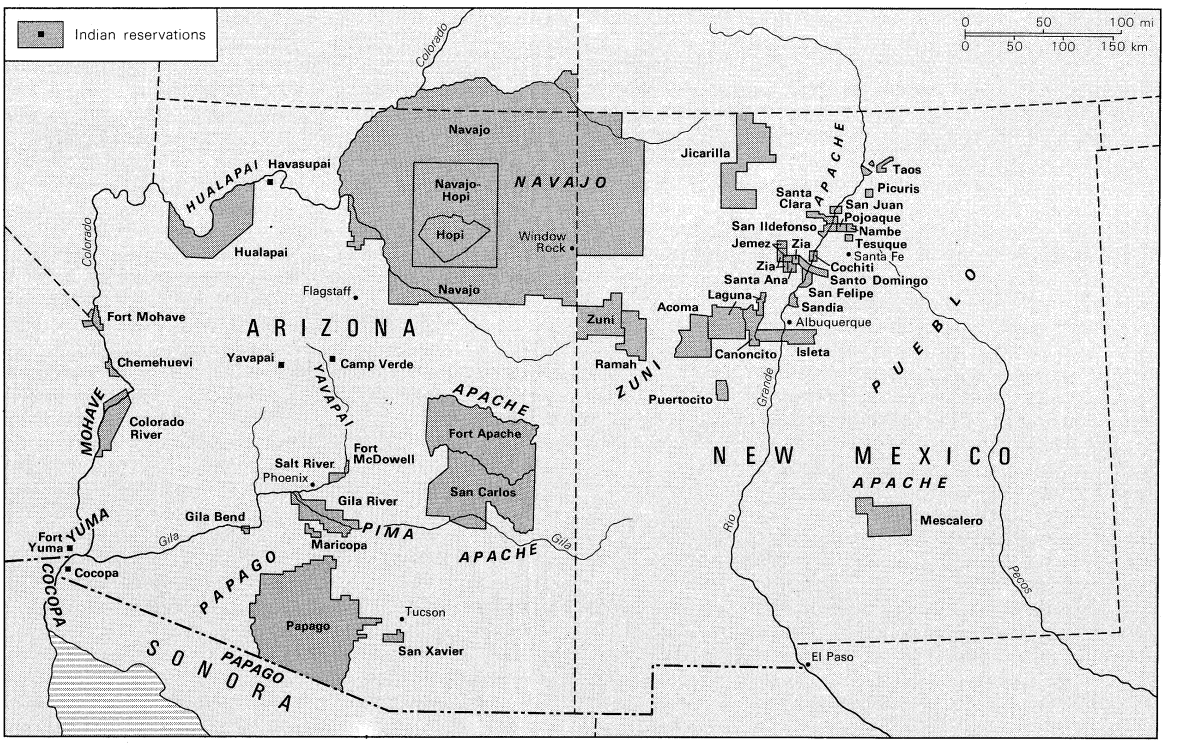
The political activities of the Africans are not always easy to determine in a context in which the South African government and the South West African whites hold power and show themselves determined to maintain it. Radical reformers of all kinds are silenced, and apartheid laws are rigorously enforced. The South African government has, however, given the territory a measure of internal stability and has been responsible for economic and financial benefits for South West Africa in which—without threatening the living standards of whites—nonwhites have had some share, even if a disproportionate one. The disputes between the United Nations and the Republic of South Africa over the policies of apartheid, however, as well as over the status and future of the territory, have served as a focus for African frustration over the continued political, economic, and social dominance of the white minority. The mood of discontent increased awareness of old grievances. The Herero people, for example, still claim lands that Germans had taken from them after the 1903 rebellion; the Ovambos dispute some boundary settlements. Leading figures from the Herero, Damara, Nama, and Ovambo people, as well as from the Rehobothers, have, over the years, petitioned the United Nations to intervene.

Various political leaders inside the territory, as well as exiles abroad, have sought to unite the various opposition elements into an intertribal organization but as yet without success. At least four African nationalist organizations are in existence—the South West Africa People's Organisation, the South West Africa National Union, the South West Africa National Liberation Front (which the first two formed), and the South West Africa National United Front. Despite this, internal opposition remains fragmented and overawed. Such acts of violence as have occurred have been isolated and sporadic. The infiltration of armed African "freedom fighters" from the north has not as yet constituted any threat to the government, or any challenge to its security forces.

For some time to come the prospect is that the future of South West Africa is bound up with that of the Republic of South Africa.

BIBLIOGRAPHY. HEINRICH VEDDER, L. FOWIE, and C.H.L. HAHN, *The Native Tribes of South West Africa* (1928, reprinted 1966), with chapters on the Bushmen, Damara, Herero, and Nama; HEINRICH VEDDER, *Das alte Südwestafrika* (1934; Eng. trans., *South West Africa in Early Times*, 1938, reprinted 1966); A.W. HOERNLE, "The Social Organization of the Nama Hottentots of Southwest Africa," *Am. Anthrop.* 27:1–24 (1925); H.R. MACCALMAN and B.J. GROBBELAAR, "Preliminary Report of Two Stoneworking Ovatjimba Groups in the Northern Kaokoveld of South West Africa," in *Cimbebasia*, no. 13 (1965); GORDON LE SUEUR, *Germany's Vanishing Colonies* (1915), concerned with the effects on colonialism of World War I; MARY E. TOWNSEND, *Origins of Modern German Colonialism, 1871–1885* (1921) and *The Rise and Fall of Germany's Colonial Empire, 1884–1918* (1930); WILLIAM O. HENDERSON, *Studies in German Colonial History* (1962), an important recent work; W.W. SCHMOKEL, *Dream of Empire: German Colonialism, 1919–45* (1964), the best recent book on the period between World Wars I and II; JON BRIDGMAN and DAVID E. CLARKE, *German Africa: A Select Annotated Bibliography* (1965); ODENDAAL COMMISSION, *Report on South West Africa* (1964), published by the South African government and therefore reflecting its views; J.H. WELLINGTON, *South West Africa and Its Human Issues* (1967), an authoritative work; INTERNATIONAL COMMISSION OF JURISTS, *Apartheid in South Africa and South West Africa* (1967).

(R.B.Ba.)



Distribution of Southwest American Indians and their reservations and lands.

Southwest American Indians

More than one-third of the rapidly growing population of American Indians live in the southwestern United States, mainly in or bordering Arizona and New Mexico. In this predominantly arid and climatically unstable region, located between the Rocky Mountains and the Mexican Sierra Madre, the Continental Divide separates the watersheds of two great river systems: the Colorado–Gila–San Juan, in the west, and the Rio Grande–Pecos, in the east.

Environment of the area

From the viewpoint of human habitation, the region's main geographic features are the two river systems, cyclical droughts, and arid lands, some low and hot, others high and cold. The hot deserts have sparse and irregularly occurring rainfall. Their long growing season supports a great variety of plant and animal communities adapted to desert conditions—creosote, sage, tarbush, and numerous cactus species, as well as such small, nocturnal, burrowing animals as the kangaroo rat. Along the river flood plains grow cottonwood, willow, mesquite, and sycamore. Basin and range landscape, from about 1,000 to 4,000 feet (300 to 1,200 metres) in elevation, predominates.

Despite its low moisture content, coarse texture, and occasional salty patches, the soil of most of the Southwest is relatively fertile.

The cold semideserts include the Colorado and other plateaus of northern Arizona. The frost-free growing season is relatively short. Much of this plateau area is covered with scrub or with piñon–juniper woodland, where rattlesnakes, rabbits, coyote, bobcat, and mule deer are found. Antelope, American elk, and mountain sheep were once plentiful. Bordering the plateau country are sheer cliffs, deep canyons, and forested mountains. Ponderosa pine, Douglas fir, blue spruce, and alpine meadows shelter weasel, deer mouse, porcupine, squirrel, and chipmunk, as well as the larger animals of the region.

In the past century most of the wild mammals have disappeared from the region. Domesticated species brought to America from Europe by the Spaniards during the 16th and 17th centuries, such as cattle, sheep, goats, horses, and burros have multiplied and destroyed delicate ecological balances, critically accelerating natural erosion processes, especially on the Colorado Plateau and in the basin and range country.

The Indian tribes that have gained a foothold in the Southwest are highly diversified culturally, linguistically, and psychologically. From a comparative viewpoint, however, they fall roughly into four groupings, each of which is characterized by living patterns designed to cope with the practical problems of communities attempting to survive and prosper in the diverse geographical zones of the region.

The peoples

The ancient Yuman tribes inhabit the floodplains on both sides of the Lower Colorado River (Yuma, Mohave) and the Middle Colorado highlands (Havasupai, Hualapai [Walapai]), as well as the lower Gila (Cocopa, Maricopa) and the Rio Verde (Yavapai). The Pima and Papago, constituting the second group, live along the middle Gila River and in the basin and range country west of the Santa Cruz River, as well as in Sonora across the Mexican border.

The Colorado Plateau and the Middle Rio Grande, with its tributaries, have long been the home of the third group, the Pueblo (village-dwelling) Indians, who, although highly diverse linguistically, share many basic cultural traits. These form three subgroups: the western Pueblos (Hopi, Hano, Zuni, Acoma, and Laguna); the central Pueblos (Jemez, Santa Ana, Zia, Cochiti, Santo Domingo, and San Felipe); and the eastern Pueblos (San Juan, Santa Clara, San Ildefonso, Nambe, Tesuque, Taos, Picuris, Isleta, and Sandia).

Finally, also on the Colorado Plateau, completely surrounding the Hopi villages and Hano and separating them from the other Pueblos, dwell the fast-growing Navajo, a branch of the Athabascan-speaking Apache who are relatively late comers to the Southwest. Their nearest linguistic relatives are found in parts of California, Canada, and Alaska. The Apache, inhabiting the mountains of the plateau rim, form a wedge between ancient Pueblo inhabitants of the region and the Pima–Papago. The major Apache tribes are the Western Apache, the Chiricahua, the Mescalero, the Jicarilla, the Lipan, and the Kiowa Apache.

TRADITIONAL CULTURAL PATTERNS

Differences among the four native groupings

General characteristics: ethos. Even though the four native groupings share some patterns of behaviour, they also evidence considerable variety in the ways in which they adapt to life and perceive life. Moreover, all of the distinctively Indian points of view differ radically from Euro-American patterns of thought. For example, River

Yuman tribes, who combined collecting edible fruits and seeds with farming rich bottomland, were relatively free of the food anxiety common to most desert dwellers. Yuma tended to focus attention on the acquisition of supernatural power by means of singing songs acquired through dreaming and illustrative of myth sequences. The individual's success in life, including leadership in warfare and curing rites, was considered to be entirely dependent on this unusual form of solitary dreaming.

On the other hand, the desert-dwelling Papago tempered their reliance on magical power achieved through singing, vision seeking, and intoxication, with concentration on such practical living problems as the defense of village communities dependent on flood farming, hunting, and collecting wild foods. Their wary approach toward any sort of personal commitment and their preference for perceiving the world in terms of flexible wholes and a continuous range of inductions, rather than the two-sided options characteristic of Euro-Americans, directly reflected a long-range, successful adjustment to an exceptionally inhospitable environment.

The Hopi Indians expressed an extreme variation of Pueblo culture developed in response to their unusually dry niche of the high, semiarid Colorado Plateau. The Hopi tribe's solution to the well-nigh insoluble living problem of village farmers without a permanent water supply was embodied in a flexibly balanced social system that stressed not only group self-defense, sobriety, and peaceful relations but also dry and arroyo flood-farming techniques developed to a high degree of skill. Their technological expertise was supplemented by formal ceremonial activities and reinforced by a conception of the world requiring, in order to keep the system in working order on behalf of the mutual welfare, active collaboration of each individual Hopi, with superhuman powers embodied in deities, animals, and plants. Consequently, a Hopi tended to develop a special type of planning mentality with a complex approach that took into consideration well-formulated whole problems with their effective parts in balance.

Another version of the Pueblo culture was found at Zuni, which differed from Hopiland in having sufficient water to meet the needs of the pueblo. From the time of its European discovery (1539–1540), Zuni was under strong pressures from Spanish explorers, soldiers, Franciscan missionaries, and governors to repudiate its indigenous customs, especially its religious beliefs and activities. The Zuni's solution to this persistent cultural harassment was figuratively to draw a magic circle around their innermost beliefs and ceremonies in order to keep them hidden from prying foreigners. Outside this sacred core they allowed certain changes to occur in behaviour and attitudes. Thus, the famous dances of kachina (masked impersonators of deities) functioned as a screen for the perpetuation of ancient rites. When the wholeness of their culture was destroyed by the intrusion of outside influences, the Zuni became preoccupied with a rote-perfect rendition of detailed ceremonial procedures, each fraught with magical efficacy. Zuni tended to perceive and emphasize tiny details rather than complex wholes, as among the Hopi, or global wholes, as among the Papago, but this tendency is giving way to a more integrated approach toward the external world.

In contrast to the Pueblos, the Navajo, who entered the southwestern region more recently than other Indian tribes and settled there by expropriating crops, herds, and rangelands, expressed an aggressive, mobile way of life. Their culture pattern, bolstered by an animistic conception of the world, fostered in the individual a matter-of-fact mental approach that singled out for attention large, discrete details with little interest in relationships between them. Such an approach relied both on adjusting to new situations by adapting segments from other cultures and on avoiding long-range planning. The Navajo were well able to make and carry out complex planning projects, however, if they believed such activities would enhance their welfare.

Social structures and economic systems. The most favourable habitation sites of the region, from the indigenous viewpoint, were probably the basins of permanent rivers, especially the floodplains of the Lower Colorado. Toward the end of the 18th century, the most desirable patches of bottomland were densely settled by village-dwelling Yuma Indians. Each village comprised a relatively stable and autonomous band composed of a number of large families living in brush shelters or rectangular, sand-covered houses. The male head of each family participated in an informal village council that settled disputes (often over land ownership) and made decisions on community problems. The wisest, ablest, and most powerful senior member of the band was its acknowledged leader. A number of bands constituted the loosely organized but strongly identified tribe, which fostered friendly relations and trade with neighbouring tribes (Mohave, Yavapai, Papago) and maintained a state of hostility or open warfare with others (Cocopa, Maricopa).

Yuma social and economic patterns

Unlike most other southwestern tribes, the Yuma did not suffer from a water problem, despite a minimum of rainfall and a hot desert climate, Although they collected edible fruits and seeds, fished, and hunted small game, the river was their lifeline. In small, irregular fields that were flooded and silted in spring when the Colorado overflowed its banks, they cultivated, with simple hand tools, several varieties of rapidly maturing maize, as well as beans, pumpkins, melons, and grasses (later, wheat and cowpeas acquired from the Spaniards). Such fields were privately owned and were inherited in the male line. Abundant harvests, supplemented by gathering and storage of food in large baskets, coupled with a simple inventory of material possessions, allowed the Yuma to concentrate on a rich world of dream and fantasy.

Pima–Papago social and economic patterns

Although the culture of the Pima–Papago showed a certain resemblance to that of the Yuma, it represented an alternate version of the ancient Southwestern desert tradition. In the Pima–Papago case this tradition was modified and enriched by a prehistoric influx of immigrants who built canals to irrigate larger areas of cropland than had been possible by flash flooding alone. After the first summer rains, fast-growing varieties of maize, as well as tepary and lima beans, pumpkins, and cotton, were planted to take advantage of the moisture before it evaporated. During this time (July and August) the Papago lived in several (perhaps 50 or 60) scattered *ranchería* villages near their fields. After the harvest, when water and food became scarce, each village community moved from its "fields" village to its "wells" village in the hills, where fresh water and game were obtainable.

Each Papago village was composed of several related extended families living in widely spaced wattle-and-daub houses and led by the oldest active man. A village council, attended by all the adult men of the community, met nightly to discuss and resolve local problems, such as land distribution and defense. Each village had its "keeper of the smoke," a religious leader (shaman) who presided over the village roundhouse and led its communal ceremonies. A number of villages formed a dialect group, and six or seven such groups constituted the tribe—a nonpolitical entity united by traditional, linguistic, and genetic bonds.

Western Pueblo social and economic patterns

Compared to the Pima–Papago and Yuman tribes, the Pueblo villages were more independent as well as more sedentary, spatially compact, and highly structured both socially and ceremonially. Although they shared a common theocratic life-style based on complex interweaving ties between kinship and religious groups and also a similar pattern of perceiving the world of nature, they differed considerably both in language and in the way the members of each village organized themselves to cope with the vicissitudes of their particular environment. Life for the western group of Pueblo farmers on the high, semiarid Colorado Plateau was particularly hazardous and nowhere more so than on the mesas of Hopiland. Here the well-being, even the survival, of whole communities long depended on what seemed to the casual observer to be minute, even insignificant variations in climate and topography—especially in the amount and seasonal rhythm of precipitation, the configuration of floodplains and ero-

sion of ephemeral stream beds, the presence and flow of freshwater springs, even the piling of sand into fixed dunes that conserved moisture.

In all of the western Pueblos the key social grouping was the maternal family line, or lineage. This was a kinship group centred in a core of blood-related women but conceived as timeless, extending backwards into the remote past and forward through generations yet unborn. Several related lineages formed a corporate clan with descent through the female line.

Among the Hopi, sets of linked clans, considered to be ceremonial "partners," formed what have been called phratries. Under this system a person was not allowed to marry into the lineage of his mother, his father, his mother's father, or into his own phratry or set of clans. The Hopi clan owned croplands and ritual paraphernalia, and its oldest active woman functioned as its "real" head, while her brother assumed the responsibilities of ceremonial leader. Although the ceremonial leader supervised the annual re-enactment of various ceremonies, the performance of a major ceremony was the responsibility not of a single clan but rather of a voluntary secret society that drew its predominantly male initiates from the entire pueblo. In each pueblo were several such societies. Their elaborate ritual activities were centred in a number of kivas (underground ceremonial chambers).

The kinship system was extended symbolically beyond the human community into the world of nature, linking clans with certain kinds of animals and plants as well as with other classes of natural and supernatural phenomena into a supersociety that included all aspects of the Pueblo world considered important to its well-being.

The basic local unit in the social organization of the western Pueblos was the extended family. This was a female-centred group occupying a common household of two or more adjoining rooms in a three-storied terraced communal structure. It was built of stone, faced the sun, and partially surrounded a central dance plaza. A household normally consisted of a core of women related through the female line, together with their husbands and unmarried sons (and unattached male relatives of the women), as well as their daughters, sons-in-law, and maternal grandchildren. In Hopiland the women of the household cared for the children, cultivated spring-irrigated gardens, produced baskets and pottery, and had charge of the cooking and storage of food as well as of certain clan fetishes. On the other hand, the men of the household raised field and dune crops of corn, squash, beans, and cotton (later apricots and peaches) and wove clothing. They were also the sheep herders. They "belonged ceremonially" to their mothers' households, however, to which they returned frequently to fulfill religious responsibilities.

The Hopi fired their painted clay pots and warmed their ceremonial chambers and homes with soft coal, which they mined by hand methods from open veins near their villages. Coal mining (the Hopi were the only American Indians to have discovered it) helped to create favourable wintertime conditions for building a surplus of skillfully fabricated craft goods—baskets, pottery, cloth woven of cotton (later of wool), and ceremonial paraphernalia. Such goods formed the basis of a trade network that supplied food and other necessities to the mesas.

Central and eastern Pueblo social and economic patterns

Except for Acoma and Laguna, the Keresan-speaking central Pueblos are located on the banks of the Rio Grande and its tributaries, with a permanent and dependable water supply that is used for household purposes and irrigation. In contrast to the pueblos in the west, the Keresans and their eastern-Pueblo neighbours inhabited adobe, rather than stone, houses and built circular, rather than rectangular, kivas. These villages also cooperated in interpueblo ceremonies, but each formed an economic unit engaged in agriculture, gathering, and hunting.

Except for the Tewa, knowledge of the social structure of the central and eastern Pueblos is incomplete and often conflicting, partly because, in reaction to the Spanish system of administration, they tended to disguise their native rites and beliefs (while still holding on to them strongly). It is known, however, that the Tewa-speaking eastern Pueblos exhibited a unique type of dual organization of social and ceremonial life—one division associated with summer and the other with winter and each, on the basis of solar observations, regulating various seasonal activities and interpersonal village rivalries. In general, they divided communal responsibilities.

Navajo and Apache social and economic patterns

The most aggressive Southwestern tribes were its most recent (about the 11th century) Indian settlers—the Athabascan-speaking Apache and Navajo. (The Navajo are generally believed to be descendants of a band of hunters and gatherers closely related to the Apache.) By the beginning of the 17th century the Navajo on the Colorado Plateau achieved a relatively settled way of life. From the neighbouring Pueblos, they had learned to grow corn, to weave, and to care for livestock. Toward the end of the 17th century, however, when the Spaniards and Pueblo Indians attempted to re-establish themselves in the area, the Navajo developed their propensity to spread into and prey upon settled communities to acquire food, sheep, horses, and cattle. The movement of Navajo groups westward into and around Hopi lands seriously threatened the precarious economic situation of that tribe.

The basic socio-economic unit of the Navajo was the extended family, which consisted of one or more related women, their husbands and unmarried sons, their daughters, sons-in-law, and grandchildren. Certain other unattached relatives might also reside with the group. Each "nuclear" family of husband, wife, and children—or each wife and her children, if a man had more than one wife—occupied a separate hogan (circular, earth-covered lodge), but the entire extended family cooperated in activities such as farming, sheepherding, and family ceremonies. Navajo extended families were grouped into about 60 clans, with descent reckoned through the female line; and marriage within one's clan was considered incestuous. Neighbouring extended families cooperated in solving land-use problems, such as range management, farmland development, and water use, under the direction of a noncoercive type of headman chosen for his leadership qualities. Such a community (which might occupy 12,000–80,000 acres [5,000–30,000 hectares]) was called an outfit. All the outfits together constituted the Navajo tribe, which was not a political unit but was unified by traditional, linguistic, and cultural bonds.

The Apache were probably the least understood of Southwestern Indians. Their culture ranged from that of the Western Apache group, who resembled Pueblos and Navajo in the practice of agriculture, in clan organization, and in the elaboration of mythology, to that of the Chiricahua Apache, who retained their mobile, war-geared organization and their hunting economy until the late 19th century. The Chiricahua consisted of three bands, which were associated respectively with the northern, central, and southern parts of their vast semiarid tribal territory ranging from the Rio Grande to southwestern Arizona and from northern Mexico to Zuni and Acoma. Within each band's hunting grounds, sites favourable for defense, hunting, and food collecting were used as temporary encampments and rallying points for smaller local groups the members of which cooperated in such undertakings as war and important ceremonies. Each local group consisted of a number of widely scattered extended families similar to the families of the Navajo, except that descent was reckoned through both parents. The Chiricahua family, however, was more mobile than that of the Navajo.

As among the Navajo, the family in all the Apache groups was centred in related women who cared for the children; gathered and processed edible seeds and other wild plants, such as yucca fruit and mescal; collected firewood and water; produced buckskin clothing, baskets, and pottery; and constructed the dome-shaped, branch-covered shelters that housed the group. The men of the family, on the other hand, hunted, fought, raided, and made weapons, such as arrows, slings, and shields. The most persuasive, strong, and successful family heads became Apache leaders whose authority depended primarily on their personal charisma and success in warfare.

Socialization and education. In all the Southwestern tribes, child training was a serious adult responsibility. The Indians were conscious of a maturation process, which was regarded in several tribes as a "path" along which the individual made his "journey through life."

Without exception in these tribes, a child was treated with warmth and permissiveness during the period of infancy (which might last from one to three or four years). He was given the breast whenever he cried. Weaning was gradual, and training in cleanliness was delayed until a child could walk or even longer. Care was taken not to agitate him unduly, and he was protected from harm (often by means of magical prophylactics) and accepted as an integral part of the family. In some tribes the infant spent much of his time laced onto a cradle board which, among the Navajo at least, was hung up in the hogan so that his head was level with the other members of a seated family circle. The effect of this treatment was apparently to augment the child's physical safety and to enhance his feeling of security. Among the Hopi the use of the cradle was viewed as an early conditioning to the restrictions of the difficult Hopi way.

Adolescence

After the infancy period, tribal socialization patterns began to diverge. From the beginning of childhood there was a strict sexual division of labour. Little boys were given chores appropriate to developing strength, such as collecting firewood or tending animals, while little girls early began to learn food processing and child care. But the child soon learned that every individual in the group was expected to pull his own weight, at every age grade, according to his sex, strength, and capacities.

When they were between five and seven years old, the boys in all the tribes began to associate almost exclusively with the men of their households, who from then on directed their education into masculine tasks and lore, while the girls remained with their mothers and aunts, taking on increasing responsibilities for the exacting tasks of the household. Among the warrior tribes the physical strength, stoicism, and skill needed for the warpath were stressed, and warrior training intensified as the youth grew to young manhood. Even among the peace-oriented Pueblos, however, boys learned agility, endurance, and speed in running. Racing was important ceremonially because it was considered to possess magical efficacy in helping plants, animals, and human beings to grow.

Among the Papago–Pima a child early learned the fundamental principles of family life: the precedence of males over females; seniority and respect for age, regardless of sex; solidarity among all family members under the leadership of the oldest active male; and respect for the role, function, and opinion of every member of the group. Within these limitations the Papago child, male or female, was allowed more freedom of action than was customary elsewhere among the Southwestern tribes.

By contrast, the Pueblo child, after infancy, was probably the most restricted, the degree of harshness in his training apparently depending on the severity of the living problem that the particular Pueblo community had to resolve. The transition from each phase of the Pueblo life cycle to the succeeding one was marked with a major ceremony. Any child who failed to move along the path of life at the culturally designated pace was pulled or pushed in formalized ways to meet the standard. For example, during the secret kachina ceremony, which marked the initiation of all Hopi children into the tribe at about seven years of age and their introduction to the Hopi gods (impersonated as masked dancers), it is reported that all the novices were ceremonially whipped to exorcise evil influences. "Bad" boys, however, were whipped more severely than the others.

Whereas the Pueblos stressed life-crisis ceremonies that offered symbolic solutions to the major problems faced by the entire community of humans, animals, and plants, the warrior groups in this region emphasized rites by which an individual might ward off sickness and acquire magical power for personal success in war and the hunt. In still another manner, the Papago youth was expected to go alone into the desert in quest of a power-giving vision to help him through life.

Belief and aesthetic systems. The life-crisis rites reflected the religious belief systems of the several tribal groups. The spectacular, communally centred Pueblo ceremonies for rain and growth reflected a conception of the universe in which every person, animal, plant, and supernatural being considered significant to the Pueblo lifestyle had its place and role. Without the active participation of every individual in the group, it was believed that the life-giving sun would not return from his "winter house" after the solstice; the rain would not fall nor could the crops grow. In fact, without human help in an annual cycle of ceremonies, the cosmic order would be in danger of breaking down.

Importance of rites and ceremonies

To the Yuma, on the other hand, the universe was pervaded by a single animating principle that was the source of all supernatural power. There was only one medium, namely dreaming, for the personal acquisition of power that was considered necessary for success in life. Sequences of traditional myths acquired through dreaming were converted into songs and acted out in ceremonies. So great was Yuman interest in this kind of dreaming and the power that it could bring to an individual religious or war leader that all other activities—farming, food collecting, and even hunting—were of secondary concern.

The religion of the Papago seems to reflect their position between the Yuman tribes and the Pueblos. Not only did they "sing for power" and go on individual vision quests, but they also held regular communal ceremonies to "bring down the clouds" and keep the world in order.

By contrast, the Apache conceived of the universe as composed of a great many different kinds of power entities, such as animals, plants, witches, superhuman beings, rocks, and mountains. Each power source could exert force in the world for good or ill, with reference to human beings, and each had to be separately propitiated. Each was personalized, talked to, sung to, scolded, or praised. Apache ceremonies were concerned mainly with the magical coercion of these power entities for the curing of disease and the acquisition of personal success in hunting and warfare.

Navajo ceremonial was apparently based on an elaboration of a similarly animistic view of the universe, with the power sources both diffused and specified. Power was localized in a great many autonomous personal beings who were dangerous and unpredictable. These were of two classes: Earth Surface People (human beings, ghosts, and witches) and Holy People (supernaturals who could aid or harm Earth Surface People by sending sickness).

With the acquisition of techniques of farming, herding, weaving, and silversmithing, and the abandonment of hunting and warfare, the Navajo turned their attention to elaborate and colourful ceremonies (called "sings") that aimed, by means of compulsive magic, to cure sickness and bring an individual into harmony with his family group, nature, and the supernatural.

CHANGES AND MODERNIZATION OF THE CULTURES

Heritage of integrating and disintegrating factors. Each local community tried to maintain a delicate balance between population and natural resources. If the population outgrew the capacities of its base within the limitations of the group's technology, a segment or faction might split off and form a colony in a favourable habitat resembling that of its parent group. Under normal conditions the daughter colony was so constituted to reproduce as far as possible the parent culture even in its most esoteric aspects. If prolonged drought (as in 1272–1299) dried up springs needed by Hopi villages for drinking water, an entire community might migrate.

Traditional adaptations to change

By contrast, human pressures from without, such as raids by marauding bands or aggressive missionary activity, caused some pueblos to consolidate or to move to less vulnerable mesa sites. In 1680, for instance, all the Pueblos united to kill or drive out the Spaniards (who, however, eventually re-established Catholic missions in all the pueblos except those of the Hopi). Papago settlements consolidated into large compact villages for defense against the Apache, but they spread out again after the raids ceased.

Religious reactions

Constant or intermittent efforts by missionaries, educators, and administrators to undermine native ceremonies and to change indigenous beliefs were resisted in various ways. The mobile Navajo–Apache tribes rejected all efforts of the Spaniards to reduce or convert them. The more accessible sedentary Pueblos employed various devices, under cover of practicing the outer forms of Catholicism, to hide their sacred religious paraphernalia and to perpetuate their essential beliefs and ceremonies. The Papago produced their own Christian sect, a blend of native and mission practices known as Sonoran Catholicism.

Unless totally destroyed, a pueblo of the western type, which was relatively decentralized, did not surrender structurally to foreign control. In fact, pueblo organization was characterized by a permanence and continuity that would not accommodate externally imposed social change. Alterations in socio-political or ceremonial patterns due to foreign intervention occurred only at superficial levels, while basic religious beliefs and practices tended to persist (even to the present day).

By contrast, pueblos of the eastern type, with more centralized control, proved more vulnerable to attack at both the socio-political and the ceremonial levels. If not utterly destroyed, most of them incorporated aspects of the foreign system into their own structures so that the result was a blend of the two.

The Navajo and the Western Apache adapted their mobile hunting cultures by borrowing from the Pueblos whole complexes of traits, such as agriculture, herding, weaving, and sand painting, and later from their industrialized neighbours, such traits as lumbering, mining, and commercial enterprises.

In virtually all the Southwestern tribes, however, the native languages are still spoken in the family groups, which continue to organize themselves according to traditional principles and values. Native ceremonies are still practiced, often secretly, in all the tribal groups, even those considered to be at least nominally Christian.

Effects of modernization

Current cultural conditions. Although the seeds of "modernization" probably were sown earlier, the process did not get under way until the end of the 19th century and the beginning of the 20th. The isolation of many of the tribes (the Havasupai at the bottom of the Grand Canyon are still the most isolated Indians in the United States), the vastness, aridity, and poor natural endowment of many of their lands, and, of course, the ferocity of the Athabascans all played a part in delaying modernization of the Southwestern tribes. Indeed, military defeat, the loss of traditional lands, and missionary efforts to change their religious beliefs and practices left a heritage of rejection and bitterness among these Indians.

Under the administration of the United States, many changes have been made on the reservations in the interests of Indian welfare. Tribal government has been developed and empowered with legal authority. Schools and hospitals have been constructed. Irrigation systems have been built and crops increased in quantity and quality. Roads and highways have been improved and communications by telephone and radio, bus and airplane inaugurated. Ranges have been protected and erosion controlled, while livestock has been upgraded by scientific breeding practices. Timber has been commercialized on a sustained yield basis, and sawmills have been constructed. Uranium and copper have been mined, oil and gas leases have been granted, and electronic as well as many other industries have been established.

Still there remains an "Indian problem." Indian incomes are generally low compared to those of the rest of the American population, especially among the Papago, Hopi, Fort Apache, and some of the highland Yuman tribes. Unemployment is high among Indians both on and off the reservations. Absenteeism and irregularity among those with jobs is high. Indifference and hostility to white ways, alcoholism, and crime are significantly present.

Biculturalism

In the opinion of some authorities, however, the Indian problem in this region is not poverty or lack of industriousness or even apathy. The primary problem rather is biculturalism. Today, the tribal reservations are the physical and spiritual homes of the several Indian groups, half of the members of which may reside outside their boundaries. All except a few who have identified with the surrounding American culture periodically return to their family households on the reservations for ceremonies and for deep religious and social nourishment. Contrary to what may appear on the surface, almost all these Indians —regardless of their degree of education and training, the work they do, and the measure of their integration with the surrounding industrialized civilization—remain Indians at heart, each according to the attitudes, values, conception of the universe, and socialization patterns of his tribe.

BIBLIOGRAPHY. EDWARD H. SPICER, *Cycles of Conquest: The Impact of Spain, Mexico, and the United States on the Indians of the Southwest, 1533–1960* (1962); and FRED EGGAN, *The American Indian*, ch. 5 (1966), are the most comprehensive, authoritative recent accounts. Among other significant studies of the region as a whole are: EMIL W. HAURY (ed.), "The Southwest Issue," *Am. Anthrop.*, 56:529–731 (1954); JOHN COLLIER, *Patterns and Ceremonials of the Indians of the Southwest* (1949); RUTH M. UNDERHILL, *Ceremonial Patterns in the Greater Southwest* (1948); ELSIE CLEWS PARSONS, *Pueblo Indian Religion* (1939); RUTH BENEDICT, "Psychological Types in the Cultures of the Southwest," in the *Proceedings of the Twenty-Third International Congress of Americanists*, pp. 572–581 (1930); and GEORGE FRANCIS CARTER, *Plant Geography and Culture History in the American Southwest* (1945). Recommended tribal studies are: for the Yumans, C. DARYLL FORDE, *Ethnography of the Yuma Indians* (1931) and LESLIE SPIER, *Yuman Tribes of the Gila River* (1933, reprinted 1970); for the Pima-Papago, WILLIAM H. KELLY, *The Papago Indians of Arizona* (1963); RUTH M. UNDERHILL, *Singing for Power* (1938, reprinted 1968); ALICE JOSEPH, R.B. SPICER, and J. CHESKY, *The Desert People* (1949); for the Pueblos, FRED EGGAN, *Social Organization of the Western Pueblos* (1950); EDWARD P. DOZIER, *The Pueblos of the Southwestern United States* (1960); ALFONSO ORTIZ, *The Tewa World* (1969); CHARLES H. LANGE, *Cochití* (1960); DOROTHEA C. LEIGHTON and J. ADAIR, *People of the Middle Place: A Study of the Zuni Indians* (1963); and LAURA THOMPSON and ALICE JOSEPH, *The Hopi Way* (1944, reprinted 1965). Other viewpoints regarding Pueblo personality development are found in ESTHER S. GOLDFRANK, "Socialization, Personality, and the Structure of Pueblo Society (with Particular Reference to Hopi and Zuni)," *Am. Anthrop.*, 47:516–539 (1945); and DOROTHY EGGAN, "General Problem of Hopi Adjustment," *Am. Anthrop.*, 45:357–373 (1943). Recommended for the Navajo are CLYDE KLUCKHOHN and DOROTHEA LEIGHTON, *The Navaho* (1946); DOROTHEA LEIGHTON and CLYDE KLUCKHOHN, *Children of the People* (1947, reprinted 1969); for the Apache, MORRIS OPLER, *An Apache Life-Way* (1941).

(L.Th.)

Sovereignty

Sovereignty may be defined as a theory of politics that deals with an ultimate overseer, or authority, in the decision-making process of the state and that explains the rules of maintaining order (internal or political sovereignty). In a more limited but related sense, sovereignty is the power to make or change the law. In the context of international law and relations, sovereignty is the freedom of a state from external control (external sovereignty).

Etymology of the term

If the etymology of the word—its derivation, through Old French *soverainеté*, from the medieval Latin *supremitas*, or *suprema potestas* ("supreme power")—indicates the meaning of the theory, it also reveals the difficulties inherent in it. Qualities such as "supremacy" or "absoluteness" that have been attached to sovereignty oversimplify and often obscure the issue. The problem is complicated by the many aspects of sovereignty that have emerged in political theory, including internal, external, legal, political, popular (in the sense of all power deriving from the people), and limited sovereignty.

THE THEORY OF SOVEREIGNTY

The locus of sovereignty. To speak of entities in which, according to various theories, sovereignty does or should reside, provides a shortcut to the ideas of particular thinkers. The locus of sovereignty may be the monarchy, the people, Parliament, parliamentary and judicial

institutions, or shifting pluralist groups. A commonly held position is that the locus of sovereignty is important because it determines the relationship of the individual to the state and government. In fact, sovereignty itself states that relationship. That is what sovereignty is about. Locating sovereignty in the people, the constitution, or the various institutions violates the logic of the concept: such plausibility as is possible is achieved by isolating concepts of "power" and "authority" from all other considerations. It is important to the concept of sovereignty that it not be given a locus. "Where ought the sovereign power of the state to reside?" asked Aristotle. "With the people? With the propertied classes? With the good? With one man, the best of all the good? With one man, the tyrant? There are objections to all these."

Limitations on the exercise of sovereignty. Perhaps the most misunderstood attribute of sovereignty concerns its so-called unlimited character. Among limiting factors, natural and positive law, constitutions and "practical" limitations are usually mentioned. Objections raised against sovereignty imply that they are limitations on sovereignty, whereas in fact they are denials of the latter. Instead of clarifying the subject, they explain the concept away. By being labelled unlimited, or omnipotent, sovereignty is transformed into a shibboleth, an old-fashioned, abandoned doctrine. Pronouncements by critics have obscured the original meaning: the definitions of the "classics" are often quoted, but neither their assumptions nor the implications of their ideas (or their relevance) are adequately considered. Thus, to dwell exclusively on an often-quoted definition by the French jurist Jean Bodin, "Majestas est summa in cives ac subditos legibusque soluta potestas" ("highest power over citizens and subjects, unrestrained by laws"), and to play down his qualifications (that the sovereign is limited by natural and divine laws and by the laws of the realm) is to miss the intent of the doctrine, which was to make the sovereign non absolute.

Bodin. According to Bodin, the definition of the state as "lawful government" is "the very end and scope of the matter propounded." This statement can be easily overlooked if one tries to separate sovereignty from the law that gave rise to it. On the other hand, the problem with juristic theories in general is the attempt to reconcile irreconcilables—government under law with the concept of the lawgiver who must be above and outside the law. Since Bodin, two types of reconciliation have been most popular: the identification of government and the governed (popular sovereignty), and the identification of government and law (constitutional sovereignty). Neither of these was Bodin's solution. He presupposed that human beings want to live under law; that the problem of government is to define authority so that desire can be fulfilled; that a lawmaker is necessary; and that the argument for the need of adherence to the law or will of the people serves no purpose when the law is being questioned. Logically, there can be only one supreme authority. If there were more than one, the emergence of more than one set of laws would be likely to ensue and conflict with the basic human desire to live under law. What emerges from this conception of sovereignty is that it implies two different views of law: law under which people want to live, and law as the command of the sovereign.

Although the necessity of obeying the second type of law logically derives from the desire to have the first, the two types are not equivalent. There is nothing in Bodin's doctrine either to compel the sovereign to observe the first or even to give him a clear idea of what it is. The problem with Bodin's theory is not that there is no legal restraint on the sovereign but that the sovereign has no way of knowing when he is fulfilling his function, whereas his subjects know—but cannot do anything about it.

Hobbes. In Hobbes's theory of sovereignty this difficulty does not arise because in his theory it is not the desire for law that leads to the acceptance of sovereignty but rather the desire for a particular type of lawmaking.

The natural law conceptions that underlie Bodin's theorizing—his concept of sovereignty was a part of natural law—play a different role in Hobbes, who detaches natural law from divine law. Hobbesian "natural laws" are laws of self-interest and their enforcement is made part of the self-interest of the sovereign. Hobbes's sovereign knows what he has to do to be a sovereign—he is in fact closer to being a function than an actual human being with a physical existence—whereas Bodin's sovereign is a supreme lawgiver who is in a curious position of having no clear notion of the type of laws he is supposed to proclaim. Little wonder that, subsequently, jurists have wanted to convert such a sovereign into the law itself. The assumption about the desire to live under law could then be reconciled with something approximating the rule of law.

Assumption of Hobbes's theory

Hobbes's theory of sovereignty, like Bodin's, emerged during a civil war, a fact that was conducive to an opinion that the theory is an argument for absolutism, for a power strong enough to impose a social order on those who, in Hobbes's own argument, lack a natural desire to live together in peace and harmony. For Hobbes, civil war is not an accident of circumstances but an ever-present threat; man is utterly egocentric and does not feel bound by any sense of obligation resulting from benefits conferred in the past; unless constrained by a superior force, he will keep no agreements nor observe any laws that are not to his advantage.

According to Hobbes, when men begin with the assumption of basic equality and egoism and reflect on the consequences of living together without imposed restraints, they recognize that life would be "nasty, brutish and short." To him, the social contract was not a legal fact of life in society but a personal—and essentially selfish—line of thought that each man follows when he raises the question of his relation to society. It is a continuous process rather than a hypothetic historical event the consequences of which have been imposed on later generations. In a world in which no one can be trusted to keep any agreement not to his advantage, each man agrees that mutual interests are so necessary to life in society that their requirements must be enforced. Because everyone is, or at least feels, equal in the state of nature and hence equally capable of violating agreements when personal advantage is at stake, an artificial inequality must be created—a sovereign superior in power to everyone else who can ensure that mutual interests are enforced.

There can only be one such supreme power, and it cannot be bound by any terms in the contract. Although the contract may seem to create a sovereign with unlimited powers, his power is not absolute but conditional. It exists only so long as expectations are fulfilled that the sovereign power can maintain law and order, so long as agreements be kept, and norms of the community be observed. No party to the contract has agreed that the will of the sovereign (as a private will) shall prevail over his private will; for a contract then would make the hypothetical state of nature an inescapable fact of life under sovereignty. If man desires power as a good in itself, however, there is a danger that granting of special powers to the sovereign may lead to their unlimited extension. Hobbes dealt with this aspect of absolutism when he said:

> I put for a general inclination of all mankind, a perpetual and restless desire of power after power, that ceaseth only in death. And the cause of this, is not always that a man hopes for a more intensive delight, than he has already attained to; or that he cannot be content with a more moderate power: but because he cannot assure the power and means to live well . . . without the acquisition of more.

This has been widely interpreted as a statement about an innate power drive. But to Hobbes power hunger is a product of circumstances—"because . . . [man] cannot assure the power and means to live well." The purpose of the contract was to ensure this and hence preclude "the general inclination of mankind" from being a component element in the sovereign power.

Rousseau. Rousseau's analysis, whereas it has fitted democratic concepts of popular sovereignty and has had political impact, has been less influential in political science than the Hobbesian view.

Unlike Hobbes, Rousseau's hypothetic natural man was not utterly self-centred: his sense of self included a capacity for sympathy. There was nothing in the state of nature impelling men to establish a power over themselves capable of limiting the injuries they inflicted on each other. Indeed, there was nothing in the state of nature that made it imperative for men to leave it, so that Rousseau either had to assume that man was still in a state of nature or that there was a capacity for corruption—either in man or in social orders—that led man away from the state of nature. Because it was impossible to believe that contemporary European societies represented the state of nature, Rousseau chose the second possibility; of the alternatives in the second possibility, he chose the view that the social order itself can corrupt. (If the capacity for corruption lies within man himself, there is no escape: man would be better off as a solitary animal, even though that would be contrary to his nature.) But not all social orders corrupt. Because of man's nature, some kind of social order is inevitable, and what needs to be done is to ensure the existence of the kind that does not corrupt.

Rousseau's concept of the general will

Rousseau's conception of man's nature comprises the idea that each man has within himself a knowledge of the good and a desire to act by it and that this good looms larger than self-interest because man is not purely egoistical. This idea is not the same as the will of all because it is only part of any man's will—each having selfish interests—and there is no reason to suppose that the mere counting of "ayes" will ensure that the will, which Rousseau calls the General Will, will prevail. The problem is to find a form of society in which this will shall be sovereign. Democracy alone will not secure it, unless selfish elements are eliminated from the individual will. Rousseau offers here no plausible solution, which provides the reason why interpretations of his theory have oscillated between two extremes epitomized in the title "Rousseau—totalitarian or liberal?" Indeed, a dictatorship is no impediment to the "reign" of the General Will because here again each man has the same impulses to act as have all of the others and purely selfish desires need not override the demands of the General Will. The weakness of this conception of sovereignty is that Rousseau has not been able to suggest a method whereby the General Will actually becomes sovereign.

Austin. To grasp the nature of Austin's argument on sovereignty it is necessary to understand why he discusses it in a book on the province of jurisprudence. Although the concept of sovereignty supplies a ready definition of law as the command of the sovereign, it is not a logical necessity to the jurist. Austin introduces the concept in order to distinguish positive law from other types of law. As he puts it, "In order to an explanation of the marks which distinguish positive laws, I shall analyze the expression *sovereignty*. . . ." Inevitably, this purpose colours his discussion. One should remember that Austin is not discussing a theory of government but a theory of law. Because of his interest in using sovereignty to define the fact of law, he has no need to begin with the concept of "contract" that played a central role in earlier discussions of sovereignty. He begins, instead, by distinguishing sovereignty from other powers. He argues: "The *generality* of the given society must be in the *habit* of obedience to a *determinate* and *common* superior: whilst that determinate person, or determinate body of persons must *not* be habitually obedient to a determinate person or body. It is the union of that positive, with this negative mark, which renders that certain superior sovereign or supreme. . . ." The terms he italicized are essential to his theory; significantly, not one of them is mentioned or implied by his predecessors. Austin alone feels that the number of people in a given community who acknowledge the sovereign power is significant. The "habit" of obedience is also a term peculiar to him; it has no function in the Hobbesian concept. It is important in Austin because it describes the essential nature of obedience, which —unlike Hobbesian obedience—is nonrational and (usually) non-motivated. An interesting feature of Austin's definition is that, whereas the attributes of the sovereign are correlatives of those of the subject, the sovereign is described as "not habitually obedient."

Austin's distinction between sovereignty and other powers

The problem implied by Austin's theory is: Are we to define sovereignty in terms of the state or international law? Is sovereignty being granted by the recognition of other states: is it always de jure? Or is it established irrespective of whether other states exist? With regard to the state in its international relations, the difference between Hobbes and Austin is that the former's sovereign is not dependent on whether he is recognized by other states. For Hobbes, de facto and de jure sovereignty are the same thing. Austin's sovereignty, however, has to be de jure, or it will cease to exist.

SOVEREIGNTY AND INTERNATIONAL LAW

Distinction between internal and external sovereignty. In international law sovereignty means autonomy or independence of a state. Either type of sovereignty (internal or external) implies that the state's claim to it is exclusive and that there is no agency higher than the state. But while a sovereign state is master of its acts, the mastery does not extend to all possible acts. The nature of external sovereignty is essentially negative in the sense that—except when it is "abrogated" by a treaty —it entails absence of subordination to a foreign authority. Internal sovereignty is more positive: it establishes the ultimate authority over individuals or groups within the state territory. It is possible for a state to have full external sovereignty but to lack internal sovereignty. Extraterritoriality in an otherwise sovereign state is an instance of this condition. The possession of internal (but not external) sovereignty is more common, as in protectorates and suzerainties. Apart from being an assertion of the supremacy of law internationally, the argument that all states are now bound by the law of nations has some basis in the logic of sovereignty. Internal sovereignty refers to a locus of power and authority such that only one power can assert supremacy and maintain its claim. In a world divided into sets of such powers, external sovereignty can refer only to such residual areas and mutual accommodations as result from the sovereign powers' consciousness of their mutual interdependence and inability to extend their claims to internal sovereignty over larger geographic areas.

Positive nature of internal sovereignty

Sovereign equality of states. Conventional legal theory considers the idea of sovereignty obsolete, "contrary to facts," and lacking in positive content (bereft of the power to give unconditional orders), and stresses the de jure equality of states and their de facto inequality. For the contention to be valid would mean that Hobbes did not succeed in producing an argument for sovereignty. Significantly, Hobbes began with just such assumptions about equality and inequality with regard to individuals. It would seem strange that, with respect to the relations between states, de jure and de facto equality or inequality are supposed to establish that sovereignty is contrary to fact. Clearly, there is a difference between the assumptions made by Hobbes and those of conventional modern jurists. Hobbes's principal argument derives from equality, but by leaving room for conquest as a possible method of achieving sovereignty—in addition to contract—he makes allowances for inequality. To him, equality without normative implications is the primitive condition—neither sought nor asserted, but simply existing as a fact. In Hobbes, equality is the condition that makes the normlessness of the state of nature intolerable, and it is the latter which is given up for the sake of normative order. Only when this has "happened" is the "second method" of establishing sovereignty—conquest (which sanctions inequality)—possible. The situation is reversed in international law. The "equality" of states is the primary condition, but it is pure fiction. States are, in fact, unequal. The equality of states has some other source: the law. The jurist sees that states in fact are unequal, yet has to begin with the fiction of equality in order to apply a law universally. In practice, advanced societies usually exempt certain classes—minors, idiots, madmen—from the full rigour of the law, but the basic meaning of justice demands that it be blind to inequalities—that "equality" be a juris-

tic fiction. Within the state, this fiction operating under sovereignty allows the law to be applied in all cases. In the absence of sovereign power, it makes the law itself a fiction between states. International law is not the result of an accommodation between equals.

As regards equality, it would seem that this concept, which law considers necessary with respect to internal sovereignty, has been uncritically applied to questions of law relating to external sovereignty. One authority has proposed the substitution of "independence" for equality. But it is also not true that states are in fact independent in any meaningful sense. In order to replace equality by independence, it would be necessary to supply an analysis of the new term that would show it capable of making international law applicable to all states.

Sovereignty and recognition. Sovereignty is related to the recognition or nonrecognition of sovereignty. According to one view, international law makes recognition of the state a precondition of its sovereignty. According to another, sovereignty and recognition are not related. In the latter view, claims that the assertion that states have no legal existence before recognition are not supported by practice; sovereignty is neither created by recognition nor destroyed by nonrecognition.

The problem arises when a nonsovereign state happens to assert its independence and demands recognition as a state. The matter is by no means academic. Thus, for many, it is at present an open question whether Rhodesia is an independent state with a legitimate government, or is not a state and therefore has a government that is illegitimate—an issue of special importance to anyone trading with this country, or trying to emigrate from it.

Limitation of sovereignty. War is so often viewed as an expression of power, and power as the essence of sovereignty, that a limitation on the right to war can be regarded as an instance of a voluntary limitation on sovereignty. But it comes closer to being "diplomacy carried on by other means" than it does to being an expression of power, and sovereignty is not to be equated with power.

Sovereignty and the League of Nations

It is often maintained that the Covenant of the League of Nations changed the foundations of law with reference to the place of war in international law. Before the establishment of the League of Nations, international law had no alternative but to accept war. The Covenant created express obligations to use peaceful means for settling disputes; it also established an organization whose task was to see that those obligations were observed by the states and to apply sanctions if necessary. It would appear that international law thus created a power capable of limiting the force of sovereign states in relation to war. This in fact, was not the case and little was changed by transferring the League's authority to the newer United Nations.

At present, "agreeing to abide by decisions of international bodies" is a special case of treaty limitation. No nation has as a rule agreed to submit to the jurisdiction of an international body, except in a restricted sense in particular cases. Thus, when nations agree to the jurisdiction of the International Court of Justice, they in effect make an agreement with each other and abide by the decisions of the Court for the same reason as they abide by treaties. Such reasons can be very complicated and quite unrelated to the problem of sovereignty.

Effect of treaties on sovereignty

It is plausible to examine the extent to which states limit their freedom of action and thereby limit their sovereignty. Nations normally plan to keep their treaty obligations despite the absence of an enforcing agency. The rationale behind treaties is the advantage obtained if both parties keep the agreement. Sovereignty, then, cannot be said to be limited or transferred—there is no one to transfer it to—unless it can be demonstrated that some element of coercion exists. There will always be some pressure—if only that of self-interest—behind all such arrangements, but on what premises one could erect the needed criteria of coercion is not clear. The only way in which a sovereign body can limit sovereignty is by an explicit decision. It cannot do so accidentally, and it would seem misleading to regard law—domestic or international—as being a limitation on sovereignty. Whether it derives from custom, natural law or treaties, law is the expression and evidence of sovereignty, rather than an infringement or limitation of sovereignty. The fact that, domestically, states do not expropriate property owners without compensation or that there is a system of independent judiciary is not regarded as a limitation on sovereignty. Whether voluntary limitations on action are undertaken on the national level or internationally, there is no reason for treating them as limitations on sovereignty.

Sovereignty and intervention. International organizations have limited powers as far as the right of intervention is concerned. Thus the competence of the UN is to "consider," "discuss" and "recommend" with reference to breaches of human rights. Any interference in internal sovereignty that may result is justified on the ground that the action is not enforceable and no intervention within the meaning of the UN Charter (Article 2:7) is involved. On a regional level, the Council of Europe in 1953 set up the European Convention for the Protection of Human Rights and Fundamental Freedoms. A list of basic rights and freedoms subject to protection was drafted and an institutional machinery was created in the form of a Commission and a Court of Human Rights. The most serious limitations on the Court's action stem from the fact that there is no obligation binding member-states to accept the Court's jurisdiction (as is the case with the International Court of Justice); nor, as a rule, can action be initiated by any party other than a state. Although it has been held that one case (*Lawless* v. *Ireland* [1961]) upheld the right of individual petition against a sovereign state before an international court, the known principles of international relations do not seem to confirm this. There is no legislative machinery to enact international law which would be binding on the states or their peoples. States often refrain from ratifying the decisions signed by their representatives at conventions and the reluctance of public opinion to give support to international legislation jeopardizes the process. One of the implications is that the rise of democracy, which has given public opinion more weight in decision making, has interfered with the development of international law.

The UN Charter contains numerous references to human rights, notably Article 1 (which describes the promotion of human rights and fundamental freedoms as one of the purposes of the United Nations), Article 55, and Article 56 (which refers to "joint and separate action" that individual members pledge to undertake in this respect in cooperation with the UN). The Charter does not define the rights and in its "domestic jurisdiction" clause, Article 2:7 states as a general principle that there is nothing in it to "authorize the United Nations to intervene in matters which are essentially within the domestic jurisdiction of any state." This provision raises the question of the extent to which the UN can investigate alleged violations of human rights. Although the Universal Declaration of Human Rights adopted in 1948 by the General Assembly of the UN (which defined some rights, such as rights to life and liberty and to a fair trial) remedies the lack of a definition of rights and freedoms, it is a declaration, not a treaty or legal instrument. If it were a treaty, it would have a major effect on international law. As a declaration, it is but a guide to rights and freedoms.

But it is usually held that even if the Charter is binding in the same way as any treaty, it does not bind the signatory states to recognize human rights as a legal obligation. Apparently, even if a country feels that human rights are being violated in another state, international law does not allow it to proceed against the violator either alone or jointly with other states. Indeed, an act designed to stop a violation of human rights would be an act of aggression according to international law. Human rights exist and are recognized, but their defense is hardly possible.

According to Article 24 of the UN Charter, the responsibility for the maintenance of international peace and security is conferred by its members on the Security Council, whose decisions—provided they accord with the Charter—the members agree to carry out. It is clear,

Effect of decisions of the UN Security Council on state sovereignty

however, that the Security Council can make decisions which member states, and even some members of the Security Council, may not accept. There have been several instances of this situation, such as the Korean War and the Suez Crisis. Great Powers have often refused to be bound by decisions with the formulation of which they did not concur. Technically, the sovereignty of Security Council members has not been limited but extended over states outside the Security Council. Members have a legal right to infringe on the sovereignty of other states for those purposes which they themselves define. Normally, there is no threat to their own sovereignty, or even interests, since they have veto power.

SOVEREIGNTY IN THE 20TH CENTURY

After World War I, some jurists held the view that, with the establishment of the League of Nations, the theory of sovereignty had become academic and impracticable and the sovereign state obsolete. The argument seemed to be that equality before the law—reflected in the expression "equal members"—eliminated the distinction between sovereign and nonsovereign states. The observation that a state bound by treaties was not sovereign was supposed to reinforce this position. The reasoning is unacceptable to many authorities, who argue that there cannot be such a thing as equality before the law in fact, unless there is a sovereign power to ensure it; that the League was not a suprasovereign power any more than is its successor, the United Nations. In this view, all that can be said is that, with regard to their relations with the League, all members were equal—which is not the same as saying that as regards their relations with each other all were equal. These relations were not markedly changed by the existence of the League. The League's task was to deal with relations between states as if there were no difference between a sovereign and a nonsovereign state. When this conflicted with the interests of the states concerned, the League collapsed. It, not sovereignty, proved obsolete.

The above argument is a special case of a commonly asserted view that sovereignty is limited in fact by the presence of treaties between states and hence that sovereignty—in the form required by theory—does not exist. The consequence of this view is that if a treaty diminishes sovereignty, it is conceivable that the conduct of foreign relations will eventually lead to the elimination of sovereignty. Yet the theory of sovereignty has not been affected by such views, although they may well have lain behind some forms of isolationism in the 1930s.

The question remains whether sovereignty can indeed become obsolete. The expression seems to imply that although a function is now performed by sovereignty, that function can be performed by a different agency, or that conditions can change and there will no longer be any need for fulfillment of the function. There is nothing inevitable, so far as one knows, in what human beings create, whether it is sovereignty or anything else. Indeed, the concept of sovereignty presupposes choice, and behind its classical Hobbesian form is the assumption that there have been times when there was no sovereignty and that there will be such times in the future; the possibility of a system without sovereignty is inherent in the argument for sovereignty. If one changed one of the basic assumptions or the system of values itself, sovereignty would become superfluous. Few people nowadays believe that man is as egoistic as Hobbes maintained: numerous Utopian communities have been set up on the premise that man is unselfish (their fate lends weight to Hobbes's view, although it does not necessarily prove he was right). One need only begin with a different, more sanguine assumption about man's nature in order to make sovereignty obsolete. But most authorities believe that few persons are willing to be so sanguine: presumably not because they are cynics by nature, but because they argue that no reasonable person would found a society or advocate change on the basis of mere hope. Sovereignty can probably be made "obsolete" in many ways, and it is the task of political theory to analyze them and to ask whether —in the light of their probable ultimate consequences— it is advisable to choose any of them.

Recent criticisms of the concept

Since World War II the concept of sovereignty has been under concerted attack from philosophers, social scientists, jurists, and students of international relations. Attempts have been made to discredit sovereignty on three grounds: semantic, empirical, and moral. Critics of the concept have dwelt on the diffuseness of sovereignty, the inability to indicate its locus with any precision, and the "anti-democratic" connotations of the concept. They have shown a tendency to take some of the statements of "classical" thinkers at face value, often missing the true intent of the doctrine as revealed not merely by the main body of the latter, but also by its attendant qualifications. They may have shown a utopian bias by projecting their dreams of human nature and international society into the world of contemporary power politics, the realities of which they often choose to ignore. Others argue that what has escaped the notice of such theorists is the logical function of the concept, and the ability of sovereignty as a theory to integrate a number of distinct concepts, such as power, authority, legitimacy, obligation, community, thus providing the foundation for a type of discourse without which political theorizing must come to an end.

BIBLIOGRAPHY. Classic statements of the theory of sovereignty may be found in JEAN BODIN, *Six livres de la Republique* (1576; Eng. trans., *Six Bookes of a Commonweale*); THOMAS HOBBES, *Leviathan* (1651); JEAN-JACQUES ROUSSEAU, *The Social Contract* (1762); JOHN AUSTIN, *The Province of Jurisprudence Determined* (1832); and ALBERT V. DICEY, *Introduction to the Study of the Law of the Constitution* (1885).

The following are general accounts: C.E. MERRIAM, *History of the Theory of Sovereignty Since Rousseau* (1900); S.C. CHEN, *Recent Theories of Sovereignty* (1929); and HYMAN EZRA COHEN, *Recent Theories of Sovereignty* (1937). The pluralist position emerges from the three works of HAROLD LASKI, *Studies in the Problem of Sovereignty* (1917), *Authority in the Modern State* (1919), and *The Foundations of Sovereignty and Other Essays* (1921); and FRANCIS G. WILSON, "A Relativistic View of Sovereignty," *Political Science Quarterly*, 49:386–410 (1934). For the juristic approach, see JOHANNES MATTERN, *Concepts of State, Sovereignty and International Law* (1928); and J.L. BRIERLY, *The Law of Nations*, 6th ed. rev. by C.H.M. WALDOCK (1963), a standard introduction to international law. Recent articles giving juristic views on sovereignty are K.W.B. MIDDLETON, "Sovereignty in Theory and Practice," *Juridical Review*, 64:135–162 (1952); H.W.R. WADE, "The Basis of Legal Sovereignty," *Cambridge Law Journal*, pp. 172–197 (1955); GEORG SCHWARZENBERGER, "The Forms of Sovereignty," *Current Legal Problems*, 10:246–295 (1957); and HANS KELSEN, "Sovereignty and International Law," *Georgetown Law Journal*, 48:627–639 (1960). A comprehensive discussion of the relationship between law and sovereignty, with reference to the legal thought of various traditions, is contained in A. LARSON *et al.*, *Sovereignty Within the Law* (1965). Perceptive, balanced modern statements on sovereignty are J. DICKINSON, "A Working Theory of Sovereignty," *Political Science Quarterly*, 2 pt., 42:524–548, 43:32–63 (1927–28); and W.J. REES, "The Theory of Sovereignty Restated," *Mind*, 59:495–521 (1950). YVES R. SIMON, under "Sovereignty in Democracy," in his *Philosophy of Democratic Government* (1961); and BERTRAND DE JOUVENEL, *De la souveraineté* (1955; Eng. trans., 1957), discuss sovereignty within the context of original political theories. A powerful defense of sovereignty is provided by F.H. HINSLEY in *Sovereignty* (1966), which offers a speculative reconstruction of relevant facts relating to different societies throughout history. See also Hinsley's subsequent article, "The Concept of Sovereignty and the Relations Between States," *Journal of International Affairs*, 21:242–252 (1967); and W.J. STANKIEWICZ, "In Defense of Sovereignty: A Critique and an Interpretation," and "The Validity of Sovereignty," in *In Defense of Sovereignty* (1969), a critical symposium that contains many of the articles mentioned in this bibliography. MAURICE KEENS-SOPER in "Sovereignty Revived?" *Political Studies*, 18:147–150 (1970), assesses the case for sovereignty advanced by the symposium. Among statements by the most outspoken opponents of the concept are JACQUES MARITAIN, "The Concept of Sovereignty," in his *Man and the State* (1951); and STANLEY I. BENN, "The Uses of 'Sovereignty,'" *Political Studies*, 3:109–122 (1955).

(W.J.St.)

Soviet and Socialist Legal Systems

Soviet and Socialist Law is a system of public order devised by Communist politicians after the Russian Revolu-

tion of 1917 and emulated subsequently in Asian and eastern European states where Communist parties became dominant. Although it was classified by some legal scholars as a subdivision of the Romanist, or civil-law, system, its adherents claim independent status for it. This claim was recognized in practice after World War II by election to the International Court of Justice of judges representative of the Soviet legal system in accordance with requirements of the International Court's statute (article 9) that it include judges who are representative of the main forms of civilization and of the principal legal systems.

HISTORY

Origin. Since Karl Marx and Friedrich Engels prescribed no pattern of law for a state organized to achieve the aims set by their *Communist Manifesto* of 1848, it was left for Lenin and his colleagues in the Russian Communist Party after the Bolshevik Revolution of 1917 to improvise a legal system. General guidelines provided by Marx and Engels were that law be regarded as an instrument of the state, not as a limitation upon those who make policy, and that it enunciate rules of public order facilitating the transition to Socialism and ultimately to Communism. Its major tasks were to be two: (1) elimination of the political power of the bourgeoisie (the property-owning middle class) by depriving them of their ownership of productive resources and (2) education of citizens in the disciplined pattern of life claimed by Communists to be requisite to the achievement of the social order they desired. Completion of these two tasks would, in the Communist view, assure both abundant production suitable to realize the Communist aim of distribution according to need and also the self-discipline of citizens necessary to make coercion unnecessary in preserving order. The police, the army, and courts would become unnecessary, and law would, in Marxist terminology, "wither away." Citizens would perform social obligations, expressed in morals and unsanctioned administrative regulations, out of conviction of their desirability rather than out of fear of punishment for their violation.

The new principles were embodied in the first constitution of the Russian Soviet Federated Socialist Republic (R.S.F.S.R.), promulgated July 10, 1918 (article 9), as follows:

> The basic task placed during the present transitional moment on the constitution of the R.S.F.S.R. is the establishment of the dictatorship of the city and village proletariat and of the poorest peasantry in the form of a powerful all-Russian Soviet authority with the objective of complete suppression of the *bourgeoisie*, the exploitation of man by man and the installation of socialism, under which there will be neither division into classes nor a State authority.

Early implementation. Lacking any precise pattern for a legal system designed to achieve the purposes stated in the first constitution, the new government issued only a few decrees designed to establish a framework for a new society and a primitive institutional structure to enforce it. The decrees deprived private individuals of the ownership of land, banks, insurance companies, merchant fleets, and large-scale industry in implementation of the policy of expropriation of the bourgeoisie, created restrictions on the employers of labour, and secularized marriage and divorce. The enforcement instruments, called "people's courts," operated without benefit of professional prosecutor or bar.

"Revolutionary conscience" as a guide

To permit regulation of the multitudinous social relationships for which no new law was prescribed, judges were directed at the start to apply Russian imperial laws, but only to the extent that they had not been revoked by the Revolution and were not contrary to the revolutionary conscience of the judges. Perhaps because Lenin and his colleagues were too busy to do otherwise, or perhaps because they preferred to let develop in practice a new legal order, no all-inclusive systematized body of law was prescribed until 1922. Local judges were guided during the first years, apart from the basic decrees indicated, only by suggestions issuing from the people's commissar of justice in the form of instructions, procedural manuals, and journal articles praising some court decisions as the proper application of revolutionary conscience and denouncing others as improper.

Political enemies of the Communists were not brought before the courts but were condemned by political bodies created at the same time as the people's courts and called revolutionary tribunals, or they were arrested and imprisoned without public hearings by the political police (Cheka). Lenin and his jurists professed only temporary resort to treatment of opponents outside the regular court system. Nevertheless, the conduct of revolutionary tribunals and of the Cheka greatly influenced the course of Soviet law long after their formal abolition in 1922, for they stimulated lack of respect for formality and strict adherence to law even among those charged with enforcing and administering the law.

New instrumentalities emerged after 1924 within the People's Commissariat of Internal Affairs (NKVD), now the Ministry of Internal Affairs (MVD), under the personal dictatorship of Lenin's successor, Joseph Stalin, to rid him of opponents. Following Stalin's death in March 1953 the Communist Party's first secretary, N.S. Khrushchev, disclosed that many innocent persons had been convicted and sentenced to long prison terms and even to death by special boards of the MVD. The Soviet method of preserving order was revealed officially to have been dual, with courts on one side dealing publicly with nonpolitical offenses and social disputes, and administrative boards and the security police punishing secretly on the other side those whom Stalin chose to call his enemies.

Codification. Reintroduction of private enterprise in limited areas and subject to strict controls followed the declaration, in 1921, of the New Economic Policy aimed to speed reconstruction to overcome the devastating effects of world and civil war. This reintroduction necessitated, in the Communist view, stabilization of law if capitalists were to be induced to invest. Lenin's commissar of justice, D.I. Kurskii, considered these measures retrogressive but necessary, while N.V. Krylenko, later to become federal commissar of justice, called them the natural evolution of five years of Soviet law. Whatever their origin, stabilization efforts took form in the first systematic codification, and a complex institutional framework of courts was constructed.

A judiciary act of October 31, 1922, established within the Russian republic a three-stepped system of general courts with civil and criminal jurisdiction.

1. The name "people's court" was retained for the lowest level, but the institution was less primitive than that of 1917. Its bench comprised a full-time judge appointed annually and two lay judges selected for service of a few days in rotation from a panel of intelligent and politically trustworthy citizens. Lay judges shared responsibility with the professional judge in deciding issues of law and fact.

2. Appeals from this court lay to a new provincial court evolved from the Congress of People's Judges, which, under the prior rules, had gathered periodically in each province to survey the work of the local courts. The provincial courts also received jurisdiction to try offenders against the security of the regime and other serious civil and criminal cases. As an appellate court the provincial court presented a bench of three professional judges. As a trial court it was composed like the people's court, except that the lay judges were selected from a panel of more experienced and politically sound citizens.

3. To coordinate policy in all provinces there was created the Supreme Court of the republic, evolving from a control department established earlier in the Commissariat of Justice. This court was authorized not only to hear appeals from cases tried in provincial courts but also to discipline lower courts, issue rulings interpreting the codes, and even to try cases of an unusually important character.

To provide specialized treatment of crimes relating to military matters and to the disruption of transport, then deemed critical to the success of the regime, special courts that had evolved earlier were continued by the judiciary act and subordinated to the Supreme Court. An office of public prosecutor and a college of defenders to perform

the functions of a bar were established to aid the judges, following experimentation with institutions designed to conduct prosecution and defense on a nonprofessional basis.

Substantive law and procedure to be applied by the new courts were established by codes enacted throughout 1922 and early in 1923, these being criminal, civil, family, land, and labour codes and also codes of criminal and civil procedure. Drafters of these codes followed patterns essentially similar to those of the Romanist states of the European continent, *i.e.*, all of the European states except the Scandinavian, to varying extents.

Patterns of legal institutions and substantive and procedural law enacted by the R.S.F.S.R. were copied with little variation in the other Soviet-type republics that had emerged in peripheral regions of what had been the Russian Empire. Federation of the Soviet republics into the Union of Soviet Socialist Republics on December 30, 1922, caused no change in the judicial system except the establishment, as a coordinator of all practice, of a Supreme Court of the U.S.S.R. The first federal constitution, adopted provisionally on July 6, 1923, and permanently on January 31, 1924, extended authority to the federal government to enact general principles of law to be followed by the republics in maintaining their codes. While this development might have caused a change in the Soviet legal system, it did not; the first federal judiciary act of October 29, 1924, confirmed the system of courts then existing in each of the republics, and no change was made in substantive or procedural law. The sole innovation was that the military and transport courts were placed under sole control of the Supreme Court; thus was created a self-contained system of federal courts, with lower branches functioning throughout the republics but beyond the reach of republic officials. A sharp line was thus drawn between courts dealing with matters vital to the security of the regime and those concerned with social disputes.

The pattern established for Soviet courts and codes of law in 1922 and 1923 continued until Stalin's death, although the second federal constitution of December 5, 1936, transferred authority to the federal level to enact codes of law to replace those of the republics. Although a second federal judiciary act was issued on August 16, 1938, new codes were not enacted. On February 11, 1957, the constitution was amended to restore the original relationship between republics and federation, namely that they enact their own codes but conform to general principles established federally. The first of the new general principles were enacted for criminal law and procedure on December 25, 1958; for civil law and procedure on December 8, 1961. Others were enacted at intervals thereafter, and republic codes were revised subsequently. The major innovation of the 1958 enactments was the requirement that punishment be ordered only by a court in accordance with the rules of the procedural code. This provision was heralded by Soviet jurists as preventing return to extralegal procedures such as those exercised by the MVD during Stalin's dictatorship.

APPLICATIONS OF SOVIET LAW IN PEOPLE'S DEMOCRACIES

Application of the pattern of law evolved within the U.S.S.R. to other areas began when Outer Mongolia declared establishment of its "people's government" on July 11, 1921. Mongolia was then a nomadic society without industry. Nationalization of productive resources had little effect on nomadic life, an exception to nationalization being made in the area of primary concern, namely the private ownership of cattle herds. Nevertheless, the base was laid for following Soviet legal patterns in the first constitution of the Mongolian People's Republic promulgated on November 26, 1924. Thereafter, state-owned industry was developed after Soviet patterns, and following adoption in the U.S.S.R. of the second federal constitution, a second Mongolian constitution was promulgated on June 30, 1940. Its provisions for courts and prosecutors were like those of the U.S.S.R., and the subsequently adopted law on land use of February 6, 1942, provided the same formula of state ownership of land allocated for use gratis and in perpetuity to the nomadic herdsmen for their privately owned cattle. A labour law of February 14, 1941, followed the Soviet pattern, as did a social insurance law of June 22, 1942.

The extension of Soviet law

The major opportunity to install the Soviet legal system outside the U.S.S.R. came only after World War II, when Soviet-type governments were created for European states occupied by the Soviet army in pursuit of retreating German forces. The system was also established in China, North Korea, and North Vietnam as military victories placed the Communists in power. In each country there was some variation, related to the degree of economic development, from the Soviet pattern evolved in the U.S.S.R.

Throughout all the states that became Communist, law was given the functions previously established in the U.S.S.R. Thus, the constitution of the Hungarian People's Republic of August 20, 1949, declared (article 41):

> Courts of the Hungarian People's Republic punish enemies of the working people, defend and secure the state, economic and social structure of the people's democracy, its offices and the rights of the toilers, educate the toilers in the spirit of observance of the rules of socialist intercourse.

Regulations governing the organization of the people's courts in the People's Republic of China, promulgated September 3, 1951, specified that courts were "to consolidate the people's democratic dictatorship, uphold the new democratic social order and safeguard the fruit of the people's revolution." The People's Republic of Bulgaria was even more outspoken than the U.S.S.R. in clarifying the function of law; in its constitution of December 4, 1947, it granted the right to citizens to create associations, but (article 87) only "if they are not directed against state and public order established by the present constitution," and then stated:

> It is forbidden and will be punished by law to form organizations having as their purpose taking from the Bulgarian people the rights and freedoms won by the people's uprising of September 9, 1944, and guaranteed by the present constitution, or to limit these rights and freedoms, to place under threat the national independence and state sovereignty of the country or to propagate open or concealed fascist or antidemocratic ideology, or to facilitate imperialist aggression, and also to participate in these organizations.

The Chinese Communist Party leader, Mao Tse-tung, in his 1949 speech "On People's Democratic Dictatorship," indicated that he had adopted in effect what had been Stalin's interpretation, enunciated in 1930, of the Marxist concept of the withering away of the state. Mao declared that there could be no withering in China as long as there was opposition to the state. The constitution of the People's Republic of China, promulgated on September 20, 1954, nearly five years after proclamation of the people's republic on October 1, 1949, fixed as final the system of courts evolved in emulation of the U.S.S.R. There were to be lay assessors, full-time professional judges, Soviet-type prosecutors, and three levels of courts; "judges were independent and subject only to the law," as they had been declared to be in the U.S.S.R. constitution.

Judicial patterns of the same type were established in the constitutions of the other people's democracies: the Democratic Republic of Vietnam (November 8, 1946), the People's Republic of Albania (March 14, 1946), Socialist Republic of Romania (April 13, 1948, and September 24, 1952), the Republic of Czechoslovakia (May 9, 1948), the Democratic People's Republic of Korea (September 8, 1948), the German Democratic Republic (October 7, 1949), and the Polish People's Republic (July 22, 1952).

The applicable law was stated in a small number of specific decrees in each republic, but otherwise determination of law was left to the judges. The Chinese government, among its first actions, abrogated all legal codes of the prior government. In October 1950 the Chinese premier, Chou En-lai, listing the documents to serve as guides to the judges, included not only specialized decrees of the government but also the Common Program of the Chinese Communist Party.

The provisional regulations on people's courts of 1951

(article 4) read, "Where no regulation is applicable, the policy of the Central People's Government will be adhered to." A law codification commission was established in the spring of 1950, but no basic codes emerged, only specialized statutes such as the land reform law of June 28, 1950; the law on marriage, April 13, 1950; the law on questions of employment, July 25, 1952; and the organic law for the court system, September 21, 1954.

The Soviet-type states in eastern Europe moved more rapidly toward codification than had China, and in contrast confirmed the existence of prewar civil codes, although with modifications. Poland adopted a General Statute on Civil Law with 119 articles on July 18, 1950. As in the Soviet Civil Code of 1922, the codifiers sought to avoid possible injury to working-class elements; article 3 provided: "It is forbidden to use one's rights to violate the fundamentals of social intercourse of the people's government." A Polish family code was adopted on June 27, 1950, and a decree on planned contracts on April 26, 1950.

Before the expulsion of the Yugoslav Communist Party from the Cominform on June 28, 1948, Yugoslavia, under the constitution of January 31, 1946, had followed the Soviet legal pattern. Under Tito, Yugoslavia accepted the concept of monopoly of power in a Communist Party and the necessity of state ownership of productive resources but refused to accept dictation from the Soviet Union on every detail, especially on centralization of industrial control and the speed of collectivization. The Yugoslav government experimented with bringing workmen into the managerial process of public corporations and in broadened autonomy for local government councils. Yugoslav legal philosophers rejected the view that the state could not begin to wither away until full achievement of Communism; they wanted to start relaxation of coercion immediately. Stalin's attempts to unseat Tito failed, and on Stalin's death his successors attempted reconciliation for a time, indicating acceptance of the idea that there could be many roads to Socialism.

Changes after the death of Stalin

Stalin's death in 1953 and his downgrading by the Communist Party of the Soviet Union at its 20th congress in 1956 opened a new era in the law of Communist-led states. The People's Democracies sought to be relieved of Soviet tutelage. Risings in Hungary and Poland forced the Soviet leaders to acquiesce in the evolution of divergent legal systems. Polish jurists, including Communists, demanded humanistic attitudes toward legal procedure with greater flexibility and an increased regard for the individual, and other east European states followed their lead. Even Soviet legal philosophers in 1958 endorsed a new humanism and declared it always to have been fundamental to Lenin's thought.

Only Chinese Communists resisted the trend toward strengthened legal procedures, denouncing Soviet policies of humanism and legality in an open letter published in 1963. After 1957 Chinese leaders abandoned their policy of copying Soviet legal patterns and ceased efforts to draft codes of law, declaring that they wanted flexibility in court. They evolved a legal system seemingly inspired by relics of traditional Chinese attitudes as modified by their own guerrilla experience before 1949 and by their study of Soviet practices during the period of War Communism before 1921.

Concurrently with revival of concern for humanistic values, Communist states other than China revised their constitutions and law codes to give formal support to their claim to have passed beyond capitalism to Socialism. Mongolia and Czechoslovakia declared themselves Socialist republics in constitutions adopted July 6 and 11, 1960, respectively. North Vietnam introduced its new constitution on December 31, 1959. Yugoslavia promulgated its revision on April 7, 1963; Romania on August 20, 1965; and the German Democratic Republic on April 9, 1968. New civil codes designed for Socialist economies were introduced in Poland and Czechoslovakia in 1964, the latter incorporating terminology that departed sharply from Romanist tradition.

Several features, however, remained unchanged in the law of these states even after the introduction of variations. Most notable was continuation of Lenin's principle that no basis remained after the Russian Revolution for the Romanist division of law into public and private spheres. To Lenin the Soviet state's concern with every detail of social intercourse required maintenance of the state's right and opportunity of intervention in any matter at any time. Consequently he held that in the Soviet system all law must be public, that is, reflect the state's vital concern in the social relationships governed by law. This attitude of state direction of all matters, sometimes called *dirigisme* by French scholars, remained characteristic of Soviet law.

CHARACTER OF SOVIET AND SOCIALIST LAW

Montesquieu's concept of separation of powers as embodied in the Constitution of the United States of America was rejected by the founders of the Soviet legal system. The legislature was made supreme over the executive and the judiciary. In principle only the legislature's enactments were a source of law.

Practice made the executive a source of law, however, for the Council of Ministers often acted without convening the legislature. The second federal U.S.S.R. constitution of 1936 attempted to stop this practice in violation of the constitution of 1924, but the executive soon again usurped legislative authority. The post-Stalin reforms again sought to restore monopoly lawmaking power to the legislature. In practice it was not the legislature in full assembly but its derivative body, the presidium, elected from its membership to legislate during intervals between full sessions, that created the day-to-day changes in law. Although subsequent ratification from the full body was sought, revocation of presidium action was virtually impossible because of the passage of time. The presidium thus became the most important source of law, except for the state economic plan, the budget, and enactment of general principles for republic codes.

Legislative domination

Communist Party orders were in the technical sense not a source of law, for to be binding on courts these had to be enacted by state agencies. Still, the party provided the initiative for legislative action. Some vital laws, such as those reforming agriculture or industry, were given preeminent status by signature by the party secretary as well as by the chairman of the presidium when they were published as law.

Having rejected judicial lawmaking with assertion of the legislature as the sole lawgiver, Soviet jurists accepted no concept of judicial precedent like that of the Anglo-U.S. common law. Nevertheless, the Supreme Court of the U.S.S.R., in exercise of its authority to enunciate principles of interpretation based upon examination by its full body of a course of judicial practice proclaimed such far-reaching principles as to make of these a source of law. Even individual decisions of the Supreme Court, although lacking the authoritative force of the general rulings, were given careful attention by lower courts, since they were published as guides. The Soviet legal system in practice adopted much the same approach to judicial decisions as have Romanist systems generally.

The Supreme Court of the U.S.S.R., being subordinate to the legislature, had no right to declare a law unconstitutional nor to create any limitation upon the legislature or the executive on the ground that some higher or general principles of law had been violated.

Control over the judiciary was preserved for the legislature at all levels above the people's court by giving it appointive and recall power, through the Soviet operating at the court's level. Only people's judges were elected by the people under provisions introduced by the 1936 constitution, although elections were not begun until after World War II. Elections followed the pattern established for all representative bodies in the U.S.S.R. by Stalin, namely the one-candidate ballot, the candidate being chosen by local professional groups with the guidance of the Communist Party members in the groups. In practice, judicial decisions had to accord generally with Communist Party policy or the judge risked recall.

In light of the foregoing, the constitutional provision that "judges are independent and subject only to the law"

(article 112) had to be understood as denying to local officials the right of intervention in the formation of an individual decision on personal grounds, but as not prohibiting Communist Party intervention when such intervention was properly formulated in criticism of a line of decisions out of keeping with party wishes.

Judges were not deprived entirely of initiative by the rule of legislative supremacy or by the leading role played by the Communist Party. In both the Civil and Criminal codes adopted in 1922 there was authorization to depart from the rigid provisions of the code under special circumstances. The Civil Code opened with article 1, reading, "Civil rights are protected by law except in cases where they are exercised contrary to their social and economic purpose." This rule permitted judges to apply their concept of state policy, when it was not specifically stated, to achieve an appropriate result in a given situation, but the decision created no rule of law for future cases. The article was used frequently to justify judicial refusal to protect the rights of capitalists during the period of reconstruction (1922–28), but after 1930 it fell into disuse. Soviet jurists argued for its deletion from the code as unrelated to a stable society from which all capitalist elements had been eradicated, but it was retained in the civil-law general principles of 1961 (article 5).

The Criminal Code of 1922, and as revised in 1926, permitted the judge to exercise discretion not only in withholding conviction where a crime had clearly been committed, provided that he found no social danger, but also in punishing a citizen who had committed no crime defined as such by the code. The formula for such judicial initiative was that the act be found socially dangerous in a manner analogous to another act specifically defined as a crime. This provision, which led to unpredictability, was eliminated from the criminal law by the general principles adopted in December 1958.

SUBSTANTIVE LEGAL PROVISIONS

Soviet jurists and their colleagues in eastern Europe and the Communist-led states of Asia claimed uniqueness for their system because of its constitutionally defined function of promoting Socialism. The end, in their view, transformed the means. While the novelty of the system was, therefore, primarily in its philosophical base, the substantive law was revised to introduce some novel institutions.

Property. The *Communist Manifesto*'s interpretation of property ownership as the source of political power was accepted by all Communist parties as the guiding principle to be followed in revision of law. Consequently, private ownership of productive wealth was everywhere abolished in some degree as soon as state officials felt competent to administer production efficiently. The second federal constitution of the U.S.S.R. (1936) set forth the ideal, denying to private individuals the right to employ labour in productive enterprise (article 9) and reaffirming Lenin's decree of February 19, 1918, nationalizing all land (article 6).

Lesser achievement of the goal was realized elsewhere. Polish and Yugoslav Communists continued to permit peasants to own farmlands and also allowed private citizens to conduct service trades, although under limitation on the number of employees. China alone permitted private ownership of factories to continue, but subject to the limitation that owners retain no financial link to foreign capitalists and share their ownership with the state as partner. Profits to such owners became a fixed return on investment, indicative more of loan capital than equity, and even this payment was often regained by the state on demand for various fund-raising campaigns for social organizations.

The various legal codes reflected the preponderant importance given state-owned productive property. For example, more severe penalties for theft of state property than of private property were prescribed by criminal codes. The U.S.S.R. labour codes were concerned solely with state employment. The provisions of civil codes relating to sales, contracts, property damage and inheritance governed only consumer goods in traditional Romanist legal fashion, while new forms were developed to govern transactions in state property. Title to state property was subject to restrictions: no sales or leases, with limited exceptions for land allocated to peasant households disrupted temporarily by military service or departure for temporary industrial employment.

The ownership of land and personal property

Land, although declared state property only in the U.S.S.R. and the Mongolian People's Republic (constitution of 1924), was brought increasingly, in all Marxian Socialist states except Poland and Yugoslavia, under the administration of cooperative-type associations structured on the Soviet collective farm model. Peasants were induced, often under strong pressure, to assign to these the use of plots originally owned by or assigned to them. The variation introduced in China (1958) was called a "commune" and united both local government and land administration in one administrative unit supervising "brigades" (groups of villages) and their subordinate "teams" (individual villages). By law, land ownership in China was placed not in the state as in the U.S.S.R. but in the "team" to stimulate peasant initiative in its development.

The Soviet model collective farm placed ownership of machinery, tools, commonly used farm buildings, and crops in the cooperative; but member families were declared owners of their family homestead, including hand tools and barnyard animals. Small plots of state-owned land were assigned each family for private use as a household garden. Farm government was carried on by a membership assembly that elected management, usually introduced to the members by central government authorities. Remuneration was in a share of the produce graduated according to value of work performed computed on an accounting unit called a "labour day."

Agricultural cooperatives were generally regarded as transitional forms designed to guide peasants toward a superior type of organization patterned on industry so that peasants would become farm employees receiving wages and social insurance benefits. Only China rejected this ultimate goal, preferring to perfect the "commune" as a model for rural productive units.

Nonproductive property was left to private ownership as were artisans' tools, all being labelled "personal property" to distinguish them from income-producing "private property." Early efforts to equalize ownership of consumer goods were abandoned by Stalin (1930), and egalitarianism was denounced as a petty bourgeois utopian idea. Shortage of goods forced a change in attitude. Wages differentiated according to work performed became the means of stimulating production, a differentiation that led progressively to weakening of restraints on inheritance and ownership of luxury items.

The second federal constitution of the U.S.S.R. (1936) entrenched the property incentive system by guaranteeing private ownership of savings accounts and inheritance. To hamper reversion to capitalist practices, various restrictions were placed on use of personal property. Thus, dwellings were limited in size to satisfy only family needs; sales of dwellings and apartments could occur only at long intervals; members of a family sharing a joint household might own only one dwelling and might lease excess space only temporarily and occasionally and at rentals not exceeding those charged for occupancy of state-owned premises. Sales of privately owned automobiles and motorcycles could be effected only through state commission stores at fixed prices. Property purchased with income gained illegally could be expropriated. Neglected property might be expropriated if repairs were not made following warning.

The determination to eliminate income from investment of private funds weakened under various paramount state pressures. Thus, interest was paid on state savings bank accounts to stimulate deposits as a means of accumulating funds for state investment and to reduce the threat of an inflation of currency. Investment in state bonds also was encouraged, but returns were paid not in conventional interest but as lottery winnings. Inheritance from abroad was paid to heirs within the U.S.S.R. Sums received through such channels were used to purchase dwellings, apartments in cooperatively constructed structures, expensive consumer goods, and tours.

Marxian Socialist states outside the U.S.S.R. established comparable restrictions: Czechoslovakia going furthest to require that the source of personal ownership be "honest," that is, derived mainly from work for the benefit of society. China, under Mao Tse-tung, took the strict position, opposing extensive use of property incentives among individuals and espousing asceticism.

Communal ownership was reserved in Marxian Socialist states, except China, for the distant future when abundance should have been achieved, money would no longer be used as a medium of exchange, and the state had "withered away," in accordance with the prophecy of Marx and Engels in the *Communist Manifesto*.

Meanwhile, the state represented the community and sought to enlarge its segment of productive property. Yugoslav Communists attempted to create a variation, rejecting the Soviet model of state ownership of industry to create in 1950 "social ownership" of factories, under which title to industrial property was conceived to be not in the state but in the plants' workmen, organized into a "workers' council." Management was appointed not by a central ministry, as elsewhere in the Marxian Socialist states, but by the "workers' council," acting in concert with local government and an industry-wide chamber representing all workers' councils in the republic in the same branch of industry. The aim was to overcome evils thought inherent in centralized bureaucracy.

Economic planning. Production was planned in all Marxian Socialist states, although in varying degree and through different types of institutions. Stalin favoured detailed planning and centrally made decisions by the Council of Ministers, advised by the State Planning Commission (Gosplan). The Yugoslav opposition to a centralized bureaucracy placed the emphasis on planning by local government and "workers' councils," subject to influences exerted by the central government through allocation of major capital investment and establishment of national goals.

Stalin's heirs experimented after 1957 with forms of decentralized decision making through regional economic councils (Sovnarkhoz) for industries not directly related to national defense, hoping to stimulate managerial initiative. Most central industrial ministries were abolished from 1957 to 1965, but Gosplan and specialized financial and technical institutions retained control over allocation of key resources. Return to centralization with abolition of the Sovnarkhoz and reconstitution of industrial ministries in 1965 was in a form differing from Stalin's model in that responsibility of republic authorities was heightened and plant managers were directed to search out local sources of supply and markets.

Encouragement of local initiative went farther in some Marxian Socialist states, Czechoslovak specialists even proposing to dramatize the change in emphasis by declaring the public corporations administering state-owned property as the "owners" of the buildings and tools rather than only the administrators. This extreme experimentation was terminated by Soviet intervention in 1968, although the Yugoslav Communists continued to rely on their system of local ownership and direction subject to indicative plans having no force of law.

Contracts as tools for economic planning

An interesting legal feature connected with production in all Marxian Socialist states was the contracts executed by plant managers for the implementation of production plans. These contracts prescribed quantity, design, technical specifications, price, and delivery date, using, as a basis for negotiation, plan requirements. Increasingly, after 1957, elements of consumer choice were considered. These contracts served also as the basis for dispute settlement by permanently functioning state tribunals called State Arbitration. These tribunals might also require execution of contracts to implement the plan if parties could not agree on terms. China alone provided a variation, establishing no permanent tribunal but preferring mediation through ad hoc committees established by local economic commissions linked to the planning apparatus.

Labour. While production and distribution occurred under compulsory orders, labour remained free to move as it wished, except during the emergency conditions of World War II in the U.S.S.R. Subsequently, the Soviet labour supply was again regulated by monetary and other inducements but subject always to centralized determination of job classification, hours and wages, and in China to vigorous political pressure designed to force urban dwellers back to the farms.

The role of collective labour agreements changed as the central authorities progressively prescribed details of employment. They fell into disuse in the mid-1930s but were restored after World War II to mobilize the labour force for production. Since job classification, wages, and hours were established by law, the agreements concentrated on betterment of local working conditions and indicated the labour union's preferences regarding the use of bonuses collectively earned as incentives for exceeding the plan. Collective bargaining was expected to reappear in the U.S.S.R. in more traditional form after the post-1965 reforms as plant managers exercised their new authority to reduce unneeded staffs and rearrange wage scales to reflect heightened productivity.

Strikes and grievances

Since production was planned and, therefore, was by definition rational, strikes were given no place in Marxian Socialist states. Remedies sought by management or labour unions against apparent injustice had to take the form of petitions to higher authority for legislative or administrative rectification. Strikes did occur in Poland and Hungary in 1956 when Communist influence weakened, and reforms became necessary to hold power. Collective agreements were enforced primarily through moral suasion, strengthened by penal code provisions applicable to state managers who violated the code, the collective agreement, or legal activity of labour unions, and to labour unions who disrupted production so severely as to commit the crime of "wrecking."

The "right to work" was established by all constitutions of Marxian Socialist states, but no system implemented it by court order on an employment office to provide a job. The concept came to require only elimination of discrimination on the basis of race, sex, age, or domicile in filling such jobs as might exist.

Individual employment contracts reflected statutory provisions governing the employment relationship and its termination. Since individual bargaining in establishing employment terms was eliminated, written contracts were often omitted in practice. A grievance procedure under labour union supervision was established by which aggrieved parties could appeal to courts for review of the reasonableness of managerial decisions on payment and dismissal controversies. Labour becoming redundant with automation's advance was to be retrained for other employment by state centres.

Yugoslavia's "workers' self-government" presented a variation in the theory of the employment relationship since workers were said to employ themselves, being both owners and workers. Their activity was regarded as a social duty rather than a labour relationship, but in practice standards of employment were established to which management was required to adhere, these being generally the same as in other Marxian Socialist states.

Accident insurance. Social insurance was introduced in all Marxian Socialist states to protect employees injured on the job or elsewhere. No coverage was offered collective farmers until 1966, nor to housewives and children, forcing them to rely for recovery on suits based on the Civil Code's provisions on obligations. These rules conformed to traditional Romanist legal models with some variation. Thus, the first Russian Civil Code (1922) required no fault for recovery, but this provision became conventional through court practice. Fault became necessary for liability except when the instrument causing injury was classified as extra hazardous. The list of such instruments included motor cars.

The social utility of negligence suits

Social insurance benefits were originally inadequate because of the impoverishment of the state treasury. Consequently, victims of accident were stimulated to sue for the difference between social insurance payments and total losses. Soviet policy makers finally concluded it desirable to continue this practice even when social insurance funds were increased, the argument being that state managers

would be prompted to exercise care because of the threat of suits, which would be likely to reduce the profit margin and their eventual chances for promotion. Social insurance benefits thus continued to be set at levels lower than damages suffered.

Damages were limited to loss of wages or anticipated wages for unemployed victims. Poland and Czechoslovakia continued ancient practice to allow recovery for pain and suffering additionally. No capitalization of damages was permitted, in part to avoid transfer to irresponsible individuals of large sums and in part to discourage accumulation of wealth likely to provoke speculation.

No personal liability insurance was made available to individuals in the U.S.S.R., except to foreigners, in the belief that its availability would reduce the deterrent effect created by civil suit for accident. Czechoslovak and Hungarian civil codes retained pre-World War II provisions, however, creating the opportunity of purchasing liability insurance. State insurance provides state enterprises in all Marxian Socialist states with protection from suit by employees injured by extra hazardous equipment, unless charges of fault are made. The burden of proving absence of fault is on the defendant.

Liability for injury was attached to state activity even when not of an economic character, as with ministries and local government, although until 1961 in the U.S.S.R. sovereign immunity for governmental functions had to be waived explicitly by statute to permit suit. Thereafter, liability was extended to all types of state administration unless immunity was claimed by specific law. The Czechoslovak Civil Code (1964) established a new approach by drawing no distinction between types of state activity in establishing a rule of liability to suit.

The Chinese penchant for mediation over litigation developed after 1957 to eliminate possibilities of suit for injury in accident, but in cases of reprehensible conduct the police were authorized to impose administrative sanctions for minor infringements of citizens' personal or property rights, or require payment of medical expenses and property damage under the Security Administration Punishment Act of 1957.

A duty to protect state property (article 131) and to observe rules of Socialist conduct (article 130) was created by the second U.S.S.R. federal constitution (1936). It initiated judicial practice establishing an obligation of a state enterprise whose property had been protected to pay damages to a victim injured while attempting to perform his duty. This principle was written into the U.S.S.R. Fundamentals of Civil Law in 1961. No implementation was provided, however, for failure to rescue another in performance of the obligation to observe the rules of social conduct. Soviet authors concluded that this duty was moral and not legal, and violation could create no civil liability in favour of a citizen or his heirs if no rescue was attempted.

The Yugoslav constitution (1963) likewise created a duty to rescue but no formal obligation to pay damages to persons injured as the result of another's failure to act in their protection. The Czechoslovak Civil Code (1964) was the first to establish not only the duty to rescue but also a right in the injured party to recover damages from an individual who failed to perform his duty. It also established the right of the rescuer to recover from the person in whose interest he acted such damages as he might have suffered.

Artistic creation and invention. The legal rights protecting authors, artists, and inventors presented problems in socialized economies unless such persons were state employed. Practice indicated that authors, artists, and inventors were most imaginative when unimpeded by state supervision. To provide the requisite freedom coupled with stimulation of initiative, Soviet codifiers created a status for innovators midway between Western concepts and those of the Socialist ideal of production for wages. This status was called "a law of personal relationships rather than of property relationships."

Inventors in the U.S.S.R. were offered two alternatives: (1) an "author's certificate" or (2) a patent. The first was granted after proof of novelty to the state patent office and qualified the inventor to receive payments computed on the saving to the state enterprise from use of the invention, the computation being based on the best of the first five years of use. Additional benefits were tax exemption on a portion of the receipts, priority in the allocation of scarce living space, and acceptance into a technical school. The patent form was traditional, except that restrictions on private enterprise limited its use to assignment to state enterprise. In practice only foreigners chose patents. The international convention protecting patents was ratified by the U.S.S.R. in 1965.

Although other Marxian Socialist states in Eastern Europe also offered a form similar to the "author's certificate," patents retained vitality until 1963 when the trend toward the Soviet model became marked. All states, however, retained the patent form for local inventions likely to have appeal abroad so as not to discourage potential foreign licensees unlikely to purchase a novel form of protection.

Protection of artists' rights

Authors and artists were provided more conventional protection than inventors, but they were deprived of some traditional rights of a copyright owner. They could not bargain for royalties nor sell their work to foreign publishers except through a state trading enterprise. Royalties were paid in accordance with a state tariff based on type and size of an edition. In addition to monetary rewards, payable for life and to heirs for a term of years, moral protection was offered in the form of a right to control editorial changes in a manuscript and its use. Exceptionally, community interests were exerted against an individual author's will if the work was regarded as a national treasure. Thus, it could be translated or published continuously after a first disclosure by the author. Royalties were paid, however, except when the language of translation was other than Russian on the ground that minority peoples needed culture for which they could not afford to pay.

The U.S.S.R. ratified no international copyright convention and has avoided the obligation to obtain consent and pay for publication of foreign authors within its boundaries. Other eastern European Marxian Socialist states, however, have accepted the Brussels Copyright Convention.

Nonmonetary inducements, including medals and honorific titles, were also utilized to stimulate innovation.

Domestic relations. Women's equality with man and secularization of marriage were the major aims of the Russian revolution in regard to the family. Other Marxian Socialist states, where these goals had not yet been achieved at the time of Communist assumption of power, had the same program. Before 1926 Soviet family law presented little variation from western European models, but the 1926 code introduced unregistered marriage and divorce. Marriage could be recognized as legally constituted if cohabitation as if in marriage were proved orally. Divorce would be registered on the request of one party on proof of termination of marital relationships.

A trend toward discouragement of promiscuity emerged in the mid–1930s on evidence of increasing juvenile delinquency resulting from the breakup of homes. A law of June 27, 1936, required both parties to appear for divorce if the care and maintenance of children was in issue and sought to discourage repeated divorces by establishing a system of graduated fees for each succeeding divorce. The 1926 provisions accepting proof of marriage or divorce other than by registration were finally revoked by law of July 8, 1944. Thereafter registration was compulsory to establish marriage, and courts were given sole jurisdiction over divorce. Judges were required, however, to attempt reconciliation and to grant divorce only when convinced that restoration of the family was impossible. Children born out of wedlock were denied the father's name and maintenance or inheritance. Unwed mothers were given state aid.

The compromise on marriage and family

Hostility to the 1944 system mounted with succeeding years. A compromise position was formulated and adopted by law of June 27, 1968. While registration of marriage remained a requirement, divorce might be registered without prior court decision if both parties consented and

there were no offspring. In other circumstances court proceedings were compulsory. The paternity suit was restored, but only to provide the father's name and maintenance rights if a child had been born from a union having all characteristics of marriage except registration. Offspring of casual unions remained under the regime established in 1944.

While other Marxian Socialist states generally followed the U.S.S.R. model, all introduced variation to preserve national cultural patterns. None but China adopted the U.S.S.R.'s early attitudes on unrestrained divorce, and even China discouraged resort to its simple procedure of granting divorce at the request of either party. All Marxian Socialist states sought to preserve the conjugal family against lighthearted decisions to separate. The formula was to establish no precise grounds for divorce but to authorize divorce only in the event of complete and permanent breakdown of the marital relationship. Yugoslav law prohibited suit for divorce by the spouse responsible for creating an unbearable situation, but in practice marriages were considered broken if cohabitation ceased. Generally children born out of wedlock were accorded the same rights as those born legitimately, with the exception indicated in the U.S.S.R.

The marital property regime established in the U.S.S.R. after early experimentation with individually owned property was established in 1926 as a system incorporating the concept of community of acquests; *i.e.*, property owned individually by the spouses before marriage remained their individual property, but that acquired during marriage was communally owned, unless received by one spouse as a gift from outside or by inheritance. The U.S.S.R. model was adopted elsewhere among the Marxian Socialist states.

Criminal law. Novel definitions of crime emerged in Marxian Socialist states to reflect Socialist principles. To the traditional crimes against the person and property retained from the definitions of pre-Socialist codes, in 1932 the U.S.S.R. added "economic crimes." This development coincided with the time when the private enterprise system, permitted during the period of the New Economic Policy, was terminated in favour of the system of state monopoly of large scale production and state planning. Under these altered priorities, merchandising became a crime. The 1936 U.S.S.R. constitutional prohibition of all private employment of labour for production purposes was reflected in the Criminal Code to penalize employers. Even artisans employing no labour were forbidden to engage in various production activities unless customers provided the raw materials. Later, leasing of unused space in small private dwellings at rates exceeding those established for tenants of state-owned apartment buildings became criminal.

Crimes against the state

Fear of opposition to Communist Party leadership caused the introduction by draftsmen in the Russian republic's Code of 1922 of a chapter on "counter-revolutionary" crime. It remained in the successor code of 1926 and was replaced by a chapter on "state crime" by a law of December 25, 1958. While all legal systems punish attempts to subvert state authority, the definitions and practice under Stalin exceeded the norms of liberal Western democracies in vagueness of expression and severity of penalties prescribed. Further, in keeping with the principle established by the Russian republic's constitution of 1918 and continued in subsequent constitutions that limited the exercise of civil rights to activities in the interest of Socialism, prosecutions for anti-Soviet agitation occurred under circumstances that elsewhere would not have constituted criminal sedition. After Stalin's death there was a decade of lessened severity, but a new epoch was initiated in 1966 when authors were prosecuted for publication abroad of allegorical manuscripts that had been rejected by Soviet censors. The ground given for this prosecution was that the authors should have anticipated the hostile use to be made of their writings by enemies of the U.S.S.R.

In implementation of Marxist attitudes toward religion as an intolerable superstition to be discouraged, the conduct of religious schools and of social activity unrelated to worship was prohibited in the U.S.S.R. Strong discouraging measures adopted during the 1920s when the religious hierarchies opposed Communist programs were relaxed during World War II to gain peasant support, but basic policies remained unchanged. An order of March 18, 1966, prescribed punishment for religious activity deemed harmful to health, and in practice was interpreted broadly to hamper worship by minority sects.

While other Marxian Socialist states introduced less restrictive laws, their leaders sought to discourage religion. Chinese Communists expressed belief in re-education of monks and priests. Eastern European leaders were initially militant in prosecuting clergy but relaxed their vigilance as religious institutions became more tractable, Yugoslavia even going so far as to sign a concordat with the Vatican in 1966.

Procedural law. Romanist influence was strong in the codes of criminal and civil procedure, which copied the main features of the French codes, although retaining more possibilities of control by the state.

Emphasis in criminal procedure was placed on investigation before trial, as in French practice. The examiner had to establish more than a prima facie case required of the grand jury in the traditional common-law system. The suspect had to be given an opportunity to testify and to produce witnesses and evidence in his defense. The examiner became a preliminary judge authorized to bring the case to trial only if he became convinced of guilt, so that the public trial consisted primarily of verifying the examiner's work rather than hearing the defendant's case for the first time. Contrast with the French model was provided by the fact that the examiner was not a magistrate but subordinate to the office of prosecutor, and until 1958 no attorney was permitted the defendant in the U.S.S.R. during the preliminary investigation. Even thereafter only juveniles were given an absolute right to an attorney.

The "inquisitorial" system

Trial procedure followed the French model closely in that the judges were instructed to establish their personal conviction of guilt. To do so they might go beyond evidence presented by the prosecution and defense. The court might call witnesses, seek its own experts, examine material evidence, and visit places connected with the crime. The judges did most of the interrogating in the courtroom, relying on the prosecutor and the defense only to bring out points overlooked. No rule of evidence of any kind bound the court, except the rule of relevance.

Civil procedure required observance of rules of evidence only for certain formalities established by the civil codes, such as those for such documents as contracts or wills. Otherwise, the judges were as free in civil as in criminal matters to seek or hear what they desired to establish their personal conviction. Soviet jurists claimed an innovation in the prosecutor's authority to intervene in a civil case at any stage when thought necessary for the interests of society. This power sometimes brought in the prosecutor to protect a party whose attorney appeared incompetent.

Appellate courts finding violation of procedural requirements of substantive law or unconvincing evidence were required to remand for new trial. Revision of a sentence or of a decision without retrial in a court of original jurisdiction was permitted only when the original court had no jurisdiction, where an amnesty had been improperly applied, or where there had been no basis for trial. A penalty could not be increased on defendant's appeal.

A singular feature of Soviet procedure was the authority given to Supreme Court presidents and to the prosecutors of the republics and of the U.S.S.R. to protest a decision, whether civil or criminal, after it had become final on appeal or in absence of appeal. This procedure permitted reopening of convictions or even of acquittals if the highest legal authorities found grave error in their periodic audit of inferior-court activities. Prosecutors frequently requested such reopenings, even of convictions obtained by inferior prosecutors. This created the opportunity of petition to superior prosecutors or to court presidents after all remedies as of right had been exhausted. At the same time it gave state authorities a second chance, not given the convict or the party to a civil suit, to change an otherwise final decision.

Soviet law in perspective. The industrialization characteristic of the 20th century required legal systems in all states to accept increasing state intervention in social relationships. Employers were no longer left free to execute labour contracts as they wished; property owners were subjected to increasing restraints on use and in some countries were subjected even to the loss of farms or dwellings if these were not used productively; nationalization in banking, industry, and transport was widespread; compulsory insurance against injury to third parties became commonplace, as did social insurance against industrial accidents; and secularization of marriage and divorce nearly everywhere prevented citizens from conducting their marital relationships solely within religious rules.

Soviet and Socialist law, in its reflection of the state's pervading presence in human relationships, appeared to be but an extreme extension of trends common to legal systems everywhere. Yet many legal philosophers concluded that Soviet law differed in more than degree from other systems. While its techniques were those of the Romanist system, its difference in degree was interpreted as a difference in kind, justifying the conclusion that Soviet law was a unique system requiring separate categorization.

BIBLIOGRAPHY. H.J. BERMAN, *Justice in the U.S.S.R.*, rev. ed. (1963), an introduction to Russian legal history and Marxist philosophy to measure their impact upon the evolution of Soviet law; A.P. BLAUSTEIN (ed.), *Fundamental Legal Documents of Communist China* (1962), translations of principal statutes and codes; V.M. CHKHIKVADZE (ed.), *The Soviet State and Law* (1969), an elementary text prepared by Soviet law professors to introduce foreigners to fundamentals of Soviet legal philosophy, state structure, and legal procedure; J.A. COHEN, *The Criminal Process in the People's Republic of China 1949–1963: An Introduction* (1968), translations of statutes, reports of interviews with refugees as to practice, and the editor's explanation of the key elements in the Chinese Communist approach to law; R. DAVID and J.E.C. BRIERLY, *Major Legal Systems in the World Today* (1968), evaluation of the Marxian socialist system of law in comparison with other legal systems; G. FEIFER, *Justice in Moscow* (1964), reports of cases observed by an American student in Soviet courtrooms in the 1960s, light but informative; V. GSOVSKI and K. GRZYBOWSKI, *Government, Law and Courts in the Soviet Union and Eastern Europe*, 2 vols. (1959), a symposium by legal specialists treating all branches of Marxian socialist law, extensively documented; J. HALASZ (ed.), *Socialist Concept of Human Rights* (1966), an Eastern European exposition of Marxist approaches to the philosophy of human rights and the protection accorded them; J.N. HAZARD, *Communists and Their Law* (1969), comparison of all systems within the Marxian socialist legal family to determine the common core, much judicial practice included, and *Settling Disputes in Soviet Society* (1960), a history of the evolution of Soviet legal institutions with emphasis on policy debates and practice during the formative years; M. JAWORSKYJ (ed.), *Soviet Political Thought: An Anthology, 1917–1961* (1967); S. LUBY (ed.), *Le droit civil tchécoslovaque* (1969), a symposium by Czechoslovak jurists on matters treated traditionally in Romanist civil codes, explaining theory and statutory provisions; G.G. MORGAN, *Soviet Administrative Legality* (1962), examination of the function of the Procurator General of the U.S.S.R. as a channel for protection of the constitution and the laws; S. ROZMARYN (ed.), *Introduction à l'étude du droit polonais* (1967), a symposium by Polish jurists on various branches of Polish law, brief but informative; SHAO-CHUAN LENG, *Justice in Communist China* (1967), analysis of doctrinal writing, statutes, and newspaper accounts to determine how much Chinese legal tradition remains in Marxist-oriented China, succinct but thorough; J.F. TRISKA (ed.), *Constitutions of the Communist Party States* (1968), translations with historical commentary of constitutional documents of all 14 Marxian socialist states; A.Y. VYSHINSKY (ed.), *The Law of the Soviet State* (1948; orig. pub. in Russian, 1938), a treatise for Soviet law students by Stalin's spokesman stating fundamental principles espoused during Stalin's purges.

(J.N.H.)

Soviet Union

By area the world's largest country, the Union of Soviet Socialist Republics (U.S.S.R., or, popularly, the Soviet Union) covers a vast segment of the Northern Hemisphere: one-sixth of the world land surface and two-fifths of the Eurasian continent. Its 8,650,000 square miles (22,400,000 square kilometres) are seven times the area of India, two and one-half times that of the United States, and 60 times that of Japan. It covers 11 of the 24 world time zones; while citizens of the Soviet Far East are greeting the dawn, their fellows in Moscow, the capital, are sitting down to dinner. The total population reached an estimated 257,824,000 in 1977.

Neighbouring countries

The Soviet Union has the world's longest frontiers, and no other country is bounded by so many nations. Norway, Finland, Poland, Czechoslovakia, Hungary, and Romania lie to the west; Turkey, Iran, Afghanistan, China, Mongolia, and North Korea lie to the south; the eastern and northern frontiers are coastal. Conventional lines linking the Rybachy Peninsula and Big Diomede Island to the North Pole form the boundaries of the Soviet sector of the Arctic. Twelve seas, belonging to three oceans (Arctic, Atlantic, and Pacific), wash the country's coast.

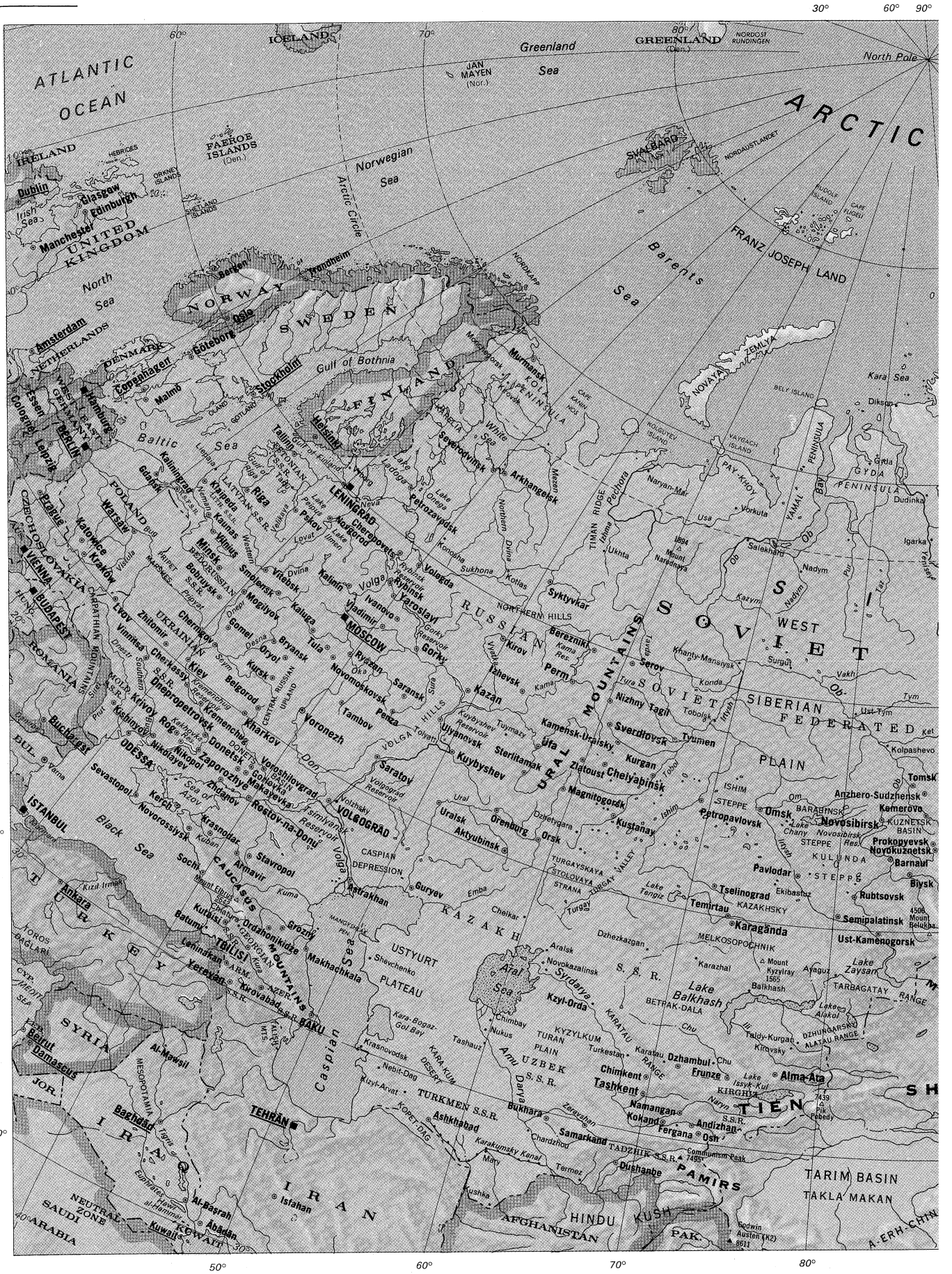
The Soviet Union was formed on December 30, 1922, and it now comprises, in addition to the 15 union republics, 20 autonomous republics, 8 autonomous regions, 10 autonomous districts, 6 territories, and 121 regions. Under the constitution adopted October 3, 1977, the political foundation of the Soviet Union is formed by the Soviets (Councils) of People's Deputies, with the Communist Party of the Soviet Union (CPSU) the "leading and guiding force of Soviet society and the nucleus of its political system." The economic foundation is "Socialist ownership of the means of production."

The varied terrain of the Soviet Union harbours a wealth of raw material, fuel, and power resources. The Soviet period has seen the accelerated development of these resources and the emergence of the country as an advanced industrial power. The proportion of the population living in cities has risen from 18 to more than 60 percent, and the population as a whole is growing at the rate of some 2,300,000 a year.

The following article offers a survey of the contemporary Soviet Union. It is divided into the following major sections and subsections:

I. The land
  The natural landscape
  The traditional regions
  Settlement patterns
II. The people
  National composition
  Social characteristics
  Demographic trends
III. The national economy
  The Soviet economy in world context
  The extent and distribution of resources
  Sources of national income
  Management of the economy
IV. Transportation
  The Soviet system in its national context
  The component systems
V. Administration and social conditions
  State and society
  The judiciary, armed forces, and public order
  The system of public organizations
  Social conditions and services
VI. Cultural life and institutions
  The cultural milieu
  The individual arts
  Cultural and educational institutions

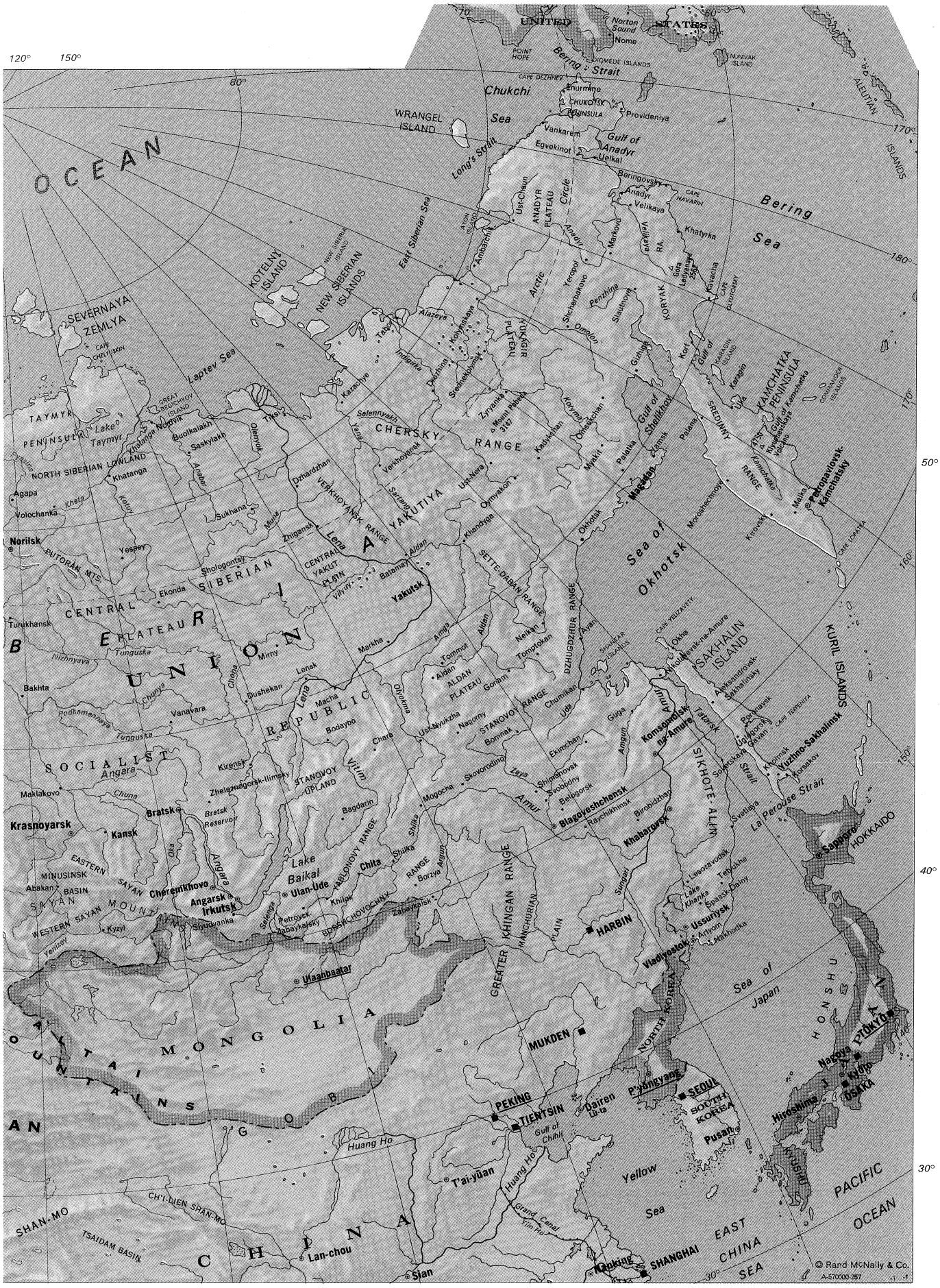
Historical aspects are covered in the article RUSSIA AND THE SOVIET UNION, HISTORY OF; and additional historical information may be found in BALTIC STATES, HISTORY OF THE; EUROPEAN DIPLOMACY AND WARS (*c.* 1500–1914); INNER ASIA, HISTORY OF; MONGOLS; SCYTHIANS; and URARTU AND ARMENIA, HISTORY OF. Geological descriptions and information on the natural environment are also dealt with in the articles EUROPE and ASIA and in articles on the major physical features of the country (*e.g.*, VOLGA RIVER; BLACK SEA; URAL MOUNTAINS). Descriptions of the traditional cultures of the country are found in such articles as ARCTIC PEOPLES AND CULTURES and SIBERIAN CULTURES. The following article is also supplemented by separate articles on each of the union republics and on major Soviet cities. (I.V.Ko./Ed.)



SOVIET UNION

Size of symbol indicates relative size of town • ◉ ■

Elevations in metres


UNITED STATES
Norton Sound
Nome
POINT HOPE
Bering Strait
DIOMEDE ISLANDS
NUNIVAK ISLAND
ALEUTIAN ISLANDS
CAPE DEZHNEV
Chukchi Sea
Anadyr
CHUKOTSK PENINSULA
Provideniya
WRANGEL ISLAND
Vankarem
Gulf of Anadyr
Egvekinot
Ueikal
Long's Strait
Beringovsky
Bering Sea
Ust-Chaun
ANADYR PLATEAU
Arctic Circle
Anadyr
Markovo
Velikaya
Khatyrka
CAPE NAVARIN
Aanbarchin
East Siberian Sea
Yerkuol
Shcherbakovo
Penzhina
Slautnoya
KORYAK RANGE
Kolyuchinskaya
Alazeya
Tabor
Indigirka
Dzhardzhan
Srednekhyrsk
Zyryanka
Mount Pekora 3147
Kolymskaya
Ostrovnoy
Gizhiga
Gulf of Shelikhov
Gulf of Korf
KAMCHATKA PENINSULA
Karaginskiy Island
Karaga
Tigil
SREDINNYY RANGE
Palana
Klyuchevskaya Volcano 15,863
Kronotskaya Volcano 12,739
Kronotskiy Gulf
Kronotskoye Lake
Petropavlovsk-Kamchatskiy
Malka
Morshechnoye
Kirovskiy
Cape Lopatka
KOMANDORSKIE ISLANDS
KOTELNY ISLAND
NEW SIBERIAN ISLANDS
SEVERNAYA ZEMLYA
CAPE CHELYUSKIN
Laptev Sea
GREAT BEGICHOV ISLAND
TAYMYR PENINSULA
Lake Taymyr
Khatanga
Dudinka
Saskylakh
Olenyok
Anabar
Yana
CHERSKY RANGE
Selennyakh
Verkhoyansk
Ust-Nera
Omolon
Kolyma
Ohotskchan
Myakit
Magadan
Yamsk
Palana
Kamchatka
Mount Klyuchevskaya 15,863
Kamchatka Valley
NORTH SIBERIAN LOWLAND
Agapa
Volochanka
Khatanga
Kheta
Kotuy
Sukhana
Muna
Zhigansk
Verkhoyansk
Oymyakon
Khandyga
Okhotsk
Sea of Okhotsk
Norilsk
Yenisey
PUTORAN MTS
Sholgontsy
Ekonda
CENTRAL SIBERIAN PLATEAU
CENTRAL YAKUT PLAIN
Vilyuy
Batamay
Aldan
Yakutsk
SETTE-DABAN RANGE
Turukhansk
Tunguska
Nizhnyaya
Mirny
Lensk
Markha
Amga
Tommot
Aldan
Tompolokan
DZHUGDZHUR RANGE
Aian
SHANTAR ISLANDS
CAPE TERPENIYA
Okha
Nikolayevsk-na-Amure
SAKHALIN ISLAND
Aleksandrovsk-Sakhalinskiy
Bakhta
Chunya
Vanavara
Podkamennaya Tunguska
Dushekan
Chona
Lena
Macha
ALDAN PLATEAU
Gonam
Olyokma
Chumikan
Udd
Guga
Amgun
Komsomolsk-na-Amure
Tatarsk
Pogibi
Lyuli
Gavan
Vyazhno-Sakhalinsk
Korsakov
Kirensk
Zhelezegorsk-Ilimskiy
STANOVOY UPLAND
Vitim
Skovorodino
Zeya
Shimanovsk
Svobodnyy
Belgorsk
Blagoveshchensk
Raychikhinsk
Birobidzhan
Khabarovsk
Sovetskaya Gavan
Strait
La Perouse Strait
HOKKAIDO
Sapporo
Maklakovo
Angara
Chuna
Bratsk
Bratsk Reservoir
Bodaybo
Chara
Uskut
Nagorny
Mukzha
STANOVOY RANGE
Bomnak
Ekimchan
Amgun
SIKHOTE ALIN
Krasnoyarsk
Kansk
Oka
Abakan
MINUSINSK BASIN
EASTERN SAYAN MOUNTAINS
WESTERN SAYAN MOUNTAINS
Cheremkhovo
Angarsk
Irkutsk
Angara
Lake Baikal
Ulan-Ude
Chita
ARGUNOVSKIY RANGE
Shilka
Khilok
Borzya
Argun
Mogocha
Skovorodino
Shimanovsk
Svobodnyy
Belgorsk
Blagoveshchensk
Raychikhinsk
Birobidzhan
Khabarovsk
Sovetskaya Gavan
Strait
La Perouse Strait
HOKKAIDO
Sapporo
Kyzyl
Yenisey
Petrovsk-Zabaykalskiy
Zabaykalsk
SUDETY-CHOVOCHNY RANGE
Sungari
Lesozavodsk
Teykkha
Lake Khanka
Spassk-Dalny
Ussuriysk
Artyom
Nakhodka
Vladivostok
HARBIN
GREATER KHINGAN RANGE
MANCHURIAN PLAIN
Ulaanbaatar
Sea of Japan
NORTH KOREA
MUKDEN
HONSHU
Nagoya
Kyoto
OSAKA
Hiroshima
KYUSHU
PEKING
TIENTSIN
Dairen
Lü-ta
P'yŏngyang
SEOUL
SOUTH KOREA
Pusan
Gulf of Chihli
Huang Ho
T'ai-yüan
Yellow Sea
PACIFIC OCEAN
EAST CHINA SEA
Grand Canal
Yün Ho
Nanking
SHANGHAI
Sian
Lan-chou
CH'I-LIEN SHAN-MO
TSAIDAM BASIN
SHAN-MO
MONGOLIA
ALTAI MOUNTAINS
GOBI
AN
CHINA
© Rand McNally & Co.
A-570000-257
0 200 400 600 800 km
0 200 400 600 800 mi

MAP INDEX

Political subdivisions

Armenian S.S.R. 40·00n 43·00e
Azerbaidzhan S.S.R. 40·30n 47·30e
Belorussian S.S.R. 53·30n 28·00e
Estonian S.S.R. 59·00n 26·00e
Georgian S.S.R. 42·00n 44·00e
Kazakh S.S.R. 48·00n 68·00e
Kirghiz S.S.R. 41·30n 75·00e
Latvian S.S.R. 57·00n 25·00e
Lithuanian S.S.R. 56·00n 24·00e
Moldavian S.S.R. 47·00n 29·00e
Russian Soviet Federated Socialist Republic 60·00n 100·00e
Tadzhik S.S.R. 39·00n 71·00e
Turkmen S.S.R. 40·00n 60·00e
Ukrainian S.S.R. 49·00n 32·00e
Uzbek S.S.R. 41·00n 64·00e

Cities and towns

Abakan 53·43n 91·26e
Agapa 71·27n 89·15e
Aktyubinsk 50·17n 57·10e
Aldan 58·37n 125·24e
Aleksandrovsk-Sakhalinsky 50·54n 142·10e
Alma-Ata 43·15n 76·57e
Ambarchik 69·39n 162·20e
Anadyr 64·45n 177·29e
Andizan 40·45n 72·22e
Angarsk 52·34n 103·54e
Anzhero-Sudzhensk 56·07n 86·00e
Aralsk 46·48n 61·40e
Arkhangelsk 64·34n 40·32e
Armavir 45·00n 41·08e
Artyom 43·22n 132·13e
Ashkhabad 37·57n 58·23e
Astrakhan 46·21n 48·03e
Ayagus 47·58n 80·27e
Ayan 54·43n 110·55e
Bagdarin 54·26n 113·36e
Bakhta 62·28n 89·00e
Baku 40·23n 49·51e
Balkhash 46·49n 74·59e
Barnaul 53·22n 83·45e
Batamay 63·31n 129·27e
Batumi 41·38n 41·38e
Belgorod 50·36n 36·34e
Belogorsk 50·57n 125·28e
Berezniki 59·24n 56·46e
Beringovsky 63·03n 179·19e
Birobidzhan 48·48n 132·57e
Biysk 52·34n 85·15e
Blagoveshchensk 50·16n 127·32e
Bobruysk 53·09n 29·14e
Bodaybo 57·51n 114·10e
Borzya 50·24n 116·31e
Bratsk 56·05n 101·48e
Bryansk 53·15n 34·22e
Bukhara 39·48n 64·25e
Buolkalakh 72·56n 119·50e
Chara 56·54n 118·12e
Chardzhou 39·06n 63·34e
Chelkar 47·50n 59·36e
Chelyabinsk 55·10n 61·24e
Cheremkhovo 53·09n 103·05e
Cherepovets 59·08n 37·54e
Cherkassy 49·26n 32·04e
Chernigov 51·30n 31·18e
Chiatura 42·19n 43·18e
Chimbay 42·57n 59·47e
Chimkent 42·18n 69·36e
Chita 52·03n 113·30e
Chu 43·36n 73·42e
Chumikan 54·42n 135·19e
Dalnegorsk 44.35n 135.35e
Dikson 73·30n 80·35e
Dnepropetrovsk 48·27n 34·59e
Donetsk 48·00n 37·48e
Druzhina 68·14n 145·18e
Dudinka 69·25n 86·15e
Dushanbe 38·35n 68·48e
Dushekan 60·39n 109·03e
Dzhambul 42·54n 71·22e
Dzhardzhan 68·43n 124·02e
Dzhetygara 52·11n 61·12e
Dzhezkazgan 47·47n 67·46e
Egvekinot 66·19n 179·10w
Ekibastuz 51·40n 75·22e
Ekimchan 53·04n 132·58e
Ekonda 65·47n 105·17e
Enurmino 66·57n 171·49w
Fergana 40·23n 71·46e
Frunze 42·54n 74·36e
Gizhiga 62·03n 160·30e
Gomel 52·25n 31·00e
Gonam 57·21n 131·12e
Gorky 56·20n 44·00e
Gorlovka 48·18n 38·03e
Grozny 43·20n 45·42e
Guga 52·43n 137·35e
Guryev 47·07n 51·56e
Igarka 67·28n 86·35e
Irkutsk 52·16n 104·20e
Ivanova 57·00n 40·59e
Izhevsk 56·51n 53·14e
Izhma 65·02n 53·55e
Kadykchan 63·02n 146·50e
Kalinin 56·52n 35·55e
Kaliningrad 54·43n 20·30e
Kaluga 54·31n 36·16e
Kamensk-Uralsky 56·28n 61·54e
Kansk 56·13n 95·41e
Karaganda 49·50n 73·10e
Karatau 43·10n 70·28e
Karazhal 48·02n 70·49e
Kaunas 54·54n 23·54e
Kavacha 60·16n 169·51e
Kazachye 70·44n 136·13e
Kazan 55·49n 49·08e
Kemerovo 55·20n 86·05e
Kerch 45·20n 36·26e
Khabarovsk 48·27n 135·06e
Khandyga 62·40n 135·36e
Khanty-Mansiysk 61·00n 69·06e
Kharkov 50·00n 36·15e
Khatanga 71·58n 102·30e
Khatyrka 62·03n 175·15e
Khilok 51·21n 110·28e
Kholmsk 47·03n 142·03e
Kiev 50·26n 30·31e
Kirensk 57·46n 108·08e
Kirov 58·33n 49·42e
Kirovabad 40·40n 46·22e
Kirovsk 67·37n 33·40e
Kirovsky 54·18n 155·47e
Kirovsky 44·52n 78·12e
Kishinyov 47·00n 28·50e
Kizyl-Arvat 38·48n 56·15e
Klaipeda 55·43n 21·07e
Kokand 40·33n 70·57e
Kolpashevo 58·20n 82·50e
Kolymskaya 68·44n 158·44e
Komsomolsk-na-Amure 50·35n 137·02e
Konosha 60·58n 40·15e
Korf 60·19n 165·50e
Korsakov 46·38n 142·46e
Kotlas 61·16n 46·35e
Krasnodar 45·02n 39·00e
Krasnovodsk 40·00n 53·00e
Krasnoyarsk 56·01n 92·50e
Kremenchug 49·04n 33·25e
Krivoy Rog 47·55n 33·21e
Kurgan 55·26n 65·18e
Kursk 51·42n 36·12e
Kushka 35·16n 62·20e
Kustanay 53·10n 63·35e
Kutaisi 42·15n 42·40e
Kuybyshev 53·12n 50·09e
Kyzyl 51·42n 94·27e
Kzyl-Orda 44·48n 65·28e
Leninakan 40·48n 43·50e
Leningrad 59·55n 30·15e
Lensk 60·43n 114·55e
Lesozavodsk 45·28n 133·27e
Liepaja 56·31n 21·01e
Lvov 49·50n 24·00e
Macha 59·54n 117·35e
Magadan 59·34n 150·48e
Magnitogorsk 53·27n 59·04e
Makayevka 48·02n 37·58e
Makhachkala 42·58n 47·30e
Maklakovo 58·16n 92·29e
Malka 53·20n 157·30e
Markha 61·49n 122·20e
Markovo 64·40n 170·25e
Mary 37·36n 61·50e
Minsk 53·54n 27·34e
Mirny 62·33n 113·53e
Mogilyov 53·54n 30·21e
Mogocha 53·44n 119·44e
Monchegorsk 67·56n 32·58e
Moroshechnoye 56·24n 156·12e
Moscow 55·45n 37·35e
Murmansk 68·58n 33·05e
Myakit 61·24n 152·09e
Nakhodka 42·48n 132·52e
Nadym 65·35n 72·42e
Nagorny 55·58n 124·57e
Nakhodka 42·48n 132·52e
Namangan 41·00n 71·40e
Naryan-Mar 67·39n 53·00e
Nebit-Dag 39·30n 54·22e
Nelkan 57·40n 136·13e
Nikolayev 46·58n 32·00e
Nikolayevsk-na-Amure 53·08n 140·44e
Nikopol 47·34n 34·24e
Nizhny Tagil 57·55n 59·57e
Nordvik 72·02n 111·32e
Norilsk 69·20n 88·06e
Novgorod 58·31n 31·17e
Novokazalinsk 45·50n 62·10e
Novokuznetsk 53·45n 87·06e
Novomoskovsk 54·05n 38·13e
Novorossiysk 44·45n 37·45e
Novosibirsk 55·02n 82·55e
Nukus 42·50n 59·29e
Odessa 46·28n 30·44e
Okha 53·34n 142·56e
Okhotsk 59·23n 143·18e
Omsk 55·00n 73·24e
Omsukchan 62·32n 115·48e
Ordzhonikidze 43·03n 44·40e
Orenburg 51·54n 55·06e
Orsk 51·12n 58·34e
Oryol 52·59n 36·05e
Osh 40·33n 72·48e
Oymyakon 63·28n 142·49e
Palana 59·07n 159·58e
Palatka 60·06n 150·54e
Pavlodar 52·18n 76·57e
Penza 53·13n 45·00e
Perm 58·00n 56·15e
Petropavlovsk 54·54n 69·06e
Petropavlovsk-Kamchatsky 53·01n 158·39e
Petrovsk-Zabaykalsky 51·17n 108·50e
Petrozavodsk 61·47n 34·20e
Poronaysk 49·14n 143·04e
Prokopyevsk 53·53n 86·45e
Provideniya 64·23n 173·18w
Pskov 57·50n 28·20e
Raychikhinsk 49·46n 129·25e
Riga 56·57n 24·06e
Rostov-na-Donu 47·14n 39·42e
Rubtsovsk 51·30n 81·15e
Ryazan 54·38n 39·44e
Rybinsk 58·03n 38·52e
Salekhard 66·33n 66·40e
Samarkand 39·40n 66·58e
Saransk 54·11n 45·11e
Saratov 51·34n 46·02e
Saskylakh 71·55n 114·01e
Semipalatinsk 50·28n 80·13e
Sevastopol 53·08n 65·44e
Serov 59·29n 60·31e
Severodvinsk 64·34n 39·50e
Shcherbakovo 65·15n 160·30e
Shevchenko 43·35n 51·05e
Shilka 51·51n 116·02e
Shimanovsk 52·00n 127·42e
Shologontsy 66·13n 114·14e
Skovorodino 53·59n 123·55e
Slautnoye 63·00n 167·59e
Slyudyanka 51·38n 103·42e
Smolensk 54·47n 32·03e
Sochi 43·35n 39·45e
Sovetskaya Gavan 48·58n 140·18e
Spassk-Dalny 44·37n 132·48e
Srednekolymsk 67·27n 153·41e
Stavropol 45·02n 41·59e
Sterlitamak 53·37n 55·58e
Sukhana 68·45n 118·00e
Surgut 61·14n 73·20e
Sverdlovsk 48·05n 39·40e
Svetlaya 46·33n 138·18e
Svobodny 51·24n 128·07e
Syktyvkar 61·40n 50·46e
Syrdarya 40·52n 68·38e
Tabor 71·16n 150·12e
Taldy-Kurgan 45·00n 78·23e
Tallinn 59·25n 24·45e
Tambov 52·43n 41·25e
Tartu 58·23n 26·43e
Tashauz 41·50n 59·58e
Tashkent 41·20n 69·18e
Tbilisi 41·42n 44·45e
Temirtau 50·05n 72·56e
Termez 37·14n 67·16e
Tiksi 71·36n 128·48e
Tobolsk 58·12n 68·16e
Tolyatti 53·31n 49·26e
Tommot 58·58n 126·19e
Tomptokan 57·06n 133·59e
Tomsk 56·30n 84·58e
Tselinograd 51·10n 71·30e
Tula 54·12n 37·37e
Turkestan 43·18n 68·15e
Turukhansk 65·49n 87·59e
Tuymazy 54·36n 53·42e
Tyumen 57·09n 65·32e
Uelkal 65·32n 179·17e
Ufa 54·44n 55·56e
Uglegorsk 49·02n 142·03e
Ukhta 63·33n 53·38e
Uka 57·50n 162·06e
Ulan-Ude 51·50n 107·37e
Ulyanovsk 54·20n 48·24e
Uralsk 51·14n 51·22e
Ussuriysk 43·48n 131·59e
Ust-Chaun 68·47n 170·30e
Ust-Kamenogorsk 49·58n 82 38e
Ust-Nera 64·34n 143·12e
Ust-Nyukzha 56·34n 121·37e
Ust-Tym 59·26n 80·08e
Vanavara 60·22n 102·16e
Vankarem 67·51n 175·50w
Velikaya 64·04n 176·12e
Verkhoyansk 67·35n 133·27e
Vilnius 54·41n 25·19e
Vinnitsa 49·14n 28·29e
Vitebsk 55·12n 30·11e
Vladimir 56·10n 40·25e
Vladivostok 43·08n 131,54e
Volochanka 71·00n 94·28e
Volgograd 48·44n 44·25e
Vologda 59·12n 39·55e
Volzhsky 48·49n 44·44e
Vorkuta 67·27n 63·58e
Voronezh 51·40n 39·10e
Voroshilovgrad 48·34n 39·20e
Vyborg 60·42n 28·45e
Warsaw 52·15n 21·00e
Yakutsk 62·13n 129·49e
Yamsk 59·35n 154·10e
Yaroslavl 57·37n 39·52e
Yerevan 40·11n 44·30e
Yeropol 65·15n 168·40e
Yessey 68·29n 102·10e
Yuzhno-Sakhalinsk 46·58n 142·42e
Zabaykalsk 49·38n 117·19e
Zaporozhye 47·50n 35·10e
Zhdanov 47·06n 37·33e
Zheleznogorsk-Ilimsky 56·37n 104·08e
Zhigansk 66·45n 123·20e
Zhitomir 50·16n 28·40e
Zlatoust 55·10n 59·40e
Zyryank 65·45n 150·51e

Physical features and points of interest

Alakol, Lake, *salt lake* 46·10n 81·45e
Alazeya, *river* 70·51n 153·34e
Aldan, *river* 63·28n 129·35e
Aldan Plateau, *upland* 57·00n 127·00e
Altai Mountains 48·00n 90·00e
Amga, *river* 62·38n 132·32e
Amu Darya, *river* 43·40n 59·01e
Amgun, *river* 52·56n 139·40e
Amur, *river* 52·56n 141·10e
Anabar, *river* 73·08n 113·36e
Anadyr Plateau 67·00n 174·00e
Anadyr, *river* 64·55n 176·05e
Anadyr, Gulf of 64·00n 179·00w
Angara, *river* 58·06n 93·00e
Aral Sea 45·00n 60·00e
Arctic Ocean 85·00n 90·00e
Argun, *river* 53·20n 121·28e
Ayon Island 69·50n 168·40e
Baikal, Lake 53·00n 107·40e
Balkhash, Lake 46·00n 74·00e
Baltic Sea 57·30n 20·00e
Barabinsk Steppe, *plain* 55·00n 79·00e
Barents Sea 74·00n 36·00e
Belukha, Mount, *mountain* 49·48n 86·40e
Bely Island 73·10n 70·45e
Bering Sea 58·00n 180·00
Bering Strait 65·30n 169·00w
Betpak-Dala, *plain* 46·00n 70·00e
Black Sea 43·00n 35·00e
Bolshoy Begichyov Island 74·20n 112·30e
Borshchovochny Range, *mountain range* 52·00n 117·00e
Bratsk Reservoir 56·10n 102·10e
Carpathian Mountains 48·00n 23·00e
Caspian Depression 48·00n 52·00e
Caspian Sea 42·00n 50·30e
Caucasus Mountains 42·00n 45·00e
Central Russian Upland, *mountain* 52·00n 38·00e
Central Siberian Plateau, *upland* 65·00n 105·00e
Central Yakut Plain 65·00n 124·00e
Chany, Lake, *salt lake* 54·50n 77·30e
Chelyuskin, Cape 77·45n 104·20e
Chersky Range, *mountain range* 52·00n 114·00e
Chono, *river* 62·54n 111·06e
Chu, *river* 45·00n 67·44e
Chukchi Peninsula 66·00n 175·00w
Chukchi Sea 69·00n 171·00w
Chuna, *river* 57·47n 95·26e
Chunya, *river* 61·36n 96·30e

MAP INDEX (continued)

Commander Islands 55·00n 167·00e
Communism Peak 38·57n 72·01e
Desna, *river* 50·33n 30·32e
Dezhnev, Cape 66·06n 169·45w
Diomede Islands 65·45n 169·30w
Dnepr, *river* 46·38n 32·18e
Dnestr, *river* 46·18n 30·17e
Don, *river* 47·04n 39·18e
Donets Basin 48·30n 38·30e
Dvina, *river* 57·04n 24·03e
Dzhugdzhur Range, *mountains* 58·00n 136·00e
Dzhungarsky Alatau Range, *mountain range* 45·00n 80·00e
East Siberian Sea 74·00n 166·00e
Elbrus, Mount, *mountain* 43·21n 42·26e
Emba, *river* 46·38n 53·14e
Finland, Gulf of 60·00n 27·00e
Fligeli, Cape 81·50n 59·20e
Franz Joseph Land, *islands* 81·00n 55·00e
Gorky Reservoir 57·00n 43·10e
Gyda Peninsula 70·50n 79·00e
Ili, *river* 45·24n 74·08e
Ilmen, Lake 58·17n 31·20e
Indigirka, *river* 70·48n 148·54e
Irtysh, *river* 61·04n 68·52e
Ishim, *river* 57·45n 71·12e
Ishim Steppe, *plain* 55·00n 71·30e
Issyk-Kul, Lake 42·25n 77·15e
Japan, Sea of 42·00n 135·00e
Kakhovka Reservoir 47·25n 34·10e
Kama, *river* 55·45n 52·00e
Kama Reservoir 58·52n 56·15e
Kamchatka, *river* 56·15n 162·30e
Kamchatka, Gulf of 55·35n 162·21e
Kamchatka Peninsula 56·00n 160·00e
Kanin Nos, Cape 68·39n 43·16e
Kara-Bogaz-Gol Bay 41·00n 53·15e
Karagin, Gulf of 58·50n 164·00e
Karagin Island 58·50n 164·00e
Kara-Kum Desert 39·00n 60·00e
Karakumsky Kanal, *canal* 37·35n 61·50e
Kara Sea 76·00n 80·00e
Karatau Range, *mountain range* 43·50n 68·30e
Karelia, *historic region* 64·00n 32·00e
Kazakh Uplands, *hills* 49·00n 72·00e
Kazym, *river* 63·54n 65·50e
Ket, *river* 58·55n 81·32e
Khanka, Lake 45·00n 132·24e
Khatanga, *river* 72·55n 106·00e
Kheta, *river* 71·54n 102·06e
Klyuchevskaya Volcano 56·04n 160·38e
Kola Peninsula 67·30n 37·00e
Kolguyev Island 69·05n 49·15e
Kolyma, *river* 69·30n 161·00e
Konda, *river* 60·40n 69·46e
Kopet-Dag, *mountains* 37·50n 58·00e
Koryak Range, *mountain range* 62·30n 172·00e
Kotelny Island 75·45n 138·44e
Kotuy, *river* 71·55n 102·05e
Kremenchug Reservoir 49·30n 32·30e
Kuban, *river* 45·20n 37·22e
Kulunda Steppe, *plain* 53·00n 79·00e
Kuma, *river* 44·55n 46·32e
Kura, *river* 39·24n 49·19e
Kuril Islands 46·10n 152·00e
Kuybyshev Reservoir 53·40n 49·00e
Kuznetsk Basin 54·00n 86·00e
Kyzylkum, *desert* 42·00n 64·00e
Kyzylray, Mount, *mountain* 48·27n 75·32e
Ladoga, Lake 60·08n 30·01e
La Perouse Strait 45·45n 142·00e
Ledyanaya, Gora, *mountain* 49·28n 142·45e
Lena, *river* 72·25n 126·40e
Long's Strait 70·20n 178·00e
Lopatka, Cape 50·52n 156·40e
Lovat, *river* 58·14n 31·28e
Mangyshlak Peninsula 44·18n 51·00e
Mezen, *river* 66·11n 43·59e
Minusinsk Basin 53·30n 92·00e
Muna, *river* 67·52n 123·06e
Nadym, *river* 66·12n 72·00e
Narodnaya, Mount, *mountain* 65·04n 60·09e
Navarin, Cape 62·16n 179·10e
Neman, *river* 55·18n 21·23e
Neva, *river* 59·57n 30·20e
New Siberia Island 75·00n 149·00e
New Siberian Islands 75·00n 142·00e
Nizhnyaya Tunguska, *river* 65·48n 88·04e
Northern Dvina, *river* 64·35n 40·30e
North Siberian Lowland, *plain* 73·00n 100·00e
Novaya Zemlya, *islands* 74·00n 57·00e
Novosibirsk Reservoir 54·35n 82·35e
Ob, *river* 66·45n 69·30e
Ob Bay 69·00n 73·00e
Oka, *river* 56·20n 43·59e
Oka, *river* 56·03n 101·50e
Okhotsk, Sea of 53·00n 150·00e
Olenyok, *river* 73·00n 119·55e
Olyokma, *river* 60·22n 120·42e
Olyutorsky, Cape 59·55n 170·27e
Om, *river* 54·59n 73·22e
Onega, Lake 61·30n 35·45e
Onon, *river* 51·42n 115·50e
Pamirs, *mountains* 38·00n 73·00e
Pay-Khoy, *mountains* 69·00n 63·00e
Pechora, *river* 68·13n 54·10e
Peipus, Lake 58·45n 27·30e
Penzhina, *river* 62·28n 165·18e
Pobeda, Mount, *mountain* 65·12n 146·12e
Pobedy, Pik, *peak* 42·02n 80·05e
Podkamennaya Tunguska, *river* 61·36n 90·18e
Pripet Marshes, *physical region* 52·00n 57·00e
Pripyat, *river* 51·21n 30·09e
Prut, *river* 45·28n 28·12e
Pur, *river* 67·31n 77·55e
Putoran Mountains 69·00n 95·00e
Pyasina, *river* 73·50n 87·10e
Riga, Gulf of 57·30n 23·35e
Rudolf Island 81·45n 58·30e
Rybinsk Reservoir 58·30n 38·30e
Sakhalin Island 51·00n 143·00e
Sartang, *river* 67·44n 133·12e
Sayan Mountains 52·45n 96·00e
Sea of Azov 46·00n 36·00e
Selenga, *river* 52·16n 106·16e
Selennyakh, *river* 67·48n 144·54e
Sette-Daban Range, *mountains* 62·00n 138·00e
Severnaya Zemla, *islands* 79·30n 98·00e
Severnye Hills, *mountains* 59·30n 49·00e
Seym, *river* 51·27n 32·34e
Shantar Islands 55·00n 137·36e
Shelikhov, Gulf of 60·00n 158·00e
Shilka, *river* 53·22n 121·32e
Siberia, *historic region* 60·00n 100·00e
Sikhote-Alin, *mountains* 48·00n 138·00e
Sredinny Range, *mountains* 56·00n 158·00e
Stanovoy Range, *mountains* 56·20n 126·00e
Stanovoy Upland, *mountains* 56·00n 114·00e
Sukhona, *river* 60·46n 46·24e
Sura, *river* 56·06n 46·00e
Syrdarya, *river* 46·03n 61·00e
Talish Mountains 38·42n 48·18e
Tarbagatay Range, *mountain range* 47·12n 83·00e
Tatarsk Strait 50·00n 141·15e
Tavda, *river* 57·47n 67·16e
Taymyr, Lake 74·30n 102·30e
Taymyr Peninsula 76·00n 104·00e
Taz, *river* 67·32n 78·40e
Tengiz, Lake 50·25n 69·00e
Terpeniya, Cape 48·39n 144·44e
Tien Shan, *mountains* 42·00n 78·00e
Timen Ridge, *mountains* 65·00n 51·00e
Tobol, *river* 58·10n 68·12e
Tsimlyansk Reservoir 48·00n 43·00e
Tura, *river* 57·12n 66·56e
Turan Plain 42·00n 61·00e
Turgay, *river* 48·00n 62·45e
Turgay Valley 52·00n 64·30e
Turgay Tableland, *plateau* 51·00n 64·00e
Tym, *river* 59·25n 80·04e
Uda, *river* 54·42n 135·14e
Ural, *river* 47·00n 51·48e
Ural Mountains 60·00n 60·00e
Usa, *river* 54·00n 28·55e
Ustyurt Plateau 43·00n 56·00e
Vakh, *river* 60·45n 76·45e
Vaygach Island 70·00n 59·30e
Velikaya, *river* 64·40n 176·20e
Verkhoyansk Range, *mountains* 67·00n 129·00e
Vilyuy, *river* 64·24n 126·26e
Volga, *river* 45·55n 47·52e
Volga Hills, *mountains* 52·00n 46·00e
Volgograd Reservoir 49·20n 45·00e
Vostochny Sayan, *mountain range* 53·00n 97·00e
Vytaka, *river* 55·36n 57·30e
Western Dvina, *river* 57·04n 24·03e
West Siberian Plain 60·00n 75·00e
White Sea 66·00n 40·00e
Wrangel Island 71·00n 179·30e
Yablonovy Range, *mountain range* 53·50n 115·00e
Yakutiya, *historic region* 65·00n 130·00e
Yamal Peninsula 70·00n 70·00e
Yana, *river* 71·31n 136·32e
Yelizavety, Cape 54·27n 142·42e
Yenisey, *river* 71·50n 82·40e
Yukagir Plateau 66·00n 155·00e
Yuzhny Bug, *river* 46·59n 31·58e
Zapadny Sayan, *mountain range* 53·00n 94·00e
Zaysan, Lake 48·00n 84·00e
Zeravshan, *river* 39·32n 63·45e
Zeya, *river* 50·13n 127·35e

I. The land

THE NATURAL LANDSCAPE

As a result of the country's enormous territorial extent, the complexity of its relief, and the wide range of climatic, drainage, and soil conditions, the natural landscapes of the Soviet Union exhibit exceptional variety.

Relief. Broad plains and depressions occupy the western half of the Soviet territory, while plateaus and mountain ranges prevail in the east; an almost continuous belt of high mountains runs along the southern border of the country. Altitudes range from the 24,590-foot (7,495-metre) Communism Peak in the Pamirs to the bottom of the Karagiye, a depression that lies 433 feet (132 metres) below sea level in the Mangyshlak Peninsula.

Range of altitudes

Plains and plateaus. The plains—subdivided into the enormous Russian (or East European), West Siberian, and Turan plains—stretch from the Baltic Sea in the west to the Yenisey River in the east and from the shores of the Arctic Ocean to the foothills of the southern Kopet-Dag Range.

The average altitude of the Russian Plain is about 560 feet, but elevations rising to 1,000–1,300 feet above sea level alternate with depressions. In the northern half of the plain, formerly covered by glaciers, the relief between the river valleys is strewn with lakes and swamps. In the south, however, watersheds are higher and cut into by valleys and ravines. The ancient wind-blown deposits known as loess are widespread and locally deep in the Ukraine and in the Don River Basin. In the northwest, near Scandinavia, the ancient uplands of Karelia and the mountains of the Kola Peninsula (maximum height 3,907 feet [1,191 metres]) abut on the Russian Plain, while in the south it is bounded by the moderately high (maximum 6,762 feet [2,061 metres]) mountain ranges of the Carpathians and the Crimea and the great barrier of the Caucasus. The last named, stretching from the Black Sea to the Caspian, rises to 18,510 feet (5,642 metres) at Mt. Elbrus. In the east, the moderate relief of the Ural Mountains extends for more than 1,200 miles (2,000 kilometres) from north to south, forming part of the traditional dividing line between Europe and Asia.

The Russian Plain

The upland ridges of the Urals also separate the Russian Plain from the West Siberian Plain, an area of monotonous, low relief and one of the largest lowlands in the world. In the south this is bounded by the Turgay Plateau and the Kazakh Uplands, which rise to a maximum of 5,134 feet (1,565 metres). South again, the Turan Plain occupies the greater part of Central Asia. Its relief is varied; troughs lying below sea level alternate with extensive low plateaus.

Between the Yenisey and Lena rivers lies the Central Siberian Plateau, mainly at heights of 1,500–3,000 feet (450–900 metres). This is bounded on the north by the North Siberian, or Khatanga, Lowland, beyond which the uplands of the Taymyr Peninsula rise to 3,760 feet (1,146 metres), while to the east, in the middle basin of the Lena, lies the Central Yakut Plain.

The mountainous belts. Regions with markedly dissected mountainous relief lie in the south and east of Soviet Asia. The highest ranges occur in the Uzbek, Tadzhik, Kirgiz, and Kazakh republics and in southwest Siberia, where the Pamir, Alai, Tien Shan, and Altai have crest lines well above 10,000 feet (3,000 metres). Numerous somewhat lower ranges, together with extensive plateaus and uplands, occupy a large area on either side of Lake Baikal. Farther east, the Stanovoy Range forms the watershed between the Aldan River, which flows toward the Arctic, and the Zeya, draining to the Pacific via the Amur. Southeast of the Stanovoy lie the mountain ranges of the South Primorye and the region around the Amur River. North from the Amur region, on the shores of the Sea of Okhotsk, stretches the Dzhugdzhur Range, and beyond it the mountains of northeast Siberia and the Chukchi Peninsula begin.

The Kamchatka Peninsula is also mountainous; it harbours a group of volcanoes, including the Klyuchevskaya Volcano, at 15,584 feet (4,750 metres) the highest summit in all Siberia and the Far East. The Kuril group of some 30 rather large and markedly rugged islands is part of the East Asian volcanic arc. In these islands are found more than 100 volcanoes, almost 30 of which are active. West of the Sea of Okhotsk, Sakhalin Island forms a major feature running north–south for more than 560 miles but with an average width of only about 60 miles.

Latitudinal climatic belts

Climate. *Regional distribution.* Latitudinal climatic belts are superimposed on the patterns of Soviet relief. The Far North, with most of its offshore islands, lies in the Arctic climatic belt; the northern European Soviet Union, Western Siberia above the Arctic Circle, and Central and Eastern Siberia between 70° and 60°–64° north latitude are in the sub-Arctic climatic belt; the remainder of Soviet Europe and Siberia, as well as Central Asia as far south as 40° north latitude, are in the temperate belt; and parts of Transcaucasia and southern Central Asia lie in the subtropics.

As most Soviet territory lies far from the oceans, the climate almost everywhere is markedly continental, with a clear division of the year into cold and warm periods. Continentality increases eastward and is especially pronounced in Central Asia and Eastern Siberia, where the average temperature difference between the warmest month (July) and coldest month (January) ranges from 45° C (81° F) to 65° C (117° F). The most westerly regions on the Black Sea coast and the south Crimean coast are much milder than the rest of the country. The southern Far East is subject to Pacific Ocean influences, having a monsoon climate.

Atmospheric circulation, temperature, and precipitation. Atmospheric circulation above Soviet territory depends upon the seasons. In winter, when there is a negative heat radiation balance (that is, the Earth's surface is losing heat), a region of high pressure known as the Asian anticyclone is formed in Inner Asia. West of it, a ridge of high pressure (the "Voyeykov axis") stretches over the northern edge of Central Asia and the southern Russian Plain. The northern half of Soviet Europe is then under the influence of recurrent Atlantic cyclones that bring west and southwest winds, relative warmth, and rain and snow. In the Far East, meanwhile—a region reached by air formed in the Asian anticyclone—the winter monsoon is under way, and cold, dry west or northwest winds prevail.

A cold continental air mass, formed as a result of the cooling of the land, nevertheless dominates the greater part of Soviet territory in winter; it is cold almost everywhere, and precipitation is relatively small. Average January temperatures along a line from the southwest to the northeast fall from 43° F (6° C) at Batumi, on the Black Sea coast, to a remarkable —58° F (—50° C) in the vicinity of Verkhoyansk, in the Far East region. The Murmansk coast, Kaluga, Volgograd, and Kyzyl-Orda, situated in widely different latitudes, have the same January mean of 14° F (—10° C). There are particularly severe January frosts (down to —85° F [—65° C] or below) in the Yakut A.S.S.R., while the absolute minimum of —90° F (—68° C) recorded at Verkhoyansk is the lowest temperature ever recorded on the Earth's surface outside Antarctica. Relief as well as anticyclonic conditions contributes to this "pole of cold" in the Verkhoyansk district; the heavier cold air lingers in hollows, which are much colder than the surrounding mountain slopes. The highest winter temperatures occur in southern Central Asia, but even here the January mean is only a few degrees above freezing.

Minimum temperatures at Verkhoyansk

In the cold half of the year up to eight to 12 inches (200 to 300 millimetres) of precipitation occurs in the Soviet west, but this decreases toward the east, falling to only two to four inches (50 to 100 millimetres) in Central Asia and Eastern Siberia. Winter precipitation falls mostly as snow, which lies for a few days in the Crimea, Transcaucasia, and Central Asia but covers the ground for 260 days or more in the Taymyr Peninsula and on the Arctic islands. The thickness of the snow cover is also uneven. In the steppes of the southern Ukraine and northern Caucasus and in the Central Asian deserts it is rarely as much as four inches. Snow cover is thin on the Baltic coast and in Belorussia, where there are frequent thaws, and also in the Volga region and Kazakhstan, where there is little winter precipitation. It generally thickens to the north and east to reach 40 inches or more on the east coast of Kamchatka.

In summer, because of the intense heating of the land surface (the radiation balance is now positive everywhere), a broad region of low pressure is formed over Asia and covers almost all Soviet territory except the southern Russian Plain, where pressure remains high. Westerly winds prevail in Soviet Europe, northerlies in Northern Siberia and in Central Asia; meanwhile, the east and southeast summer monsoon dominates the Far East. Summer air temperatures are above freezing everywhere (except in the mountains) and generally decrease toward the north; only on the northeast Pacific coast, under oceanic influence, is a fall in temperature observed from west to east. Average July temperatures in southern Central Asia are, in places, above 86° F (30° C). The absolute Soviet maximum temperature, recorded in the town of Termez in the Uzbek S.S.R., is 122° F (50° C). The central Soviet Union has considerably lower July temperatures—from 59° to 70° F (15° to 21° C)—while the Arctic coasts experience even lower readings. Conversely, the frost-free period increases from 45–60 days in the Siberian tundra to 270 days in Transcaucasia.

Maximum temperatures at Termez

Most precipitation falls in the warm half of the year; in the central Soviet Union, totals drop from 20 inches in the west to less than half that in Eastern Siberia. The heaviest rain (up to 47 inches, or about 1,200 millimetres) falls in the western Caucasus, the least (less than 4 percent of the 1,200 millimetres) in the southern Turan Plain.

Drainage. *River distribution patterns.* There are more than 150,000 rivers in the Soviet Union that are more than six miles long; their total length is about 2,000,000 miles (3,200,000 kilometres), one-sixth of which is suitable for navigation or floating logs. Soviet waterpower resources are the world's greatest. The density of the river system, however, ranges from 0.40–0.56 mile per square mile (0.25–0.35 kilometre per square kilometre) in the

European forest zone to a fraction of this amount in the deserts and semideserts of the Caspian Depression.

Arctic rivers

The majority of rivers drain to the Arctic Ocean. The largest of these northward-draining rivers are the Northern Dvina, Pechora, Ob and Irtysh, Yenisey, Khatanga, Lena, Indigirka, and Kolyma. The Pacific Ocean takes in all the Far Eastern rivers, including the Amur, Anadyr, Penzhina, and Kamchatka. The Atlantic Ocean ultimately receives the waters of the rivers that discharge into the Black Sea (the Danube, Dnestr, Dnepr, Kuban, and Rioni), the Sea of Azov (the Don), and the Baltic Sea (the Neva and the Western Dvina).

The eastern Caucasus, part of the Russian Plain, the Tien Shan, and the Pamirs, as well as the semideserts and deserts of Central Asia and Kazakhstan, comprise the extensive internal Aral–Caspian drainage region, more than 60 percent of which is drained by rivers flowing into the Caspian Sea. These are the Volga, Ural, Emba, Kura, and Kuma rivers. The Syrdarya and Amu Darya drain to the Aral Sea Basin, while many rivers in this region, such as the Ili, Chu, and Zeravshan, either end in lakes or vanish into the desert sands.

The river regimes. Climatic variety is reflected in river-flow characteristics, or regimes. The differing climatic zones offer varying sources of nourishment to the associated rivers and affect annual distribution of flow. Thus a distinction is made between rivers fed mainly by melted snow (the Volga, Don, Dnepr, Pechora, Ob, Yenisey, and Lena), by rain (the Amur), by glacier and snow melt (the high mountain rivers of the Caucasus, the Central Asian mountains, and the Altai), and by subterranean groundwaters (in the foothill regions of Central Asia).

Soviet rivers may be divided into three groups, depending on flow characteristics: those with spring high water (the most numerous); those with spring and summer high water (the rivers of the Far East, the Yana–Indigirka Basin, Transbaikalia, and the high mountains of the Caucasus and Central Asia); and those with a flood regime (the mountain rivers of the Crimea, the northern slope of the Great Caucasus, the Carpathians, and the Black Sea coast of the Caucasus). The rivers with the largest discharge are those in the Arctic Basin (three-fifths of the annual total Soviet river discharge), especially the Yenisey, which has an annual discharge of 150 cubic miles (624 cubic kilometres).

Most Soviet rivers are at times covered with ice; the period varies from eight to nine months in the Siberian north to one to two months in the extreme south.

Lakes and reservoirs. An abundance of lakes—they number about 270,000—includes five of the world's largest, each with an area exceeding 4,000 square miles (10,000 square kilometres): the Caspian and Aral seas, Lake Baikal (at 5,315 feet [1,620 metres] also the world's deepest), Lake Balkhash, and Lake Ladoga. An additional 14 lakes—including Onega and Issyk-Kul—have areas of more than 400 square miles (1,000 square kilometres). The lakes are very unevenly distributed, being particularly numerous in the regions of surplus moisture in the Soviet north and the Western Siberian steppes. In Karelia and the Lena–Vilyuy plain, they occupy more than 10 percent of the surface area, whereas in the forest–steppe of the Russian Plain they are almost nonexistent.

Largest lakes

As a rule, there is a flow of river water through the large northern lakes, while in some regions (notably Karelia and the Kola Peninsula) there are many rivers that start from lakes, run through lakes, and sometimes even end in lakes. Nonflowing lakes are scattered through the Soviet south; each lies in a hollow from which there is no water discharge, gathers atmospheric and groundwater from the whole region, and then expends it mainly by means of evaporation.

Soviet lakes also vary greatly in their content of mineral salts, and there are freshwater, brackish, and salt lakes. In Soviet Asia, especially in the Kulunda Steppe of Western Siberia, there are lakes whose water has a high sulfate content, from which chemicals are produced.

Lake temperatures are subject to sharp variations. In summer, the water in shallow lakes is intensely heated, while in the deep lakes the temperature falls with depth to a low of 39° F (4° C); in winter the water temperature rises with depth as the lower levels retain summer heat.

As a result of the construction of hydroelectric plants and navigation canals, there are many very large artificial lakes in the valleys of the major rivers. These include the Kakhovka, Kremenchug, and Kiev reservoirs on the Dnepr, Tsimlyansk on the Don, Rybinsk, Kuybyshev, and Volgograd on the Volga, Novosibirsk on the Ob, and Bratsk on the Angara. The Kayakkum, Chardzhou, and other lakes in Central Asia have been created by irrigation projects.

Groundwaters and marshes. The groundwaters of the Soviet Union are not homogeneous. In much of Soviet territory—from the tundra to the northern steppes—they have a relatively high content of calcium bicarbonate and dissolved silica and are also marked by abundant underground discharge, thus helping to feed rivers. In the dry steppes, semideserts, and deserts, groundwaters with high chloride and sulfate content are formed, which in places are unsuitable for drinking. In the loamy deserts of Central Asia the groundwaters lie at great depth and are usually bitterly salty. In the sand deserts the water lies nearer the surface and is not so mineralized. In the foothill zone of Central Asia, groundwaters are abundant and slightly mineralized.

Marshes occupy nearly 10 percent of Soviet territory, and they are an important element of the landscape in several regions. They are found mainly in the northern half of the country, in the forest and tundra zones, where precipitation exceeds evaporation. In central and northern Karelia they amount to 40–50 percent of the land surface, and they occupy nearly 80 percent in parts of the Western Siberian taiga. Farther south the marshland area falls off, and marshes are concentrated in such depressions as that called (in the West) the Pripet Marshes (in Belorussia and the Ukraine) and the Meshchera (near Moscow and Ryazan). In the forest-steppe zone, marshland diminishes sharply, and south of this zone it becomes quite insignificant; it covers about 1.5 percent of the Ukrainian steppes.

Soils. The Soviet plains are characterized by several distinct latitudinal (zonal) soil belts.

In the Arctic islands and on the Arctic coast, relatively infertile (and often permanently frozen) Arctic and tundra soils are dominant. The latter are thin, with a low organic content (1–2 percent) and coarse composition, and are boggy.

About a third of Soviet territory is covered by the soils of the forest zone. These include podzols and (in Eastern Siberia) frozen taiga soils (both developed under taiga-type forests), together with turf–podzol soils and huge areas of peat bog and gley–podzol soils, "gley" being a sticky clay formed in waterlogged soils. All varieties of podzols and podzolic soils are characterized by intense leaching, as a result of which easily soluble substances are washed out of the upper strata. The top stratum is gray or brown with 1–4 percent of humus, or organic matter. The underlying stratum is ashy ("podzol" comes from Russian *zola,* "ashes") and consists of the finest quartz sand grains; nutrient substances are completely washed out. Below this is a third layer, which in fully developed podzols has an accumulation of humus and leached mineral nutrients; in podzolic soils the humus is lacking, and in gley–podzols there is only an accumulation of clay minerals. In the extreme southwest (in Transcarpathia) and southeast (the central Amur and South Primorye) of the podzol soil zone, there appear brown forest soils, formed under broad-leaved and partially mixed forests. These are characterized by an accumulation of clay minerals and ferrous oxides, which impart the brown colour; a "clotty" structure; and a fairly well-developed humus stratum (3–7 percent).

Podzols

South of the zone of podzol and marshland soils stretches a narrow strip of gray forest soils, developed under leafy forests of the northern forest steppe. Gray forest soils have a distinctly formed stratum of humus, and there is no uninterrupted podzolic stratum.

Still farther south stretch the black-earth, or chernozem, lands, occupying 9 percent of Soviet territory and comprising almost half the world area of this soil type. Black-earth soil is formed under the grassy vegetation of the steppe and forest–steppe, in conditions of adequate warmth and a negative or near neutral balance of moisture. This prevents intense leaching and assists the accumulation of organic substances. In addition, meadow–black-earth soil is also widespread in the steppes. The greater part of the black-earth belt is now under the plow.

In the dry steppes and semideserts of the southeast Russian Plain and of Kazakhstan are found chestnut soils, marked by a rather low humus content, a tendency toward saltiness, and a feeble structure. The meadow–chestnut soils are formed in hollows and estuaries in conditions of heightened surface or ground moisture.

Continuing the southward succession, chestnut soils give way to other semidesert types and then, in the deserts themselves, to brownish-gray and gray-earth soils (sierozems). These are all very poor in humus and are often salty (solonetz and solonchak) and are broken by extensive salt marshes. A unique type of desert soil is the *takyr,* a clay soil found in depressions, which creates impassable mud during rain and a brick-hard crust in dry periods. These soils are so barren that they are often devoid of all vegetation except meagre lichens.

Red-earth soils

Red-earth soils in their typical form are found only on the lower mountain slopes of the Kolkhida plain (in the Caucasus). These, the most ancient of Soviet soils, are soils of humid subtropical forests, very thick, and rich in iron and aluminum oxides. Abkhaziya and the Talish Mountains are the sites of other humid subtropical forest soils, the yellow-earth soils, distinguished from the red-earth by their paler yellow tint and their inconsiderable depth. In addition to the typical gray-earth soils, brown soils are widespread in the dry subtropics, notably on the slopes of the mountains of Central Asia and the Caucasus, being found under low-growing, broad-leaved forests, sparse forests, and thickets.

Finally, in the mountain regions, comprising 31 percent of Soviet territory, the soils form a series of vertical belts. A particularly full selection of high-altitude soil belts is represented in the Caucasus, where the soils vary from black-earth to mountain pasture, subalpine, and alpine. The mountain soils are stony, well drained, and not very thick. The alluvial soils of the river floodplains also should be mentioned, because they are the youngest of Soviet soils, still being formed.

Vegetation. About 17,000 plant species and almost half of all the families of higher plants known in the world are found in the Soviet Union, which lies entirely in the area known to botanists as the Boreal kingdom. The vast Soviet plains are characterized (as in the case of soils) by clearly marked latitudinal vegetation zones. Successively, these are Arctic waste, tundra, forest–tundra, forest, forest–steppe, steppe, semidesert, and desert.

The latitudinal belts. The Arctic wastes include most of the Arctic islands and the north coast of the Taymyr Peninsula; mosses, lichens, and a variety of grasses cover scarcely half of the soil surface.

The tundra zone—divided into subzones of Arctic, moss–lichen, and scrub tundra—stretches in a continuous belt along the northern border of Eurasia, including some Arctic islands and the Commander Islands of the Pacific.

To the south lies the zone of forest–tundra, in which the barren tundra landscapes alternate with marshes and sparse forests of birch and fir (in the west) and larch (in the east).

The forest zone

Farther south stretches the broad belt of the forest zone, extending from the western borders of the country to the Pacific coast. Coniferous taiga forests of fir, pine, silver fir, cedar, and Siberian larch are most widespread to the west of the Yenisey River, while in Central and Eastern Siberia the forests are predominantly of larch. Marshes are very common, especially in the taiga of Western Siberia, and various types of meadows are also found.

Large areas of birch and aspen groves, growing on the site of former coniferous forests destroyed by felling or fire, are common in the southern taiga of the Russian Plain and in the West Siberian Plain; they do not last long, because shoots of conifers gradually appear under their canopy and choke the young trees. Mixed (coniferous and broad-leaved) and broad-leaved forests occur in the southern half of the Russian Plain forest zone. Unlike the conifers, the broad-leaved forests have many tiers, the uppermost usually consisting of crowns of oak, lime, ash, and maple. Below these come bird cherry and mountain ash, then a thick, shrubby undergrowth, perhaps of hazel; the whole is underlain by grass. Beech and hornbeam appear in broad-leaved forests in the southwest of the zone. Mixed forests are also widespread in the southern Far East, where most of the conifers are pine or whole-leaf silver fir, and common broad-leaved trees include oak, Amur linden, maple, and hazelnut. The depressions and lower mountain slopes are covered with broad-leaved forests of oak with a striking admixture of black birch.

The forest–steppe zone extends in a narrow latitudinal belt across the Russian Plain and the West Siberian Plain. Meadow steppe with various grasses alternates with small areas of oak (in Soviet Europe) or birch and aspen (in Western Siberia). Farther east, forest–steppe is encountered only as isolated islands in taiga forests.

To the south, forest–steppe gives way to mixed-grass and grassy steppe, and the forests remain only along river valleys and ravines. The herbaceous covering of the grassy steppes consists of needlegrass (*Stipa* sp.) and such other perennial turf grasses as fescue.

Farther south lies a semidesert zone extending in a continuous belt from the Yergeni hills (in the southeast of the Russian Plain) to the mountain systems of the Altai and Tien Shan. Grasses that like dryness (such as *Stipa rubens* and fescue) and semishrubs (such as saltwort and wormwood) predominate in the thin, patchy, vegetation covering.

The desert zone

Finally, the southernmost desert zone occupies the southern third of the Caspian Depression and the Central Asian plain. It is distinguished by a very sparse and poor covering of vegetation, composed of saltwort, wormwood, and grasses. Saltwort (*Salsola richteri*) and sand acacia are found in relatively light sandy soils and in river valleys. An abundance of ephemeral plants is typical of the southern part of the zone, where sand deserts, notably the Kara-Kum and Kyzylkum, predominate; in addition *Haloxylon* "forests" are common, alternating with salt-marsh wastes overgrown with saltwort.

The subtropical, Mediterranean, and mountain regions. The humid subtropics of the Kolkhida and Lenkoran depressions have a unique vegetation as well as soils; they are clothed in dense broad-leaved forests that include oak, beech, hornbeam, chestnut, and alder, with an undergrowth of evergreen plants and lianas.

Vegetation of the Mediterranean type makes a strong showing on the Black Sea coast. Forests of oak, juniper, and pine are common.

In mountainous regions, vegetation is distributed in altitudinal belts, particularly in the Caucasus, the Carpathians, and the mountains of the Primorye. Thus the steppes of the foothills of the northern Caucasus give way to mountain forest–steppe, broad-leaved forests of oak and beech, and then pine (in the west) and fir (in the east). Above the forest belt lie sparse high mountain forests, thickets of rhododendron, and subalpine and alpine meadows. In arid regions the middle mountain belt of vegetation is in some places composed of turf-grass steppes, related thickets, and sparse juniper forests. High mountain steppe and meadow are widespread in Central Asia, with turf-grass steppe giving way to areas of dark coniferous forests of spruce or to mixtures of steppe, meadow, shrub thickets, and sparse juniper forests. Coniferous taiga forests cover the Siberian mountains, with tundra above; the higher Altai, however, exhibits subalpine and alpine meadows.

Animal life. The Soviet Union lies entirely in what zoologists call the Palaearctic subregion of the Holarctic

faunal region, in which animal life is characterized by five distinct sectors of varying size and natural conditions.

The Arctic sector embraces both Arctic and tundra landscape zones. The Arctic zone (the islands and the mainland seacoasts) is poor in animal species; such mammals and birds connected with the sea as the polar bear, sea hare, seal, walrus, eider, gull, and loon are found. In summer the coastal rocks are covered with nesting colonies of seabirds. The tundra is inhabited by animals adapted to cold climate and open spaces that are under snow for long periods. They include the Arctic fox, reindeer, alpine hare, lemming, and—among its birds—ptarmigan and snowy owl. Geese, swans, and ducks migrate into the region in summer. Lizards, frogs, and other reptiles and amphibians are found only in the extreme southwest of the zone and in the forest–tundra.

The European–Siberian forest sector (covering the Russian Plain, Siberia, and much of the Far East) possesses a much richer and more varied animal life. The northern taiga harbours elk, brown bear, reindeer, lynx, sable, and alpine hare, with such birds as the crossbill, nutcracker, cuckoo, owl, and woodpecker. There are few reptiles or amphibians. Representative fauna of the broad-leaved forests are wild pig, deer, and mink, and the most typical birds are the hawfinch, oriole, owl, and nightingale. Various snakes, lizards, and tortoises are the common reptiles.

The steppe environment

In the steppes, the abundance of plant food allows such rodents as the marmot, jerboa, and hamster to predominate. Birds include the kestrel, crane, eagle, and lark; hooved animals include the occasional steppe antelope; the masked (or steppe) polecat and Tartar fox represent beasts of prey. There are few amphibians but more reptiles than in the north.

Animal life in the Mediterranean sector, found in the Caucasus and mountainous Crimea, is very varied as a result of the great landscape variety and also of the fact that the fauna was originally connected with that of the Mediterranean, Central Asia, and south and southeast Europe. The Caucasus harbours the wild goat, bison, chamois, bezoar goat, Caucasian deer, wild pig, porcupine, leopard, hyena, jackal, Persian squirrel, brown bear, and roe deer. Birds include species characteristic of both mountain (Caucasian black grouse, turkey, and stone partridge) and forest (nuthatch, jay, crossbill, woodpecker, and bullfinch). Reptiles and amphibians are numerous, ranging from cobra to tortoise.

The semideserts and deserts of Central Asia and Kazakhstan form the Central Asian sector, whose animals are specially adapted for existence in conditions of high summer and low winter temperatures, lack of water, and poor vegetation. Many of them are nocturnal. Typical desert dwellers are the Persian gazelle, sand cat, sand eel, and ground squirrel; birds include the bustard, pheasant, raven, and sparrow. Spiders, including tarantulas, are also present. Several northern forest species are widespread in the eastern part of the Tien Shan; the Kopet-Dag harbours fauna of the fore-Asian type (leopard, stone marten, partridge, and others); and the remaining part of the mountains is inhabited by animals of Central Asian and even (in insignificant measure) Indian origin: the argali (a wild sheep), snow leopard, wild goat, red wolf, bear, turkey, and others.

(V.M.St./J.C.D.)

THE TRADITIONAL REGIONS

Traditional names for large sections of the Soviet Union, such as the North, Western Siberia, and the Far East, are well established. Some of these names have been applied to the present-day official economic regions into which the country is divided for planning purposes, but the traditional regional divisions do not constitute natural or economic entities; it would be more correct to define them as natural historic regions, differing in both their geography and their economy and also in the density and composition of their populations. Their borders are quite arbitrary, following sometimes natural and sometimes administrative boundaries.

The traditional regions of the European Soviet Union are the North (or European North), Centre, Volga Lands, West, Southwest, and Urals. The Caucasus is a distinctive region abutting the European Soviet Union on the south. The Asian part includes Central Asia and Kazakhstan, which occupy the vast area south of Western Siberia and east of the Caspian Sea; the huge territory of Siberia, traditionally subdivided into Western Siberia and Eastern Siberia; and the Far East.

The North. The European North (also long known as the Russian North and equivalent to the present-day Northwestern economic region) corresponds closely to the ancient lands of Novgorod and covers a much greater area than the other European Soviet regions. Bordered by Finland on the west and the Urals on the east, it comprises the northern slopes of the Russian (or East European) Plain, running toward the Arctic Ocean and to the Gulf of Finland of the Baltic Sea. It also includes such major Arctic islands as Novaya Zemlya and the Arctic archipelago Franz Josef Land. Administratively, the North lies within the Russian Soviet Federated Socialist Republic and is composed of the Arkhangelsk, Leningrad, Murmansk, Novgorod, Pskov, and Vologda *oblasti* (regions) and the Karelian and Komi Autonomous Soviet Socialist republics. It covers some 640,000 square miles (1,660,000 square kilometres)—about one-third of Soviet Europe—and is the home of nearly 13,000,000 people, mainly Russians, with Komis (or Zyryans), Karelians, and Nenets (or Yuraks) leading the minority nationalities.

Regions of the North

The physical geography of such a huge area is naturally varied. The main overall character, however, is molded by forest and sea; the whole landscape is marked by a northern severity, and transitions are generally gradual.

A characteristic feature is the extremely uneven distribution of population and economic development. The east and north consist of taiga and swampy tundra and are thinly populated; settlement is concentrated principally in valleys of the main rivers, the Northern Dvina and Pechora. The most important resources here are timber, petroleum, and coal in the northeast and various minerals in the Kola Peninsula. The largest cities are Arkhangelsk, the country's oldest timber port and a centre of industry, and Murmansk, a major ice-free port and fishing centre.

In the southwest, conditions are less severe, and settlement was already quite dense in early times, for here such old Russian cities as Novgorod arose, as the economy developed. Here, too, is Leningrad, the second Soviet city in population and economic importance and one of the largest cultural and scientific centres.

The European North today is one of the Soviet Union's most important industrial regions and the main base of the timber industry.

The Centre. Central Russia (all of it within the Russian S.F.S.R.), with an area of 350,000 square miles (900,000 square kilometres) and a population of 44,500,000, occupies the whole central belt of the Russian Plain as far south as the Ukraine and from Belorussia eastward to the Volga Lands. Low hills alternate with broad lowlands, and the greater portion lies in a forest zone (taiga in the north, mixed and then broad-leaved forests farther south). The far south is in fact in the forest–steppe zone, noted for its black-earth soil; fields of wheat, sugar beets, and sunflowers occupy the open expanses.

The Russian nucleus

Ever since the formation of the Russian state in the 15th century, central Russia, as the nucleus of that state and the hearth of Russian national culture, has played a primary role in national growth. More than 17 percent of the Soviet population is concentrated here, in little more than 4 percent of the national area. Its proportionate share of industrial production, of scientific manpower, and of the training for the professions generally is even greater.

There are three distinct economic subregions: the Central Industrial region, the Black-Earth Belt, and the Volga–Vyatka region.

The Central Industrial region. This subregion (called the Central region in the official economic classifica-

tion and composed of Bryansk, Ivanov, Kalinin, Kaluga, Kostroma, Moscow, Oryol, Ryazan, Smolensk, Tula, Vladimir, and Yaroslavl *oblasti* and the city of Moscow) stands out sharply for its level of economic development; 28,400,000 persons (more than 11 percent of the Soviet total) make their homes here, in but 2 percent of the national territory. The region produces more than half the country's textiles and a large percentage of the total output of, among others, the products of engineering and chemistry. Among the urban centres (which contain 76 percent of the population), Moscow, the Soviet capital and leading political, industrial, scientific, and cultural centre, is of exceptional importance. Here the natural scenery has charm and variety deriving from its undulating expanses, the bright, neat birch groves, and the broad valleys of deep rivers on whose banks communities of various sizes lie amid meadows and copses.

The Black-Earth Belt. To the south, abutting the Ukraine, lies the Black-Earth Belt (officially the Central-Black Earth region and including the Belgorod, Kursk, Lipetsk, Tambov, and Voronezh *oblasti*), the home of 7,800,000 people and notable both for its extremely fertile black-earth (chernozem) soil and for the unique iron-ore deposits of the Kursk Magnetic Anomaly. The landscape has many features of the transition from northern forest to southern steppe; it is lent character by fields, oak groves, and villages strung out along river valleys and ravines or occupying flat watersheds. With only 0.75 percent of the country's area, this region nevertheless provides 6 percent of its agricultural produce and is of growing industrial significance. The largest town is Voronezh, and nearly half the population is urban.

The Volga–Vyatka region. This region (its name is the same in the official classification) occupies the eastern part of central Russia, covering the Gorky and Kirov *oblasti* and the Chuvash, Mari, and Mordvinian A.S.S.R.'s. It is the home of about 8,300,000 people. The regional landscape is given character by its great rivers, the Volga and its tributary the Vyatka. Steep river banks, wide bands of water, and water meadows lend variety to the plains. Here the Volga flows from west to east along the boundary of the forest and forest–steppe zones. On its precipitous right bank rise the "mountains," as the local people say, while on its left bank, beyond the water meadows, stretch forested lowlands. This area was often celebrated in the works of Aleksandr Herzen and Maksim Gorky, who were born there. The Volga–Vyatka region is known as the land of forests and factories. Timber is its principal natural wealth, but industry (particularly engineering) accounts for the greater part of economic output. More than 60 percent of the population is urban, and the largest industrial centre is Gorky.

The Upper Volga landscape

The Volga Lands. Near Kazan the Volga sharply changes direction to flow southward. Here the Tatar, Kalmyk, and Bashkir A.S.S.R.'s, together with six *oblasti* of the Russian S.F.S.R., form the Volga Lands, called the Povolzhye region officially. With an area of some 260,000 square miles (680,000 square kilometres) it is the home of about 19,100,000 people, approximately 7.5 percent of the Soviet total.

Stretching a thousand miles from north to south, the region extends across a series of natural landscape zones from the southern fringe of the forest zone to the semi-desert bordering the northern Caspian. Its principal natural resources are fertile land, pastures, the country's most important oil fields, and, not least, two rivers, the Volga and its tributary the Kama. As the largest waterway and the most important source of hydropower in Soviet Europe, the Volga is of enormous economic importance, and, with the building of hydroelectric stations at Kuybyshev, Saratov, and Volgograd, the Volga Lands have become the principal link in the Soviet unified power grid. After World War II the area became the country's main oil base and the main centre of the chemical industry. Agriculture, too, is important; the Volga Lands provide more grain than any other region of the Russian S.F.S.R., and Saratov wheat is among the best in the country. At the same time the region is one of the Soviet Union's main suppliers of meat, wool, vegetables, and melons. The curse of agriculture here is drought and dry winds, but the area of irrigated lands increases every year.

The Lower Volga landscape

More than two-thirds of the population is urbanized; the largest towns and industrial centres are Kazan, Kuybyshev, Saratov, and Volgograd (formerly Stalingrad).

The West. Composed of Belorussia, Moldavia, the three Baltic republics (Estonia, Latvia, and Lithuania), and the Kaliningrad *oblast* of the Russian S.F.S.R., the West occupies an area of more than 166,000 square miles (430,000 square kilometres) and contains more than 21,000,000 people (about 8 percent of the national total). Of these, about 7,500,000 are Belorussians, 2,300,000 Moldavians, 2,500,000 Lithuanians, 1,300,000 Latvians, and 900,000 Estonians. There are also about 3,200,000 Russians and nearly 700,000 Poles. The five union republics have vigorous national traditions.

The natural features here are lent character by the residual influence of ancient glaciers; the climate is relatively mild and humid, and many forests, especially pine forests, flourish on the sandy soil. Southwest Belorussia contains the world's oldest nature reserve, the Belovezhskaya (Belavezhskaya) Pushcha. The morainic ridges of residual glacial debris lend variety to the landscape, which changes imperceptibly every few miles from plain to hilly country, woods alternating with fields, meadows with marshes. Rivers connect hundreds of scattered lakes. In the Baltic republics a special element is introduced by the ruins of medieval castles that are often found on high ground, while the Baltic coastal resorts are very popular.

Effects of ancient glaciers on the landscape

The region as a whole embraces three major economic regions: the Baltic region (Estonia, Latvia, Lithuania, and the Kaliningrad *oblast*), Belorussia, and Moldavia. Fishing is highly developed (the Baltic republics provide more than 20 percent of the total national catch), as is engineering. Agriculture, specializing in raising livestock, is noted for its butter and bacon. The main type of fuel produced in Belorussia is peat, together with small quantities of oil, natural gas, and lignite. Belorussia also produces about 20 percent of Soviet flax and almost a third of its potassium fertilizers and is the main source of heavy trucks and tractors. Only about 54 percent of the population of the West is urban, figures for individual republics ranging from 37 percent in Moldavia to 66 percent in Latvia and 68 percent in Estonia. The biggest urban centres are the republic capitals—Minsk, Kishinyov, Riga, Vilnius, and Tallinn—and Kaunas and Kaliningrad (formerly Königsberg).

The Southwest. Covering an area of 233,000 square miles (604,000 square kilometres), the Southwest is coterminous with the Ukrainian S.S.R. (though in some classifications the region includes also Moldavia and the Rostov *oblast* of the R.S.F.S.R.). In the official Soviet pattern of economic regions, the Ukraine is divided into the Donets–Dnepr, Southwestern, and Southern regions. The Black Sea and the Sea of Azov are on the south, and it is through their ports that the Soviet Union maintains its trade with Asia, America, and southern Europe. The southern coast of the Crimean Peninsula and the area around Odessa are noted for their holiday resorts.

Forest–steppe and steppe occupy the greater part of the land area; both (but especially the latter) possess the fertile black-earth soil, which has attracted dense settlement from ancient times. The soils are almost entirely plowed to such crops as wheat, maize (corn), sugar beets, and sunflowers. There are also many orchards and vineyards there, especially in the Crimea, the southern coast of which forms the northern fringe of the subtropical zone.

Topographically, the Southwest is predominantly a plain (though the western Ukraine contains a portion of the Carpathians, and the Crimean Mountains rise in the south). The plain is divided into right bank and left bank regions by the Dnepr, the republic's largest river, as it flows south to the Black Sea.

Mineral resources of the Southwest

The Ukraine is rich in coal (the largest coalfield being the Donets Basin, or Donbass, which extends into the

adjacent Rostov *oblast* of the R.S.F.S.R.), iron ore (the Krivoy Rog and Kerch deposits), manganese (the largest deposit in the world is at Nikopol), natural gas, and other mineral resources. Hydropower is also significant, chiefly on the Dnepr.

Covering but 3 percent of Soviet territory, the Southwest contains more than 49,000,000 persons, or 19 percent of the national total. Of the region's population, 68 percent lives in urban areas, the proportion rising eastward from 35–40 percent in the western Ukraine to 67 percent in the eastern Ukraine. The largest Ukrainian cities are Kiev (the capital of the republic), Kharkov, Odessa, Dnepropetrovsk, and Donetsk. There are more than a dozen other cities with populations over 250,000, most of them in the Donets–Dnepr industrial zone.

The Urals. The Urals region contains the mountains that give it its name. These run southward from the Arctic shores to form the traditional boundary between Europe and Asia. The name is also given to the economic region. The region includes the Udmurt A.S.S.R. and five *oblasti* of the R.S.F.S.R., a total area of 260,000 square miles (680,000 square kilometres).

Divisions of the Urals

The Ural Mountains themselves consist of many medium ridges and rises, covered with forest. Customarily divided into the Polar, Nether-Polar, Northern, Central, and Southern Urals, they rise to a high point of 6,214 feet (1,894 metres). Natural features of the mountain region and the adjacent plains are unusually varied. The tundra and "alpine" peaks of the Polar and Nether-Polar Urals give way to the rolling taiga plains of the Northern Urals, while the fertile plowed steppes of the Central and Southern Urals are succeeded by the lake-dotted forest plains of the far (Siberian) side. The region is rich in minerals, notably metallic ores and precious and semiprecious stones. The Urals region has long played an outstanding role in the Russian economy because of its mineral wealth. In its industrial development it is surpassed only by central Russia and the Ukraine; it produces just under one-third of all Soviet iron, steel, and rolled steel. It is also a major petroleum-producing region and a base for the industrialization of the Soviet eastern regions.

The Central and Southern Urals are urbanized (65 percent of the total population) and contain such major industrial cities as Sverdlovsk and Chelyabinsk, metallurgical and chemical centres such as Magnitogorsk and Perm, and many settlements by mines, factories, and power stations. At the same time, old timber villages are to be found in the heart of the mountains. More than 15,000,000 persons, four-fifths of them Russians, live in the Urals, and there are many minority nationality groups, ranging from Tatars to the oldest inhabitants, the Komi-Permyaks.

The Caucasus. The Caucasus is the isthmus between the Black Sea and the Sea of Azov in the west and the Caspian Sea in the east. The northern boundary is the Kuma-Manych Depression, while the southern is the frontier with Turkey and Iran. The region is subdivided into the North Caucasus (predominantly plain), the Great Caucasus range, and Transcaucasia, containing the Kolkhida and Kura-Araks lowlands, the mountains of the Little Caucasus, and the Armenian Highland. The Caucasus is often alternatively divided (as it is in the formal Soviet classification of economic regions) into only two parts, the North Caucasus and Transcaucasia; the whole Caucasian sector of the Russian S.F.S.R. is included in the former, while Transcaucasia takes in the Georgian, Azerbaijan, and Armenian S.S.R.'s.

Extremes of the Caucasus landscape

The natural conditions of this comparatively small region (about 209,000 square miles [541,000 square kilometres]) are exceptionally varied, for it contains fertile steppe plains and high mountains covered with glaciers and snow (the highest point is Mt. Elbrus, at 18,510 feet [5,642 metres]); the hot and humid lowlands of Kolkhida and the dry steppes of the Kura-Araks; lava plateaus, alpine pastures, and forests on the mountain slopes of the humid west.

The Great Caucasus is an important natural boundary. North of it stretches the temperate zone of the Russian Plain, and to the south lie subtropical zones. A wide range of crops is cultivated.

The Caucasus has various minerals, notably petroleum, gas, copper, and manganese, which have contributed to the development of industry in several parts of the region. At the same time, the Caucasus is the main Soviet resort area, boasting a mild climate, warm seas, muds believed to be curative, and, most important of all, numerous mineral springs. The latter support such spas as Kavkazskiye and Mineralnye Vody in the northern foothills of the Great Caucasus, near which are the major resort towns of Kislovodsk, Yessentuki, Pyatigorsk, and Zheleznovodsk. On the Black Sea coast are numerous health and holiday resorts, of which Sochi is the largest in the Soviet Union. The Caucasus—especially around Mt. Elbrus—is also a major centre for hiking and mountaineering.

About 29,000,000 persons live in the Caucasus, and the region is unrivalled in the Soviet Union for the number of its nationalities; the Dagestan A.S.S.R. alone has at least 20 ethnic groups.

More than half the total population lives in cities, the largest of which are Baku, Tbilisi, and Yerevan, the capitals, respectively, of the Azerbaijan, Georgian, and Armenian S.S.R.'s. At the time of the 1970 census, there were 21 cities with populations of more than 100,000 and 19 with populations between 50,000 and 100,000.

The Caucasus is the home of the most ancient cultures in the Soviet Union. Numerous architectural monuments of past centuries are found, and the ruins of temples and fortresses witness to the distant past. As a consequence of this great age and the region's ethnic diversity, several distinctive local cultures may be distinguished.

Central Asia and Kazakhstan. In the heart of Eurasia, stretching from the Caspian eastward to the Altai and from the Urals and Western Siberia southward to the frontiers with Iran, Afghanistan, and China, the region known as Central Asia and Kazakhstan covers 1,550,000 square miles (4,000,000 square kilometres). This is the Soviet Union's hottest and driest region, full of striking contrasts: immense plains, with desert and semidesert depressions below sea level; very high mountains permanently covered with snow and ice; and thickly populated oases surrounded by empty deserts.

Desert and semidesert

All the Soviet deserts and almost all the semideserts (occupying in all about one-seventh of the national area) are found here. To the south and east, the desert expanses are bordered by mountain ranges, culminating in the Pamirs and the Tien Shan. To the north of the desert belt, low hills and isolated mountain ridges—the Kazakh Uplands—extend for hundreds of miles in central Kazakhstan. Many of the hills are covered with pine forests, while semideserts predominate in the depressions. North again stretches the West Siberian Plain; the north Kazakhstan portion is a zone of forest–steppe and steppe, with black-earth and dark chestnut soils.

Central Asia. In the political, administrative, and economic sense (and in terms of natural history), the term Central Asia embraces the four Central Asian republics—Uzbekistan (including the Kara-Kalpak A.S.S.R.), Turkmenistan, Tadzhikistan, and Kirgiziya. These republics cover almost 6 percent of the total Soviet area, and their 23,500,000 people represent 9 percent of the total population. Apart from the four nationalities for which the union republics are named, Central Asia is now the home of Russians, Ukrainians, Tatars, Kazakhs, Uighurs, and other people who have settled there at various times.

Settlement in the Central Asian republics is almost entirely concentrated where irrigation potential is high, in a narrow belt between the desert plains and the mountains and also in the valleys of the major rivers. The rivers of Central Asia not only irrigate the fields (the area contains more than half of all Soviet irrigated land) but also, with 15 percent of national hydroelectric resources, serve as an important source of power. Oasis agriculture specializes in fruit, grapes, silk, vegetables, rice, and cotton and other fibre crops. Central Asia provides nine-tenths of the cotton grown in the Soviet Union, 70 percent of its dried fruit, and 60 percent of its raw silk. The deserts and

mountains are areas of natural pasture for sheep and lesser numbers of cattle, horses, camels, and—in the Pamirs—yaks. Central Asia is of importance as a supplier of petroleum, natural gas, mercury, antimony, iodine, bromine, and gold and rare metals. Engineering is mainly directed to the production of cotton-picking and textile machinery and oil and chemical equipment.

Cities of the Central Asian region

The largest cities and industrial centres are the capitals of the union republics—Tashkent, Frunze, Dushanbe, and Ashkhabad—and also the oldest city of Central Asia, Samarkand, which was an important place in the 4th century BC. These ancient lands are replete with traces of the great civilizations of the past and cherish distinctive contemporary ways of life.

Kazakhstan. Lying to the north of the Central Asian republics, the Kazakh S.S.R., with an area of 1,050,000 square miles (2,700,000 square kilometres), is the second largest union republic and ranks third in population (14,300,000). Situated between the Urals, Siberia, and Central Asia, Kazakhstan finds much of its geographical and economic character molded by those of its neighbours; cold Siberia and hot Central Asia, forest and desert, broad plains and great mountain ranges meet, as it were, in Kazakhstan. It also contains important deposits of iron, copper, lead, zinc, and rarer minerals, as well as coal, oil, phosphorites, and rare metals. Industry dominates the republic's economy, but its agricultural importance increased greatly during the 1950s, when nearly 100,000,000 acres (40,000,000 hectares) were brought into cultivation in north Kazakhstan and southwest Siberia under the Virgin and Idle Lands Campaign. As a result, Kazakhstan now produces more cereals than the Ukraine, ranking second only to the R.S.F.S.R. in this respect.

The indigenous people, the Kazakhs, constitute only about a third of the republican total; Kazakhstan is also the home of representatives of almost all the major peoples of the Soviet Union, including Russians (42 percent of the republic's population), Ukrainians, Germans (deported from the Soviet west in 1941), Tatars, Uzbeks, Belorussians, Uighurs, Dungans, Koreans, and others. About 54 percent of the population lives in towns and cities, of which the largest are the capital, Alma-Ata, and the industrial centre of Karaganda.

Siberia. Siberia (all of which is part of the Russian S.F.S.R.), the giant among Soviet regions, stretches 4,300 miles (7,000 kilometres) from the Urals in the west to the mountain ranges of the Pacific watershed in the east; from north to south it extends 2,200 miles (3,500 kilometres) from the shores of the Arctic Ocean to the dry steppes of Kazakhstan and the Mongolian frontier. It occupies an area of 2,500,000 square miles (6,550,000 square kilometres), or 30 percent of all Soviet territory. Siberia lies in the Arctic, sub-Arctic, and temperate climatic belts, and its climate is sharply continental. The outlying northern parts of the region are covered by tundra and forest–tundra, with the taiga zone occupying the greater part. In the south there are zones of forest–steppe and steppe, and it is here that settlement is thickest.

Economic resources of Siberia

Siberia contains up to 90 percent of Soviet coal resources, petroleum, about half the natural gas, 70 percent of the peat, and deposits of iron and polymetallic ores and minerals. Siberia has more than 50 percent of all Soviet hydropower resources, more than two-thirds of its timber, and 230,000 square miles (600,000 square kilometres) of arable and pasture lands. Its forests contain a wealth of furbearing animals.

Just over 20,000,000 people live in this vast area, which accounts for about 8 percent of the Soviet Union's total industrial output, 15 percent of its electricity, and 25 percent of its timber. With each five-year plan the population increases, new towns are built, and production expands. The overwhelming majority of the people are Russians, who began to settle among the Ostyaks (Khants), Voguls (Mansis), Evenks, Buryats, Yakuts, Khakass, and other indigenous peoples in the late 16th century. There are several autonomous *oblasti* (regions) and autonomous (formerly national) *okruga* (districts) in which one of these peoples constitutes the chief element in the population.

Western Siberia. Western and Eastern Siberia—separated by the Yenisey River—may be differentiated (as they are in the official Soviet economic regional scheme) in respect of their physical geography and their degree of economic development. Three-fifths of Siberia's people live in the predominantly flat lands of Western Siberia (the West Siberian Plain), which cover 940,000 square miles (2,400,000 square kilometres). The greater part of Siberia's industrial and agricultural output comes from this region, the most important cities of which are Novosibirsk, Omsk, Novokuznetsk, and Barnaul.

Eastern Siberia. Eastern Siberia is predominantly plateau and mountain. It constitutes 18 percent of the Soviet total area, and only 8,000,000 people live there. It provides the country with diamonds, gold, aluminum, nickel, and other minerals, and its forests supply timber and furs. The world's largest hydropower stations are found on the Eastern Siberian rivers. The most important towns are Krasnoyarsk and Irkutsk.

The Far East. The Soviet Far East (which in some classifications is considered part of Siberia) stretches for 2,700 miles (4,300 kilometres) along the Pacific coast. It is composed of seven administrative entities: the Yakut A.S.S.R., the Primorye and Khabarovsk *kraya* (territories), and the Amur, Kamchatka, Magadan, and Sakhalin *oblasti*. Its area is 2,400,000 square miles (6,200,000 square kilometres), and its population is scarcely more than 6,500,000—a density of three persons per square mile (one person per square kilometre).

Low population density

The highest point—the Klyuchevskaya Volcano (15,584 feet [4,750 metres]—is less than 200 miles (300 kilometres) from the north end of the Kuril Trench, which at its lowest point is 34,586 feet (10,542 metres) deep. The Far East is mountainous: ridges, highlands, and plateaus are found in more than three-quarters of the region, and plains are found only between mountains and along the rivers and seacoasts. There are more than 250 volcanoes (almost 70 active) in the Kamchatka Peninsula and the Kuril Islands. The Far East is a source of gold, lead, tin, zinc, and other minerals. The customary power resources are here supplemented by the force of high ocean tides and heat generated by volcanoes and hot springs.

Most of the maritime region has a monsoon climate, with severe winters and cool summers. Because of abundant summer rainfall, the rivers, notably the Amur, are generally full. Landscapes are varied: in the north, tundra extends over large expanses; farther south lies taiga, giving way to broad-leaved forests. The Far East is second only to Siberia in its timber resources. In the south is a variety (and rare mingling) of southern and northern flora and fauna: lianas wind round the trunks of fir trees, and the northern deer meet the Siberian tigers.

Russians—who began to inhabit the region in the 17th century—constitute the majority of the population in the south. Next in numbers come Ukrainians, Belorussians, and Jews, and there is a great variety of very small indigenous population groups. Slightly more than three-quarters of the population lives in cities, the most important of which are Vladivostok, Khabarovsk, and Komsomolsk-na-Amure, the last named largely constructed by members of the All-Union Leninist Communist League of Youth, or Komsomol.

The Far East provides almost one-third of the total Soviet catch of fish and other seafoods and produces one-third of the canned fish. The region also takes a leading place in the extraction and smelting of ores, in the production of paper, and in the supply of timber. It is becoming of increasing importance as a base for Soviet exports.

SETTLEMENT PATTERNS

The basic trends. The population of the Soviet Union on January 1, 1977, numbered 257,824,000, giving an average density of about 28 persons per square mile (11 persons per square kilometre). Distribution, however, is very uneven.

Table 1: Soviet Union, Area and Population

	area		population (000)		growth (percent)
	000 sq mi	000 sq km	1959	1977	1959–77
Soviet Federated Socialist Republic					
Russian S.F.S.R.	6,591.1	17,075.4	117,534	135,600	15.4
Soviet Socialist Republics					
Armenian S.S.R.	11.5	29.8	1,763	2,893	64.1
Azerbaijan S.S.R.	33.4	86.6	3,698	5,776	56.2
Belorussian S.S.R.	80.1	207.6	8,056	9,414	16.9
Estonian S.S.R.	17.4	45.1	1,197	1,447	20.9
Georgian S.S.R.	26.9	69.7	4,044	4,999	23.6
Kazakh S.S.R.	1,048.9	2,717.3	9,295	14,498	56.0
Kirgiz S.S.R.	76.6	198.5	2,066	3,443	66.7
Latvian S.S.R.	24.6	63.7	2,093	2,512	20.1
Lithuanian S.S.R.	25.2	65.2	2,711	3,342	23.3
Moldavian S.S.R.	13.0	33.7	2,885	3,885	34.7
Tadzhik S.S.R.	55.2	143.1	1,981	3,591	81.3
Turkmen S.S.R.	188.4	488.1	1,516	2,650	74.8
Ukrainian S.S.R.	233.0	603.7	41,869	49,300	17.7
Uzbek S.S.R.	172.7	447.4	8,119	14,474	78.3
Total Soviet Union	8,598.1*	22,274.9*	208,827	257,824	23.5

*The total area of the Soviet Union is 8,647,250 sq mi (22,402,200 sq km), including the White Sea (34,750 sq mi [90,000 sq km]) and the Sea of Azov (14,400 sq mi [37,300 sq km]), which are excluded from the figures for the republics.
Source: Official government figures.

The mass of the Soviet population is concentrated in the so-called Main Belt of Settlement, which coincides with the main zone of grain production. Its base stretches along the western border from Leningrad to Moldavia; the northern edge of this belt passes through Cherepovets, Vologda, Kirov, and the north of the metallurgical and mining district of the Urals; the southern edge passes along the shores of the Black Sea and the Sea of Azov and then proceeds through Rostov-na-Donu, Volgograd, and Kuybyshev, to Chelyabinsk. Beyond the Urals the Main Belt of Settlement narrows, and to the east of Novosibirsk it becomes a narrow strip along the Trans-Siberian Railroad, spreading out only in a few places (in the Kuznetsk Basin, Minusinsk Basin, and elsewhere). This ribbon of denser population then stretches to the Pacific. To the south of this belt are important areas of denser population in Central Asia and the Caucasus.

The overall rural landscape. Of the total Soviet population, 62 percent now lives in urban settlements. Before World War II more than two-thirds was still rural, but even then industrialization had brought about both a relative and an absolute reduction in the rural population. Rural dwellers were first outnumbered by urban in 1961, but, since urban settlements occupy only about 1 percent of the country's total area, it is rural forms of settlement that chiefly mold the character of the landscape visible to the transcontinental traveller.

It has been calculated that there are about 470,000 rural inhabited localities, with a population of about 100,000,000, in the Soviet Union. Most of these settlements are connected with agricultural production. The remainder include small settlements at railway stations, at individual small mines, at centres of timber felling, at small or medium-sized processing enterprises, and also at resort settlements, inhabited localities at sanatoriums and hospitals sited in rural areas, and so on. Agricultural settlements are either old, formerly separate hamlets and villages whose inhabitants now work on collective farms, or the communities of state farms.

Collective and state farm structure

As a rule, each collective and state farm (respectively, *kolkhoz* and *sovkhoz*) has a central principal farmstead with its own settlement. At collective farms this is usually the largest (or the most conveniently sited) village in the area; at state farms, the largest farm settlement is the centre that houses machinery, fuel, and produce. These central settlements contain the collective farm board or the state farm directorate and such services as a club, one or more shops, a secondary school, clinics, and so on. Peripheral settlements contain primary schools, simple shopping facilities, and medical posts (without permanent staff) and the individual collective farm brigade (or state farm section) leadership.

Regional variations in rural population. *Soviet Europe.* In the Far North of Soviet Europe the rural population is very sparse, a reflection of the unfavourable environment. Tundra is used only as pasture for reindeer and as hunting and trapping grounds; fishing settlements are found along the rivers and the Arctic coast. The habitations (*stanovishcha*) are small in size. Several clusters of mining and port settlements have nevertheless sprung up, and railways (with their station settlements) have consequently reached these remote parts.

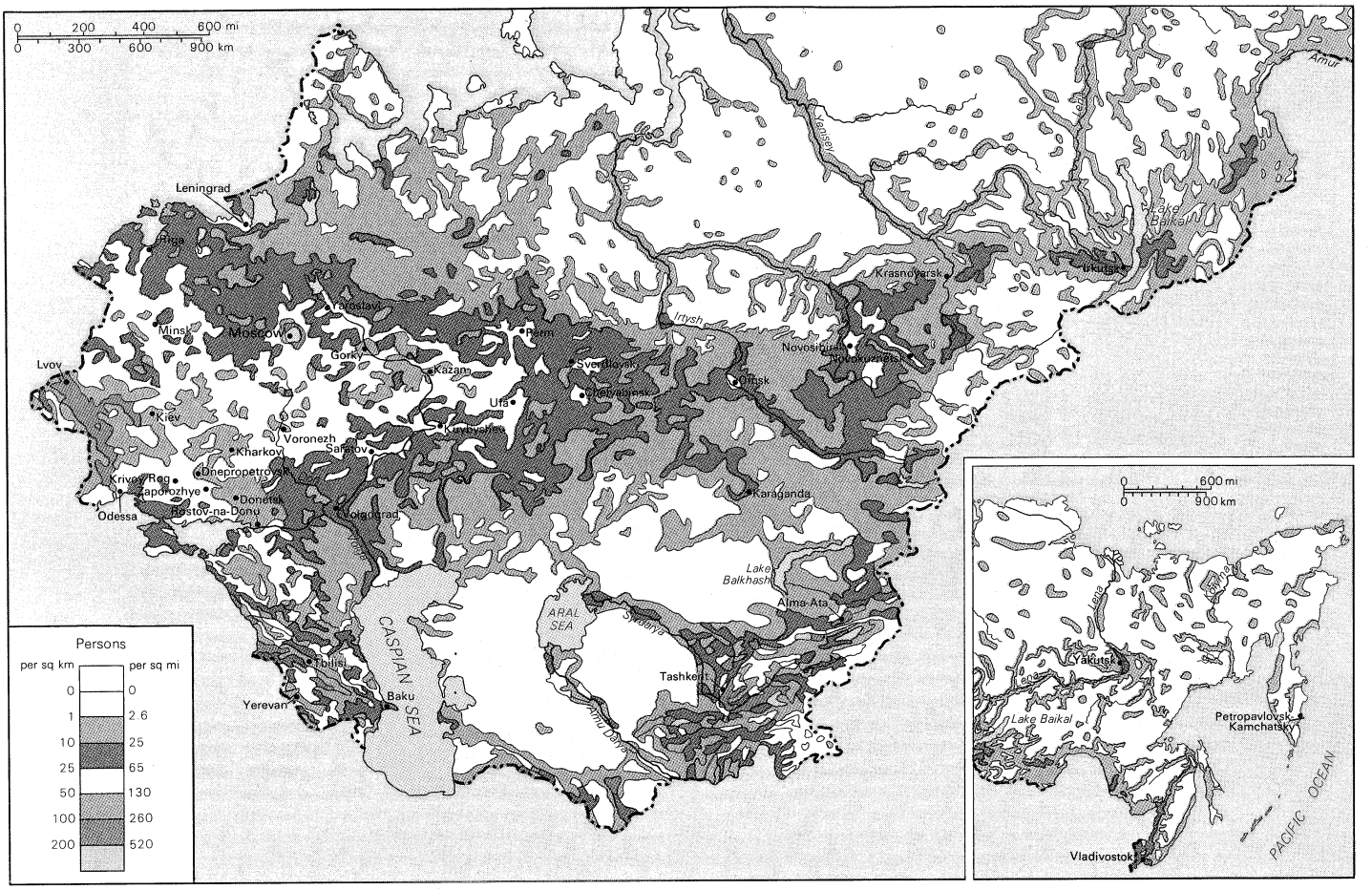
Farther south, in the northern taiga, the population is somewhat denser; agriculture acquires significance, and small villages of perhaps 100 inhabitants, centres of cleared patches of land, are strung along the fertile river valleys. Timber felling and occasionally timber processing are still the economic mainstays and have caused the creation of a great number of small settlements (populations 100–500) surrounded by felled areas; some of these lie along rivers (floating timber by raft to processing centres is common), others along railway lines and their timber-hauling (sometimes narrow-gauge) branch lines, or at ports.

In the middle belt of the European Russian S.F.S.R. (home of up to a quarter of the total Soviet rural population) settlement is still denser, and the landscape becomes one of continuous cultivation, with arable land comprising 30–60 percent of the total area; southward are fewer and fewer forests, and at about the latitude of Kuybyshev they survive as mere islets or strips of pine forest. Overlying the comparatively dense village network are strings of settlements at railway stations and also whole clusters of settlements at industrial enterprises situated outside the towns. Fairly evenly distributed medium and occasionally large villages nevertheless continue as the basic structure. From northwest to southeast there is a marked regularity in the increasing size of rural settlements, reflecting the growth in general density of the agricultural population and the associated fall in the proportion of uncultivable land. Historical circumstances—the widespread earlier distribution of large farms—were also important in the south. The predominant rural building material in this region (and also to the north of it) continues to be wood. From this, one-family houses and the necessary auxiliary structures are constructed. Apartment buildings of the modern urban type, however, are gradually appearing. The systematic enlargement of rural settlements—especially in the northwest, where villages average only 10–20 households—is an important precondition for spreading urban amenities and for raising the level of services, a process that, it is thought, requires an average settlement population of 1,500 to 3,000.

Taken together, the Ukraine, southern Belorussia, and Moldavia contain nearly 30 percent of the total Soviet rural population. The forms of population distribution—a close village network—are similar to those that have arisen in the middle belt of the R.S.F.S.R., but, reflecting greater densities, villages are generally larger and closer together. They are especially dense in the western Ukraine and in Moldavia. Nonagricultural rural settlements are mainly connected with processing agricultural raw materials, occasionally with mining, and also (on the seacoast, especially in the Crimea) with resort functions. In Belorussia and the northern Ukraine wood is still the major building material, but bricks, clay, and stone are used elsewhere; multifamily structures are spreading very rapidly.

Stanitsy and *auly*

Especially large rural settlements—Cossack villages, or *stanitsy*, numbering several thousand and sometimes 10,000–15,000 inhabitants—are typical of the plains of the North Caucasus; in the past these were Cossack communities guarding the southern frontiers. Villages are also large in the neighbouring Lower Volga region. The *auly*, settlements of the indigenous peoples in the foothills and mountains of the North Caucasus, are also quite large, their size depending on the economic capacity of the valleys in which they are situated. While the popu-



Population density of the Soviet Union.

lation distribution in the plains is fairly even, in the foothills and mountains it is in ribbons following the rivers flowing from the mountains. Population distribution has also been affected by the plowing up of steppe and by the building of irrigation systems linked to, for example, the Nevinnomyssky Kanal and the Volgo–Donskoy Kanal reservoirs.

In Transcaucasia, too, the rural population is distributed in generally large villages; their siting is largely determined by the terrain, and stone is widely used in construction. In the plains, villages are evenly distributed, locally reflecting aridity and irrigation patterns, while in areas characterized by ravines the villages keep to the valley roads, and surrounding valley land is intensively utilized for vineyards, tea plantations, and other crops. The Black Sea coast is marked by a string of health resorts.

Urals and Siberia. In the Urals and Siberia, including the Far East, where only one-eighth of the total rural population is to be found, population distribution in the Far North differs substantially from that in the more settled part of the Main Belt of Settlement adjoining the Trans-Siberian Railroad. In the Far North the population distribution is similar to that referred to above in the description of the European North, and the rural settlements are even smaller and more widely dispersed (mainly following the rivers). A greater proportion of settlements is connected with mines. Ordinary villages of the same type as those in the middle zone of the R.S.F.S.R. predominate within the boundaries of the Main Belt of Settlement; but they are more widely separated and, within them, houses are spaced farther apart, reflecting the large Siberian expanses. There are more young settlements here (particularly in the state farm communities of the steppes), and settlements connected with industry and transport are common. Farmsteads are virtually nonexistent; in the recent past small groups of farmsteads were traditional for the Yakuts, but in the Soviet period rural dwellers have been concentrated in settlements of medium size, where it is easier to provide public services and to direct the collective economy. Although the rural population is denser near the railways and large towns, only a fraction of the whole territory is intensively utilized, and cultivated areas are separated by large areas of taiga and forest–steppe.

The young rural settlements of Siberia

Central Asia and Kazakhstan. In Central Asia and in southern and central Kazakhstan (rural settlement in north Kazakhstan is similar to that of southern Siberia) live about one-seventh of the Soviet rural population. The size and disposition of rural settlements are determined by the configuration and size of the areas used for agriculture, which are mainly irrigated. In the thickly populated oases, where water makes intensive farming possible, rural settlements (*qishlaq*s and *auly*) often are composed of several hundred households. The *qishlaq*s whose population is engaged in unirrigated agriculture or who specialize in livestock (raising Karakul sheep in the semideserts or raising free-ranging livestock in alpine and dry plain pastures) are much smaller. The appearance of *qishlaq*s is unique. The flat-roofed houses are built of raw bricks (occasionally of packed clay or reeds and clay); many of them have no windows onto the street, but the inner courtyard, which gives way to the orchard, usually has an open platform with an awning that is a place for dining, welcoming guests, and so on. There is often a small pond in the courtyard. The walls surrounding the house are thick. Because of the packed clay construction and the absence of windows, a village street in this region looks like a narrow clay corridor. New building, however, is in the main made to follow standard plans, with single-family dwelling houses of European type. Public buildings are new, meeting modern technical requirements but taking note of national architectural motifs.

*Qishlaq*s and *auly*

Urban population distribution. According to official data, town dwellers constituted 18 percent of the total population in 1926, 33 percent in 1939, 48 percent in

1959, and 62 percent in 1977, reflecting an absolute increase in urban population from 26,300,000 to 160,800,000. Apart from natural growth, the increase in the urban population was determined partly by migration from the villages to the towns and cities and partly by the transformation of some rural inhabited localities into towns and cities.

In most republics an inhabited locality will not be granted city status until its population reaches 10,000–12,000, and the population must be at least 3,000 before town status is conferred. A further requirement in both instances is that the overwhelming majority of the population not be engaged in agriculture. Since the criteria are not inflexible, there are about a thousand towns and several dozen cities with populations of less than 3,000.

In the mid-1970s there were about 2,000 cities and 3,700 towns, compared with 1,679 and 2,940, respectively, at the 1959 census. Between 1959 and 1975, the total urban population rose from 100,000,000 to 153,110,000, an increase of about 53 percent. The most rapid increase was in the number of people living in the larger cities. In 1975 some 87,700,000 people (57 percent of the total urban population) lived in cities with more than 100,000 inhabitants, as against 48,600,000 (49 percent) in 1959. Over the same period, the number of such cities rose from 148 to 240. The proportion of people living in small towns is declining, although their absolute number (as well as the number of small towns) continues to rise, reflecting new development of natural resources and the transformation of rural inhabited localities into urban settlements. Between 1959 and 1975 the number of urban centres with fewer than 10,000 inhabitants rose from 3,043 to 3,674 and their combined population from 14,400,000 to 16,800,000, but their share of total urban dwellers fell from 14 to 11 percent. Medium-sized towns and cities, with populations between 10,000 and 100,000, increased in number from 1,428 in 1959 to 1,838 in 1975, and their population rose from 37,000,000 to 48,600,000 over the same period. Their share of the urban total, however, declined from 37 to 32 percent.

Between 1959 and 1970 an increase of 36,000,000 in the urban population was made up of: urban natural increase, 14,600,000; migration from rural areas to urban centres, more than 16,000,000; transformation of rural inhabited localities into towns and cities, about 5,000,000.

Major cities

By the late 1970s, 15 Soviet cities had passed the 1,000,000 population mark. Moscow had a population of 7,819,000 (or 12,000,000 if the satellite towns of the Moscow region were added); Leningrad had 4,425,000, Kiev 2,079,000, and Tashkent 1,689,000 citizens; Baku, with its suburbs, had about 1,435,000; Kharkov, Gorky, and Novosibirsk had about 1,300,000 each; Kuybyshev, Sverdlovsk, and Minsk had about 1,200,000 each; and Odessa, Omsk, Chelyabinsk, and Tbilisi had just over 1,000,000. Another 26 cities had populations between 500,000 and 1,000,000, the largest of these being Omsk, Dnepropetrovsk, Donetsk, Perm, Kazan, Yerevan, Ufa, Volgograd, Rostov-na-Donu, Alma-Ata, and Saratov. More than three-quarters of the total urban population is to be found in the half dozen highly urbanized zones listed below, and the great bulk of the cities and towns fall within the Main Belt of Settlement and the outlying population concentrations of Central Asia and Transcaucasia. Quite a few urban nuclei, however, including some of considerable size (*e.g.,* Murmansk, Arkhangelsk, Norilsk, Bratsk, and Komsomolsk-na-Amure in the north and Karaganda, Astrakhan, Orsk, and Semipalatinsk in the south), lie in isolated situations outside these zones.

There are six major zones of urban concentration: (1) the Moscow region together with the adjacent cities on the Upper Volga from Kalinin through Yaroslavl and Kostroma to Gorky; (2) the zone stretching from Leningrad southwestward across the Baltic republics and into northern Belorussia, including Riga, Vilnius, Tallinn, Klaipėda, Kaliningrad, and the Belorussian cities of Minsk, Vitebsk, and Mogilyov; (3) the cluster of cities on the Donbass coalfield, where the urban population of two *oblasti*—Donetsk and Voroshilovgrad—totals nearly 7,000,000; (4) the cluster of cities around the Dnepr bend, with a total population of 4,000,000; (5) the Urals group of metallurgical and mining cities, notably Sverdlovsk, Chelyabinsk, Perm, Ufa, Nizhny Tagil, and Magnitogorsk; and (6) the cities of the Kuznetsk basin, including those in the Kemerovo *oblast* together with nearby Novosibirsk.

Role of administrative centres

The administrative centre of the *oblast* or the capital of the union or autonomous republic is nearly always the largest city—there are only isolated exceptions in the entire country. The fact that they are administrative centres attracts to these cities institutions of higher education, public bodies, associations, and so on, all of which contribute to increasing their population. Thus most of the largest cities combine many functions—productive, administrative, and social. Cities with only one predominant function rarely grow to considerable size. Smaller urban settlements may serve only one function (sometimes connected with a single enterprise—a mine or a fish-processing plant, for example). There are also smaller settlements that act as the centres of subordinate administrative units; they have several functions, housing the bodies concerned with administration and communications within the district and also serving the surrounding rural area, processing local raw material and producing consumer goods.

Restrictions on city growth

Although the largest cities as a group grew most rapidly between 1959 and 1975, some of the largest and oldest centres grew relatively slowly. Compared with an overall urban increase of 53 percent, the population of Moscow rose by only 26 percent, that of Leningrad by 29 percent, and that of Gorky by 36 percent. This was attributable to a specific policy of restricting their growth for ecological and social reasons. The concept of "central city—green zone for recreation—peripheral satellite towns" is regarded by the authorities as much more suitable for promoting economic growth than the piling of an ever-growing number of industrial enterprises into a central city. Some of the highest growth rates in the period since 1959 have been seen in medium-sized cities, especially those in which big industrial centres are being established, and sometimes the city itself (*e.g.,* Bratsk, Volzhsky, Tolyatti) is virtually built anew. Along with these recently founded cities, several old cities have greatly expanded or changed their productive functions and have thus also grown twofold or more. They include Balakovo, Belgorod, Kirovakan, and Novgorod.

II. The people

NATIONAL COMPOSITION

The Soviet population is composed of more than 100 distinct national groups. Soviet ethnographers and demographers generally consider it inaccurate to assign national-ethnic affiliation on the basis of mother tongue, since cases of persons using as their basic everyday language the language not of their own people but of the people among whom they live, though at the same time preserving awareness of their own national identity, are fairly common; according to the 1970 census, 6.1 percent of the Soviet people indicated the language of another nationality as their mother tongue. Accordingly, Soviet censuses establish nationality and mother tongue as separate and independent indexes (according to the awareness of the person interviewed, as he declares it). Provision is also made for the recording of any second language spoken fluently.

On the basis of ethnic territories occupied (and in most cases occupied since ancient times), the Soviet Union is composed of one Soviet Federated Socialist Republic (the Russian), which includes 16 Autonomous Soviet Socialist republics, five autonomous *oblasti,* and 10 autonomous *okruga* (districts); and 14 Soviet Socialist republics, which include four A.S.S.R.'s and five A.O.'s. Yet the total of such entities does not coincide with the total number of Soviet nationalities; there is much overlap of one kind or another. The Dagestan A.S.S.R., for example, in the R.S.F.S.R., is inhabited by no fewer than 20 nationalities, within which are consolidated many smaller national groups. The ethnic diversity of the Soviet popu-

lation is, however, officially recognized in the composition of these political entities.

The eponymous nationalities of (*i.e.*, those that give their name to) the union republics comprise 90 percent of the total Soviet population. The most numerous are Russians (53.4 percent of the 1970 total), followed by Ukrainians (16.9 percent), Uzbeks (3.8), Belorussians (3.7), Tatars (2.5), Kazakhs (2.2), and numerous other smaller groups of less than 2 percent. Even smaller but still administratively distinct nationalities include the Nenets, who number only 29,000 but, being scattered over huge tundra areas, are grouped in three autonomous districts. Nationalities without a distinct ethnic area and hence without administrative status include Koreans, Poles, Bulgarians, Greeks, and Gypsies.

Table 2: The Eponymous Nationalities of the Union Republics, 1970

nationality	numbers in Soviet Union (in 000)	percentage of population in their republic	percentage of total nationals in Soviet Union living in their republic
Russians	129,015	82.8	83.5
Ukrainians	40,753	74.9	86.6
Uzbeks	9,195	64.7	84.1
Belorussians	9,052	81.0	80.5
Kazakhs	5,299	32.6	78.5
Azerbaijanis	4,380	73.8	86.2
Armenians	3,559	88.6	62.0
Georgians	3,245	66.8	96.5
Moldavians	2,698	64.6	85.4
Lithuanians	2,665	80.1	94.1
Tadzhiks	2,136	56.2	76.3
Turkmens	1,525	65.6	92.9
Kirgiz	1,452	43.8	88.5
Latvians	1,430	56.8	93.8
Estonians	1,007	68.2	91.9

A comparison of the figures in the two right-hand columns of Table 2 shows that the bulk of each nationality is concentrated in its corresponding republic. The Armenians are something of an exception: although the proportion of Armenians in the Armenian S.S.R. (88.6 percent) is the highest of all, fewer than two-thirds of all Soviet Armenians actually live in the union republic.

A cause for disquiet in some of the non-Russian Soviet Socialist republics is immigration of Russians and, furthermore, their concentration in the cities. At the time of the census of 1970, for example, Kazakhs constituted only 32.6 percent of the population of Kazakhstan and Russians made up 42 percent. Of the population of Kirgiziya, 44 percent was Kirgiz, 29 percent Russian.

Language families and groups

It is Soviet practice to classify nationalities according to their affinity with language families and groups within families. Soviet nationalities speak tongues related to seven language families; in addition, because of their isolation, the languages of Koreans (360,000) and Kets (altogether about 1,000) may be treated as linguistic families, while Paleosiberian languages fairly remote from one another are brought together formally at the family level (see Table 3); some authorities consider Ket to be a Paleosiberian language also.

The Slavic language group predominates: about three-quarters of the population speaks Slavic languages (not counting those who speak one of these languages fluently as a second language). Russian was named as their mother tongue by 141,800,000 persons in the 1970 census, and a further 41,900,000 named it as a second language; thus, taking second languages into account, Russian is the language of cultural intercourse for 183,700,000 people, or more than three-quarters of the total population. Ability to speak Russian is necessary for career advancement for non-Russians. The Turkic language group, spoken by about one-eighth of the total population, ranks next in number of speakers.

SOCIAL CHARACTERISTICS

Religion. It is, and has been since 1918, official Soviet policy to eradicate religion, viewed by doctrinal Communism as the "opiate of the people." In that year, by decree, church and state were separated and all church property was confiscated. A church has no legal status, and believers who wish to have a place to worship must sign a contract with the local soviet, which has discretionary power to permit such use of a building. Private instruction in religion is hemmed round with restrictions.

Church membership

No official statistics on church membership or religious belief are published, but Western scholars have estimated that believers make up about 40 percent of the population of the Soviet Union; many of them, however, because of the official attitude, never attend services. The number of Russian Orthodox is estimated at 50,000,000, Muslims (mainly Turkic peoples and some in the North Caucasus) at 30,000,000. Roman Catholics (chiefly Lithuanian), Eastern-rite Catholics (Ukrainians and Belorussians), Lutherans (Latvians and Estonians), and Baptists also number several million, and there are about 2,100,000 Jews and about 500,000 Buddhists. The supreme catholicos of the worldwide Armenian Church resides in Echmiadzin, in the Armenian S.S.R.

Nationality and education. It is stated in Soviet law that all nationalities enjoy equal rights in terms of participation in the country's economic, cultural, and public life. An important precondition for the attainment of such a condition is an equalizing of cultural levels, a major objective of Soviet policy since the 1920s. Literacy is virtually universal, except among some older people (in 1970, 99.7 percent of those aged nine to 49 were classified as literate). Among the working population almost two-thirds have completed secondary or higher education. The non-Slav nationalities, however, continue to be at a disadvantage in terms of education. In 1970, for example, Kazakhs constituted more than 4 percent of the total Soviet population but less than 2 percent of graduate students. Among the Kirgiz, 98.3 percent over the age of 60, 91.8 percent between age 55 and 59, and 79.5 percent between age 50 and 54 had had no secondary education. Within the overall context of Soviet education (see below *Administration and social conditions*), the widespread training of specialists is assisted by the many territorial branches of the national education system. An increasing number of scientific and other centres—including the internationally famous Akademgorodok (Academic Community) of Novosibirsk—are sited far from the European heartland of the country.

Employed and pensioned

Economic categories. The 1970 census established that a total of 115,500,000 persons, or 47.8 percent of the total population, were actually employed (not counting those working only in the household or in personal ancillary agriculture). For 33,100,000 persons (13.7 percent) the main source of income was a pension paid by the state. The total number of pensioners at the beginning of 1970 was 40,100,000, but some of these continued to work and were receiving both a pension and wages. More than 3,500,000 students (1.5 percent) were receiving scholarships from the state. The remainder of the population (89,100,000 persons, or 36.8 percent) was composed of persons who were not employed in the national economy but were engaged in housework and bringing up children, or were children themselves; and of members of the families of collective farmers, workers, and employees, engaged in personal agriculture.

The distribution of the employed population by nature of work is estimated as follows: 75.4 percent in branches of material production (including personal ancillary agriculture); more specifically, about 60.8 percent are workers and employees, 12.7 percent collective farmers, and about 1.9 percent members of the families of collective farmers, workers, and employees, engaged in personal ancillary agriculture. The remaining 24.6 percent or so work in nonproductive branches (services, education, health, the arts, science, and literature, administrative bodies, and so on). In terms of specific branches of the economy (and not taking account of persons studying and doing military service), industry and construction account for 43.7 percent of the total; agriculture and forestry, 10.8 percent; transport and communications, 10.5 percent; trade and related activities, 8.6 percent; health, education, culture and the arts, science, social security,

and sport, 18.7 percent; state administrative apparatus (including financial operations) and public organizations, 2.7 percent; and the remainder—housing and other municipal and public services—some 5 percent.

DEMOGRAPHY

In the absence of large-scale immigration or emigration, the rate of growth of the Soviet population as a whole is attributable almost entirely to natural increase (*i.e.*, the excess of births over deaths). Regional population growth rates, however, are influenced not only by regional variations in the rate of natural increase but also by inter-regional migrations, which occur on a large scale.

Natural rates of increase. Social factors have had a considerable influence on the natural increase of the Soviet population. The most important factors have been a substantial decline in the average birth rate as a result of rapid urbanization (fertility is appreciably lower in the towns than among the rural population); the widespread employment of women, on the same terms as men, in nearly all branches of economic activity, including a rise in the percentage of trained female professional workers, which has also contributed to a fall in the birth rate; provision of state aid for mothers to assist them in the raising of children, which has tended to check birth rate decline; and, not least, a progressive rise in standards of health. This last factor has influenced the natural rate of increase in a number of ways: the decline in infant mortality has not only reduced the overall death rate but has also reduced the psychological "insurance" attained by giving birth to additional children and thus has tended to reduce the birth rate; in addition, reduced mortality at all ages and a subsequent increase in life expectancy have increased the proportion of the older, non-childbearing element in the population. The net result has been a long-term decline in both birth and death rates.

Birth rate

In the 1960s, birth rate decline was particularly marked, while the death rate stabilized at a low level and then rose slightly as the average age of the population increased. This contrasts with the situation at the beginning of the 20th century, when the demographic model of the country involved a very high birth rate and a fairly high death rate, producing a rate of natural increase much higher than that of most European populations; moreover, the change of generations was relatively rapid. The present model is based on a modest birth rate and a very low death rate, giving a rate of natural increase about half the world average but still significantly higher than that in the advanced countries of Europe.

The Soviet demographic model was greatly distorted by the enormous population losses incurred during World War II. Direct losses, both in the ranks of the military forces and among the civilian population (especially in the Nazi-occupied territories), exceeded 20,000,000. If the indirect losses—fewer births than usual as well as greatly increased mortality associated with poor living conditions—are taken into account, then the total population deficit resulting from World War II may be put at approximately 50,000,000. These war losses greatly deformed the age and sex structure (see below), with results still visible in the 1970s. In addition, the accelerated decline in the birth rate that occurred during the 1960s was in part caused by the arrival at marriageable age of the small wartime generation.

In 1976 the annual average number of births per 1,000 population was 18.4 (as against 47 in 1913) and that of deaths was 9.5 (30.2 in 1913), giving a natural increase rate of 8.9 per 1,000 (16.8 in 1913), little more than half the level of the early 1960s. The birth rate in towns, at about 17.2, was notably lower than that in rural areas (20.3), though the gap had been greatly reduced. Another major long-term trend has been the decline in infant mortality; the number of children dying in their first year was 273 per 1,000 live births in 1913, 182 in 1940, and only 28 in 1974.

The rate of natural increase varies greatly between republics, mainly as a result of variations in the birth rate. In 1976 natural growth in the R.S.F.S.R. was only 5.9 per 1,000 population (with 15.9 births and 10 deaths) and even lower in Estonia and Latvia (natural increase 3.1 and 1.7, respectively). In several southern republics, however, it was very high; in Tadzhikistan it reached 29.7 (with 38.2 births and 8.5 deaths), and it was also above 20 per 1,000 in Uzbekistan, Kirgiziya, and Turkmenistan. These territorial differences reflect ethnic differences. While many groups are increasing their numbers, however, others are declining as a result of assimilation; this is particularly true in the case of peoples who do not have their own ethnic territories and are dispersed in close contact with peoples of other cultures.

Age and sex structure. The demographic trends described above have determined the sex and age structure of the Soviet population. At the time of the census of 1970, there were 130,500,000 persons (54 percent of the total population) of working age (defined as 16–59 for men and 16–54 for women, as persons above these ages receive pensions). In the overall population, 46.1 percent were men and 53.9 percent women (the ratio was 46.3:53.7 in towns and cities and 45.8:54.2 in rural areas). The preponderance of women was attributable to their predominance in the older age groups, especially those that suffered most in the war; among those aged 45 or older there were only 547 men per 1,000 women. In the young groups, there was a higher percentage of married women than married men: in the 20–24 age group, 55.9 percent of the women and only 28.9 percent of the men were married, compared with 37.1 percent and 92 percent, respectively, in the 60–69 age group.

Marriage rates

The role of migration. Some of the results of the movement of population from rural areas to towns and cities are considered above in connection with the marked trend of urbanization (see above *Settlement patterns: Urban population distribution*). Further demographic aspects of this process may also be seen in a comparison of movement in the three intercensal periods 1927–38, 1939–58, and 1959–70.

Taking into account that the first period covered 12 years, the second 20, and the third 11 and that total population grew over the whole period, it may be concluded that migration from country to town is somewhat slackening. In the early 1970s the natural growth of the urban population was already approximating the net in-migration; the natural growth was to a considerable degree maintained precisely by that migration, for it is mostly young people who move to the towns, where they create their families.

The largest and most rapidly growing towns, including those at the sites of new construction projects, have attracted people not only from localities nearby but also from remote regions. Thus country-to-town movement is accompanied by inter-region and even inter-republican migration. The motive force in such a widespread redistribution of population is the creation of new jobs in economically developing regions; the state offers incentives (particularly in the Far North) to draw potential migrants to areas with manpower shortages. The regions that most attracted settlers in the 1959 to 1970 period were, above all, Kazakhstan and Central Asia, where, although the natural local population growth is already high, several million migrants (mainly from the R.S.F.S.R.) settled. On the other hand, the movement of population to Siberia from the European regions was much less than in previous periods and affected only certain areas. The 20 percent growth experienced in the Far East (as against 11 percent for the R.S.F.S.R. as a whole) was attributable largely to migration, as was the 29 percent growth in the Tyumen *oblast* of Western Siberia, where there was intensive development of newly discovered oil and gas resources. Migration also took place to the Baltic republics, where high economic potential brought about a shortage of manpower. The regions providing settlers were mainly the Volga–Vyatka region (where the numbers of emigrants almost reached the amount of the natural growth), the Black-Earth and Central Industrial regions, and Belorussia. In the Ukraine, internal movement was from the Southwest to

Regions of out-migration

Table 3: Distribution of Nationalities in the Soviet Union According to Language Families and Groups

	Indo-European family			Caucasian family		Uralic family	
	Slavic	Iranian	other language groups	South Caucasian (Kartvelian)	North Caucasian	Finno-Ugric	Samoyedic
More than 100,000,000	Russians*						
10,000,000–100,000,000	Ukrainians*						
5,000,000–10,000,000	Belorussians*						
1,000,000–5,000,000	Poles	Tadzhiks*	Armenians* (Armenian group), Moldavians* (Romance group), Lithuanians* and Latvians* (Baltic group), Germans (Germanic group), Jews‖ (Yiddish; Germanic)	Georgians*		Mordvins†, Estonians*	
300,000–1,000,000	Bulgarians	Ossetes†§	Greeks (Greek group)		Avars‡, Lezgians‡, Chechens†	Udmurts†, Maris (Cheremis)†, Komis (Zyryans)†	
100,000–300,000			Gypsies (Indic group), Romanians (Romance group)		Darginș‡, Ingushes†, Kabardians†, Adyghians§	Komi-Permyaks‖, Kareliansț, Hungarians	
50,000–100,000		Kurds			Lakks, Tabasarans, Abkhaz†	Finns	
10,000–50,000	Czechs, Slovaks	Iranians, Tats			Rutuls, Tsakhurs, Circassians§, Abaza	Khants (Ostyaks)‖	Nenets‖
5,000–10,000					Aguls	Mansis (Voguls)‖	
Fewer than 5,000						Saamians	Selkups, Nganasans

*People forming a Soviet Socialist Republic. †People forming an Autonomous Soviet Socialist Republic. ‡The largest nationalities of the Dagestan A.S.S.R.

the Donets–Dnepr area and to the South. The characteristic link between migratory influx and a higher level of natural growth and between efflux and a lower level of natural growth is explained by the fact that it is mainly young persons that migrate.

Population density. Among the economic regions of the R.S.F.S.R. the Central Industrial region is the most populous (28,136,000) and has the highest population density; the proportion of urban population there is only slightly less than in the northwest, where it is highest of all (73 percent). The urban population also reaches a very high 71 percent in the Far East. In the Ukraine, the most densely populated areas are the industrial regions of the Donets–Dnepr area—the 5,000,000 people of the Donetsk *oblast* are 89 percent urban, and the figures for the Dnepropetrovsk and Voroshilovgrad *oblasti* are only very slightly lower (see Table 4 for corresponding data for the union republics).

The demographic outlook. Soviet scientists expect the population to reach 333,000,000 (± 25,000,000) by the year 2000. The overall natural reproduction will probably be marked by a not very high birth rate, a very low death rate, and a moderate net growth. Life expectancy will rise to 80–85 years, and three-quarters of the population may be urban. According to published Soviet forecasts, by the time the total population rises to 300,000,000 there will be 205,000,000 town dwellers (almost 70 percent); moreover, almost one-quarter of the population will be concentrated in the largest cities (those with more than 500,000 inhabitants) and almost one-third in large cities (100,000–500,000 inhabitants). The most rapid growth is forecast for this last category. A redistribution of the population in certain regions—specifically, a growth in the share of the eastern and southern regions—may also be anticipated. In rural areas, the size of inhabited localities will continue to increase, while their numbers will be reduced; by the year 2000 the bulk of the remaining rural population will probably be concentrated in well-equipped settlements with populations of at least 1,500 to 3,000 and with living conditions close to those in towns. Such an abolition of the substantial differences between urban and rural settlements is, at least, one of the major long-term Soviet goals.

(V.V.P./J.C.D.)

Moderate net growth

III. The national economy

THE SOVIET ECONOMY IN WORLD CONTEXT

Although considerable economic development took place between the 1860s and World War I, the economy of tsarist Russia lagged behind those of the western European industrialized nations and the United States. Since the early 1920s, the economy of the Soviet Union has

	Altaic family			Hamito-Semitic	Sino-Tibetan	Eskimo-Aleut	also at family level or regarded as equivalent		
	Turkic	Manchu-Tungus	Mongolian				Paleo-siberian	Korean	Ket
More than 100,000,000									
10,000,000–100,000,000									
5,000,000–10,000,000	Uzbeks*, Tatars†, Kazakhs*								
1,000,000–5,000,000	Azerbaijanis*, Chuvash†, Turkmens*, Kirgiz*, Bashkirs†								
300,000–1,000,000			Buryats†					Koreans	
100,000–300,000	Yakuts†, Kumyks‡, Kara-Kalpaks†, Uighurs, Gagauzis, Tuvinians†, Karachays§		Kalmyks†						
50,000–100,000	Altais§, Nogays, Turks, Khakass§, Balkars†								
10,000–50,000	Shors	Evenks‖, Evens (Lamuts), Nanais		Assyrians	Dungans		Chukchis‖		
5,000–10,000				Arabs			Koryaks‖		
Fewer than 5,000	Dolgans‖	Ulchis, Udegeys, Orochis				Eskimos, Aleuts	Gilyaks, Yukaghirs, Itelmens (Kamchadals)		Kets

§People forming an autonomous *oblast*. ‖People forming an autonomous *okrug*.

developed into the most powerful in the world after that of the United States. Since 1928 the course of economic development has been guided by a series of ambitious five-year plans.

The agricultural sector has become one of the world's principal sources of cereals and of a number of other farm, forestry, and fisheries products. Western specialists nevertheless regard the sector as relatively inefficient and underdeveloped, and Soviet authorities admit that they are far from satisfied with the level of some farms and some branches of agriculture, although they would doubtless emphasize that, by the 1970s, some others compared favourably with the best in the world.

Minerals and industry

In the world economy, however, the Soviet Union is more important for its mineral resources and industrial production. The country produces a very wide range of minerals and boasts huge reserves of coal, iron ore, copper, petroleum, and natural gas. Geography and climate make many of these reserves difficult to work, but the development of several new fields has placed the nation among the leading producers of iron ore, coal, oil, natural gas, and manganese. By 1975 the value of Soviet industrial production was estimated to be second only to that of the United States. The country manufactures an immense range of products, although large quantities of technologically advanced equipment are imported from capitalist countries. The nation is also among the world's three top manufacturers of a large number of basic and heavy industrial products. As a result of historical factors that made emphasis on consumer industries a relatively recent development, however, it tends to lag behind in the output of light industrial products and consumer durables. One result of this has been that consumer demand is only partially satisfied and that the quality of the goods available is, in some cases, poor when judged by Western standards.

Despite a rapid increase in industrial investment in the past, productivity and industrial technology are still judged to be below the level of the most advanced capitalist countries.

Role in world trade

Judged in Western terms, the role played by the Soviet Union in world trade is considerably smaller than its economic strength would justify. It has been estimated that it exports only some 3 percent of its national production and accounts for about 4 percent of total world trade—figures that would be more than halved if its trade with other Communist countries were excluded. Although the country has extensive influence—heightened by its foreign-aid program—in some developing countries, the principal international effects of Soviet economic power are felt in the Communist world. The eastern European economies in particular are closely

Table 4: Urbanization and Population Density in the Union Republics, 1977

	population (000,000)			percentage of population urban	density persons per sq mi	density persons per sq km
	total	urban	rural			
R.S.F.S.R.	135.60	93.72	41.88	69	20.5	7.9
Armenia	2.89	1.86	1.03	64	251.5	97.1
Azerbaijan	5.78	3.01	2.77	52	172.8	66.7
Belorussia	9.41	5.01	4.40	53	117.3	45.3
Estonia	1.45	1.00	0.45	69	83.1	32.1
Georgia	5.00	2.54	2.46	51	185.7	71.7
Kazakhstan	14.50	7.88	6.62	54	13.7	5.3
Kirgiziya	3.44	1.34	2.10	39	45.8	17.3
Latvia	2.51	1.67	0.84	67	102.0	39.4
Lithuania	3.34	1.95	1.39	58	132.9	51.3
Moldavia	3.88	1.47	2.41	38	298.6	115.3
Tadzhikistan	3.59	1.30	2.29	36	65.0	25.1
Turkmenistan	2.65	1.28	1.37	48	14.0	5.4
Ukraine	49.30	29.84	19.46	61	211.6	81.7
Uzbekistan	14.47	5.71	8.76	39	83.9	32.4
Soviet Union*	257.82	159.59	98.23	62	30.0	11.6

*Detail may not add to total given because of rounding.

integrated with that of the Soviet Union, which is the source of more than two-thirds of their imports and accounts for an even larger share of their raw material purchases.

In spite of rapid economic development in the past—a process carried out against a scarcely propitious internal and external background—Soviet living standards in the mid-1970s were lower than in some capitalist and certain Communist countries. Western economists estimate that the Soviet per capita national income puts the country somewhere between 20th and 30th place in the world. Soviet authorities aver that such a reckoning is somewhat misleading as it does not take into account the low rate of income tax, the low rents (generally about 5 percent of family income), and the provision of such free services as medicine and education.

THE EXTENT AND DISTRIBUTION OF RESOURCES

Mineral resources. *Coal reserves.* One of the most important of Soviet mineral resources is coal. Estimates of workable reserves vary; one frequently quoted official figure is 7,800,000,000,000 metric tons, of which about 5 percent is proved and 15 percent probable. A quarter of the output is accounted for by brown coal and lignite. Anthracite production has become more or less static. The country's coal output—which was scheduled to increase during the 10th Five-Year Plan—is more than enough to meet domestic requirements, and large quantities are exported, mainly to nations of eastern Europe. Reserves are widely scattered, but about a third of total output comes from three main producing areas. The most important is the Donets Basin (Donbass) region, where the seams are thin and difficult to work but the coal is of high quality. The region is also well situated in relation to a number of industrial centres. In second place is the Kuznetsk Basin region, with thicker seams and some coking coal. It is followed by the scattered fields of the Urals, which are easy to work but produce large quantities of lignite. Considerable amounts of lignite are also derived from the Moscow Basin. Other important coal-producing areas include Eastern Siberia, Karaganda and Ekibastuz in Kazakhstan (producing large quantities of coking coal), and the Pechora field in the Arctic, centred on the town of Vorkuta. Small quantities are also mined in Central Asia, Georgia, and a number of other places. There are huge and virtually unworked reserves in the Tunguska and Lena basins of the Central Siberian Plateau.

Donbass

Petroleum and natural gas reserves. According to official Soviet figures, known petroleum reserves total almost 70,000,000,000 metric tons—enough for about 200 years of domestic consumption at the current rate of extraction—and there are indications of a further 300,000,000,000 metric tons beneath land and sea. Annual output doubled between 1965 and 1975; it is more than adequate to meet domestic requirements, and large quantities are exported. The largest producing area is the Volga–Urals region, which accounts for 55–60 percent of total production. Other important fields are situated at Baku in Azerbaijan, in the North Caucasus, and in Western Siberia. Baku has lost much of its previous importance, but Western Siberia, where deposits were discovered as recently as the 1960s, has been developing rapidly. Lesser producing areas include the Emba River district, the Mangyshlak Peninsula of Kazakhstan, and western Turkmenistan (all on the Caspian Sea), the Fergana Basin in Central Asia, the Pacific island of Sakhalin, the Pechora Basin, the western Ukraine, and the Rechitsa district of southern Belorussia. Large untouched reserves lie in Eastern Siberia, mainly in the Lena Basin of Yakutiya.

Production of natural gas has increased even more rapidly than that of oil. Output increased more than sixfold from 1960 to the mid-1970s, and the 10th Five-Year Plan called for further increase. The principal gas fields are situated in the Volga–Urals region, the Ukraine, Central Asia, and the North Caucasus. A major new source was under development in the northern part of the West Siberian Plain in the mid-1970s. Shale oil is produced in relatively small quantities around Lake Peipus in Estonia and is also found in the Volga Valley and eastern Kazakhstan.

Metals. Soviet reserves of iron are estimated to account for about 40 percent of the world's total. Between 1960 and 1975, annual output more than doubled, to make the country the world's largest producer. Iron ore is mined in a number of different places, but some 70 percent of total output comes from two areas, the Ukraine (especially the fields at Krivoy Rog) and the Urals. Small quantities of ore are also derived from scattered sites elsewhere. Most ores are of fairly high quality; according to the United Nations, the average metal content of total Soviet output is approximately 60 percent. Intensive development of lower grade ores, notably those of the Kursk Magnetic Anomaly (in the Black-Earth Belt) and the Kustanay *oblast* of northwestern Kazakhstan, began in 1975.

Ores for iron and steel

The Soviet Union is also the world's leading producer of manganese. This is mined in two principal areas, Nikopol in the Ukraine—which yields relatively poor-quality ore but accounts for more than 50 percent of total output—and Chiatura in Transcaucasia, which produces a higher grade ore. Output (with an estimated manganese content of about half) is more than adequate to meet domestic demand, and large quantities are exported, especially to eastern Europe but also to some Western countries. The Soviet Union has large reserves of ores of several other important components of steels, including nickel, cobalt, molybdenum, vanadium, tungsten, and, in the Urals, chromium (sufficient to allow substantial exports).

Ores for nonferrous metallurgy

There are also large copper resources in several parts of the country. The principal producer is Kazakhstan, but there are also important mines in parts of the Urals as well as in Uzbekistan and Armenia. Monchegorsk, in the Kola Peninsula region, and Norilsk also yield quantities of copper ore, and there are huge unworked reserves—estimated to be capable of producing 90,000 tons of ore a day—around Udokan in Siberia. Output of copper is second only to that of the United States. The most important zinc and lead deposits are situated in the Altai Mountains of eastern Kazakhstan, where they occur in association with copper, and at several sites in the mountains of Central Asia and the Caucasus.

The Soviet Union is not particularly rich in bauxite; there are several rather low-grade deposits in Asia (mainly in Kazakhstan) and in the European northwest, where a major new source was under development in the Onega Valley in the mid-1970s. Aluminum production, while relying to some extent on imported bauxite (from the Balkans) and alumina (from Hungary), is chiefly supported by the use of alternative ores—nephelite from the Kola Peninsula, nephelite and alunite from Transcaucasia, and kaolinite from Central Asia. Titanium is produced in the Ukraine and in the Urals; and uranium is derived from Uzbekistan, Tadzhikistan, and Kirgiziya.

The rarer metals and minerals

The country is thought to be the world's second largest producer of gold, and a large number of other metals—including, among others, platinum, beryllium, columbium (niobium), and mercury—are also produced on a significant scale.

Nonmetallic minerals. There are large phosphate deposits at Kirovsk in the Kola Peninsula, at Karatau in Kazakhstan, in the Donbass, and around Moscow. Potash is mined in the Urals, and there are also a number of pyrite deposits. The Urals area also yields large quantities of asbestos (especially north of Sverdlovsk) and some graphite. Other important nonmetallic minerals include kaolin, mica, fluorspar, and salt.

Biological resources. Taking the vast expanse of Soviet territory as a whole, much of the soil is unsuitable for agriculture, and the climate—characterized by long, cold winters, poor water distribution, and much unpredictability in most parts of the country—is also adverse. The northernmost belt, stretching from the Kola Peninsula eastward toward Kamchatka, consists of economically unimportant barren tundra, where the subsoil is permanently frozen.

Timber resources

The northern part of European Russia, much of Western Siberia, and most of the area east of the Yenisey River is covered with forests (see above *The land: The natural landscape: Vegetation*). All in all, about 40 percent of the country's total land area is covered by forests. Total timber resources are estimated at 2,800,000,000,000 cubic feet (80,000,000,000 cubic metres), of which approximately four-fifths are coniferous. The Soviet Union is the world's leading timber producer, with Siberia and the Far East accounting for about one-third of the total output.

Agricultural activities are concentrated in an irregular and narrowing belt stretching from the Ukraine, Belorussia, and the Baltic lands toward Omsk, Karaganda, Novosibirsk, and Barnaul, and extending into Eastern Siberia. All in all, agricultural land accounts for 27 percent of Soviet land territory. Approximately 40 percent of the agricultural area consists of arable land; the remainder is made up of meadows and pastures.

SOURCES OF NATIONAL INCOME

Agriculture and fisheries. The agricultural sector (see also above *The land: Settlement patterns*) provides just over 14 percent of the country's net material product and employs some 23 percent of the economically active population. Contrary to the practice in several other Communist countries, farming in the Soviet Union is fully socialized, the two basic agricultural units being the collective and the state farm (*kolkhoz* and *sovkhoz,* respectively).

The only obligation that the collective farm has is to meet a delivery quota. It makes its own decisions on profit sharing. Weak and unprofitable collectives are being amalgamated into larger collectives or turned into state farms. As a result, their numbers have shown some decline; thus in 1976 there were 27,700 collectives compared with 36,900 in 1965. They accounted for some 13,300,000 farm households—an average of about 486 per collective—and their agricultural land area represented nearly 35 percent of the total. Just over 50 percent of this was arable land; the rest was meadows and pastures. The average collective of 16,300 acres (6,600 hectares) owned 39 tractors, 1,715 cattle, 1,837 sheep and goats, and about 957 pigs. Some 14,700,000 acres (5,900,000 hectares) of the collectives' land area is distributed in the form of small private plots among members, who are also allowed to own some livestock. Although these plots account for little more than 3 percent of all Soviet agricultural land, they account for a significant proportion of the output of a number of agricultural products (*e.g.,* over half the eggs, a third or more of total meat and milk). The collectives as a whole (which provide pasture and feed for the individual plots) concentrate, Soviet sources emphasize, on the production of grain and cotton, sugar beets, and related industrial crops.

State farms are owned and operated by the state, and their employees are paid a regular wage which is not usually significantly affected by their performance. Since the mid-1950s, the number and importance of state farms have progressively increased. By the mid-1970s, there were more than 19,600, and their agricultural area (one-third of it arable) made up more than 60 percent of the Soviet total. Their size varies considerably, but on average they employ some 550 workers and have about three times as much farmland as the average for the collectives. State farms account for rather less than 50 percent of the output of most agricultural products, and in a number of cases (notably cotton, sugar beets, and sunflower seeds) their share is very much smaller than that of the collectives.

As in most other countries of the Communist bloc, agriculture, in the opinion of Western experts, has been the Achilles' heel of the Soviet economy. Growth in agriculture has not kept up with that of the rest of the economy; gross agricultural production in the 1966–70 plan period was 22 percent above that of 1961–65, but output in 1971–75 was only about 14 percent above the 1966–70 level. Among agriculture's principal problems have been the lack of adequate resources, the absence of effective material incentives, and the low level of earnings; there has been some improvement in the situation, however.

Material incentives in agriculture

After much discussion, efforts have been made to give greater independence and responsibility to both collective and state farms and to provide some material incentives; bonus payments for target overfulfillment are widespread, and there are experiments to create within the state farms and the collectives small units that are given a degree of independence and are paid according to their own achievements as opposed to those of the collective or the state farm. The ninth and 10th Five-Year plans put considerable emphasis on agriculture. Investments were planned to continue at a high level, resulting in a good increase in the supply of agricultural machinery; further steps were to be taken to encourage initiative and to provide incentives for target overfulfillment.

Despite undoubted improvements, the farming sector is still relatively underdeveloped if compared with those of many other—including some Communist—countries. Soviet use of mineral fertilizers is still one of the lowest among the members of Comecon (the east European trade group), and agricultural machinery is still in short supply and, in some areas, relatively poor in quality. Notwithstanding a spectacular rise in productivity, labour is still inefficiently used, especially in the collectives. As a result, and as a consequence of poor climatic conditions, yields tend to be low. Productivity in livestock also leaves much to be desired; the difficulties here are aggravated by an almost permanent shortage of high-grade fodder.

Of the total crop area, about 60 percent is devoted to grain, nearly half of it sown to wheat. Fodder crops occupy another 30 percent, and the area devoted to industrial crops—mainly cotton, sugar beets, and sunflower seeds—constitutes only about 7 percent of the total. Most forms of agricultural production expanded more rapidly in the late 1960s and early '70s than in most earlier periods; there was a 37 percent increase in the total average annual value of agricultural output between 1961–65 and 1971–75. In terms of volume, the largest increases were achieved in cotton, grain, meat, and eggs, the smallest in sunflower seeds, potatoes, wool, and flax. The number of livestock also increased markedly. More than a fifth of all livestock was owned by individuals as opposed to collective and state farms.

The Soviet Union accounts for about 14 percent of the world fish catch.

Industry. Because of its wealth of mineral resources and its long-standing policy of giving priority to industrial development, the Soviet Union is among the world's foremost industrial powers. Although international comparisons of industrial production are always subject to qualifications, it is estimated that, in absolute terms, Soviet total industrial output is second only to that of the United States, and it exceeds that nation in production of

some individual items. The country's industrial strength is also illustrated by its ranking as the second or third world producer of a wide range of important industrial products.

The status of technology

Although Soviet industry can claim a number of spectacular technical achievements in certain fields, including space exploration, it is still well behind the leading capitalist countries in terms of productivity and overall industrial technology; for example, early in the 1970s the Soviet computer industry was still turning out second-generation machines while the advanced Western countries were producing third-generation computers. Similar examples can be found in several other industries—including metallurgy, chemicals, and engineering—many of which rely on relatively unsophisticated and outdated equipment and are incapable of producing sufficient quantities of the more advanced products. While imports of advanced machinery account for a large share of total Soviet purchases from the West, sales of such products are insignificant in relative terms. As a result, the five-year plans for 1971–75 and 1976–80 placed heavy emphasis on measures to increase labour productivity and to narrow the technological gap between the Soviet Union and the United States.

The industrial sector (including mining) contributes nearly two-thirds of the country's net material product and employs more than a third of the economically active population. Growth in this sector has been extremely rapid; the index of industrial production rose by 114 percent in the decade 1965–75—an increase that was considerably larger than the average for developed market economies and for the world as a whole. In fact, among the developed countries only Japan recorded a better performance during this period.

Capital and consumer goods output. As in most other centrally planned economies, the main emphasis has been on the development of the capital goods industries. In successive five-year plans since World War II, this sector received 30–35 percent of the country's total expenditure on investment, whereas only about 5 percent was directed into consumer goods industries. As a result, between 1950 and 1970 (according to official Soviet sources, which have been criticized by Western economists as exaggerating industrial growth rates) the output of capital goods was multiplied by about 7.7, whereas consumer goods production was multiplied by only about 5.4, a ratio of about 1.4:1. These figures, however, represent a considerable acceleration in the relative as well as the absolute growth rate for consumer goods production; in prewar five-year plan periods (1928–40) the ratio between the two sectors was about 4:1. Investment in the consumer goods industries has been given progressively greater emphasis in plans since World War II, and by the early 1970s the two sectors were growing at approximately equal rates.

In spite of this change of priorities, the funds devoted to the consumer goods sector remained rather small in relation to its role in the economy; it received only about one-sixth of total industrial investment, although it accounted for more than a quarter of gross industrial production and employed about a quarter of all workers occupied in industry.

As a result of the past emphasis on heavy industry, production facilities in consumer goods are generally outdated and inadequate, and labour productivity is considerably lower than in most other branches of industry. In consequence, output is not large enough to meet domestic demand, and there are also frequent complaints about quality and variety. It was hoped, however, that the 1971–75 and 1976–80 five-year plans—which devoted considerable attention to this sector—would see a significant improvement in the situation. The main objective was to double capital investment during the 1970s in comparison with the previous decade and to achieve a production increase of 95–100 percent. In the 1971–75 plan, the output of consumer goods was scheduled to grow faster than that of heavy industry (44–48 as against 41–45 percent), but in the 1976–80 plan the order of priority was reversed, to 38–42 percent growth in heavy industry and 30–32 percent in consumer goods. This indicated the failure of the plan for 1971–75 to raise the standard of living as it was meant to do. The plans also stressed the need to adopt better management techniques and to speed up the absorption of new technology.

The main branches. Within the capital goods sector, which makes up approximately three-quarters of total industrial production, the principal emphasis has been on the development of machinery building (defined as producing the means of production), metalworking (*i.e.*, using a semi-finished metal product such as plate or tube to make a component or finished product), metallurgy, and the chemical industry. Collectively these branches form the backbone of Soviet industry, accounting for about half the total industrial production.

Table 5: Soviet Output of Selected Industrial Commodities

	1960	1977	% increase
Crude steel (000,000 metric tons)	75.3	147.0	95
Rolled steel (000,000 metric tons)	43.7	118.0	170
Sulfuric acid (000,000 metric tons)	5.3	21.1	298
Caustic soda (000,000 metric tons)	0.7	2.7	286
Mineral fertilizers (000,000 metric tons)	13.9	96.7	596
Plastics (000,000 metric tons)	0.3	3.3	1,000
Machine tools (000)	155.9	236.0	51
Trucks (000)	362.0	734.0	103
Automobiles (000)	138.8	1,280.0	822
Tractors (000)	238.5	569.0	139
Grain combines (000)	59.0	106.0	80
Paper (000,000 metric tons)	2.3	5.5	139
Cement (000,000 metric tons)	45.5	127.0	179
Cotton fabrics (000,000 metres)	6,387.0	6,800.0	6
Woollen fabrics (000,000 metres)	342.0	800.0	134
Linen fabrics (000,000 metres)	559.0	805.0	44
Silk and synthetic fabrics (000,000 metres)	810.0	1,600.0	98
Leather shoes (000,000 pairs)	419.0	735.0	75
Clocks and watches (000,000)	26.0	60.7	133
Radio receivers (000,000)	4.2	8.7	107
Television sets (000,000)	1.7	7.1	318
Domestic refrigerators (000,000)	0.5	5.8	1,060
Domestic washing machines (000,000)	0.9	3.6	300
Dairy products (000,000 metric tons)	8.3	24.3	193

Between 1965 and 1975, the metalworking and machinery-building sector trebled its output and more than doubled its labour productivity—a better performance than could be seen in any other branch of manufacturing.

Chemicals

In the pre- and early post-World War II periods, the chemical industry was one of the less well-developed branches of Soviet manufacturing industry. From the late 1950s onward, however, helped by the rapid growth in the production of oil and natural gas, particular attention was paid to this sector of manufacturing, and in the early 1960s it was the most rapidly expanding branch. During the late 1960s and early '70s, there was a noticeable slowdown in the growth of the chemical industry, attributable partly to a slight decline in the rate of overall industrial growth but mainly to a change in planning priorities. During the phase of most rapid expansion, the main emphasis was on ensuring an adequate supply of such basic chemicals as soda ash, sulfuric acid, caustic soda, and chemical fertilizers and on the establishment of large-scale artificial fibres and plastics industries. This growth was aided by heavy imports of Western equipment and technology, including a large number of complete plants. Now that a chemicals industry more commensurate in scale with that of the Soviet industrial economy as a whole has been established, it is not surprising that annual growth rates should have slowed down. Nevertheless, expansion continues—between 1965 and 1975 the output of all chemical products rose by 178 percent, a rate of increase about half that of the decade 1955–65.

Metallurgy

Since the beginning of the Soviet era, a strong metallurgical industry has been looked upon as one of the main preconditions of industrial growth and has, on the whole, been given priority in the allocation of investment funds. Its share of total capital spending, although subject to some recent limitations, has been significantly higher

than its share of industrial output. Although the increase in both productivity and production has fallen well below the average for industry as a whole, output has managed to keep up with the growth of domestic demand, and steel, manganese, and a number of other metals are exported in substantial quantities.

Other important manufacturing branches include the building materials, timber, woodworking, and paper industries; at the start of the 1970s these absorbed about 14 percent of total industrial investment. Over the 1960s they saw a rapid growth in output, but, in common with most other capital goods industries, there was a marked slowdown in the period 1966–75.

The bulk of the country's industrial output comes from relatively large production units, and the average size of industrial plants is steadily increasing. In the mid-1970s about 35 percent of all enterprises employed 100 or fewer people but accounted for only 6 percent of gross production (compared with 44 percent and 7 percent, respectively, in 1960). Enterprises with a labour force between 101 and 1,000 accounted for 53 percent of the total number and for some 26 percent of output (49 percent and 40 percent in 1960). At the top end of the scale, enterprises employing more than 1,000 workers, while making up only 12 percent of the total number, provided no less than 68 percent of gross industrial production (7 percent and 53 percent in 1960).

Energy. Throughout the Soviet period, the energy sector has been given high priority by economic planners; partly in response to Lenin's dictum that "Communism equals Soviet power, plus the electrification of the whole country," particular attention has been paid to the development of the electric power industry, which in the early 1970s took 11–12 percent of all capital investment in the industrial sector. Between 1965 and 1975, total installed capacity increased by 90 percent and output doubled. The five-year plan for 1976–80 envisaged a 30–33 percent further rise. Although hydroelectricity production has also seen a rapid increase, its share of the total has declined from a peak of nearly 20 percent in the late 1950s. This change marks a realization of the greater returns obtainable from coal-, oil-, and gas-fired thermal stations in consuming areas and some lessening of the emphasis on grandiose hydroelectric projects in remote districts (though these continue to be built). The Soviet Union's first nuclear power plant came into operation in the mid-1960s, and several others have been built, but their combined output remains a minor proportion of the total.

Oil and natural gas fuels

A major development in the energy sector since the mid-1950s has been a rapid growth in the production of oil and natural gas (see above *Mineral resources*), which are now the leading fuels. Whereas in 1955 oil and gas accounted for only about a quarter and coal for some two-thirds of the country's energy supplies, by the mid-1970s the contribution of coal was down to about 33 percent and that of oil and gas had reached 65 percent (the small balance came from such minor sources as peat, oil shale, and firewood).

The development of Soviet energy resources, particularly oil, gas, and hydroelectricity, has involved problems resulting from their highly uneven distribution and the great distances that separate resources from consuming areas. These difficulties are being attacked by a large program of oil and gas pipeline construction (see below *Transportation*) and the creation of an integrated electric power grid and by bringing the geographical pattern of demand nearer to that of supply.

Finance. As in other Socialist countries, the banking sector is owned and controlled by the state. This control

Gosbank

is exercised through the powerful Gosbank (Gosudarstvenny Bank S.S.S.R., or State Bank of the U.S.S.R.), which not only fulfills most of the functions of a Western central bank but also acts as a clearing bank and is used as one of the principal instruments of control over state enterprises. It is responsible for note issue, for controlling the currency circulation, for foreign trade and other overseas transactions, and for executing the government's financial and credit policy. It administers tax receipts, pays out much of the state's current expenditure, monitors the performance of state enterprises, holds their accounts, handles their receipts and payments, and provides them with short-term credit.

A number of specialist transactions, however, are handled by other banking institutions, notably the Bank for Foreign Trade (called Vneshtorgbank, from Bank Dlya Vneshney Torgovli) and the Stroybank (Vesesoyuzny Bank Finansirovaniya Kapitalnykh Vlozheny, or All-Union Bank for the Financing of Capital Investments; the common name is derived from *stroytelstva,* "construction," which the bank finances), which is the source of much long-term credit to state enterprises and organizations. The most widely spread banking institution is the State Savings Bank (Sberegatelnye Kassy). It has about 79,500 branches distributed all over the country and provides a wide range of banking services to the general public, collectives, and other organizations. Savings deposits have been growing at a rate of just under 20 percent annually. The Soviet Union also maintains a number of banks in the West.

Foreign trade. Although the Soviet Union's foreign trade nearly tripled between 1965 and 1975, the country is still a very small foreign trader in relation to its size and economic power. In value terms its exports and imports are slightly larger than the combined foreign trade of Belgium and Luxembourg but smaller than that of Italy, the population of which is only some 22 percent of the Soviet total. The figure is also small when compared with that of other Communist countries. The country's rather poor showing in world trade is attributable in part to its wealth of mineral resources, which obviate the need for large purchases of raw materials, and in part to a marked lack of interest in specialization.

In effect, Soviet policy appears to be to attain virtual self-sufficiency in as many fields as possible and to use foreign trade as a means of making good some gaps that cannot otherwise be filled by domestic producers and to mitigate occasional shortages. As is the case in most other countries of the Communist bloc, great care is taken to maintain a balance between exports and imports or to earn a surplus.

Trading partners

Although the bulk of the Soviet Union's foreign trade since World War II has been with fellow Communist countries, their share of the total declined from nearly 75 percent to about 56 percent during 1960–75; the Comecon group accounted for 44 percent. This change was partly accounted for by reduction of trade with the People's Republic of China, but the chief cause of it was the Soviet Union's expanding trade relations with the West. By the mid-1970s trade with the non-Communist world had reached the same level as that with the Comecon group. The largest component of this was trade with developed capitalist countries, which had grown rapidly over the preceding decade. Though the Soviet Union's largest single trading partner was still the German Democratic Republic, followed by Poland, Bulgaria, and Czechoslovakia, leading capitalist partners were the Federal Republic of Germany, Japan, Finland, the U.S., Italy, France, and the United Kingdom. Trade with developing countries has also been increasing at a rapid rate, though by the mid-1970s it still represented only 12.5 percent of the total. The most important partners in this group were India, Iraq, Iran, Egypt, and Brazil.

In contrast to the healthy surplus maintained in most years in the Soviet Union's trade with Communist and developing countries, trade with several of the developed capitalist countries (*e.g.,* West Germany and the United States) has often yielded a deficit. This is covered by exports of gold, of which the Soviet Union is one of the world's leading producers.

Imports

In view of the country's abundant mineral resources, it is not surprising, as noted above, that raw material imports make up a relatively small part of its purchases abroad. About a third of these purchases are accounted for by a wide and growing range of machinery and equipment. The second largest import group is consumer products, followed by foodstuffs and ores, concentrates,

and metal products. The composition of Soviet imports from the developed Western world is more than half accounted for by machinery and transport equipment items, while a further quarter is taken up by shipments of manufactures, especially iron and steel and textile yarns.

Principal exports

Nearly a quarter of the Soviet Union's export earnings is derived from sales of machinery and allied products, mainly to fellow Communist and some developing countries. The advanced Western world takes only a minute portion of the total. Approximately 8 percent of Soviet machinery sales is made up of complete plants. Vehicles and transport equipment, communications equipment, lifting machinery, agricultural equipment, and a wide range of electrical machinery are also important. In second place come sales of metallic ores and concentrates, and fuel and electricity sales—mostly of oil and petroleum products—rank third.

The structure of Soviet sales to the West, however, differs sharply from the overall pattern. A large proportion of these is made up of inedible crude materials (hides and skins, textile fibres, metallic ores) and oil. Manufactured products account for about a quarter.

Overall, the Soviet Union is a net importer of machinery, consumer goods, and chemical products and a net exporter of fuels, ores and metals, and timber products. Exports and imports of foodstuffs are usually small and in most years cancel each other out. On a number of occasions in the 1960s and early and mid-'70s, however, large imports of grain were necessary to offset poor harvests at home.

MANAGEMENT OF THE ECONOMY

The role of government. In this, the prototype Communist economy, the government's role is all-important. Industry and most service trades are owned and controlled by the state, and agriculture is in the hands of the collective and state farms. In effect, there is no private sector; in the countryside, private enterprise is tolerated only in respect of farmers' small individual plots, while in other sectors of the economy only artists, a few artisans, etc., can earn a living without being formally employed by the state.

This means that the government is in full control of the economy. It formulates economic policy, prepares long- and short-term plans, lays down a large number of detailed targets, controls investment, determines wages and incomes, and fixes the prices of most commodities and products. It is in control of foreign and internal trade, banking, and almost every other branch and sector of the economy.

The structure of state planning

The mechanism by which such control is operated underwent extensive reform—accompanied by considerable discussion—in the late 1960s. Under the system that had evolved by the early 1970s the top planning position is occupied by Gosplan (Gosudarstvenny Planovy Komitet, or State Planning Committee), the authority of which extends right down the line through a number of regional Gosplans. Detailed control over industrial activities is exercised by all-union ministries and ministries of the union republics, which look after industries of national significance. There are also some republican ministries that control local industries and certain construction activities. In addition, there are a number of state committees, the most important of which are the State Committee for Science and Technology and the State Committee for Material and Technical Supplies. The latter works through regional supply authorities and shares the task of material and resource allocation with Gosplan. There is also a State Committee for Prices, but this comes under Gosplan rather than under the Council of Ministers. Agriculture is the responsibility of the Ministry of Agriculture, while credit, finance, banking, and foreign exchange are under the control of the Ministry of Finance and the Gosbank.

At the top of the planning mechanism is the Council of Ministers, determining the broad lines of economic policy and the principal plan priorities. These guidelines are passed down to Gosplan, which is charged with incorporating them into a fairly detailed draft plan, setting out the principal tasks for the various sectors and branches of the economy. The plan is then passed on to the appropriate ministries, the local Gosplans, and other regional authorities, where—after extensive consultations with individual enterprises—more detail is added. The draft plan, suitably embellished, then goes back to Gosplan, which—after the necessary checking, coordination, and balancing—works out a detailed program. This is then submitted to the Council of Ministers and thence, after amendments, to the Supreme Soviet. After any further amendments, it becomes the all-important official current plan.

Taxation. The two most important taxes are the turnover tax and the tax or levy handed over to the government by state enterprises. The latter is made up of the difference between the enterprises' profits and the amount they are allowed to keep, as well as a charge on their capital assets. Receipts from enterprises became increasingly important during the 1960s, when they rose from less than a quarter to more than a third of total budgetary revenue. The percentage of net earnings deducted varies from firm to firm, but in the late 1960s probably averaged from 76 to 80 percent. Subsequent moves toward a greater degree of self-financing (as opposed to reliance on budgetary grants) have diminished the importance of receipts from enterprises to less than a third of revenue. Turnover tax is collected at the factory or the wholesale distribution centre and is levied on a wide range of consumer goods. The rates vary considerably, but since the late 1960s the relative yield of this tax has seen a decline to a third or less of total revenue.

Turnover tax

Among other taxes, the most important group covers direct levies on the general public. Most of the amount raised in this way is derived from income tax, which is payable on incomes in excess of roughly two-thirds of the average wage. It is slightly progressive; on wages it rises from 5 percent to 13 percent, but the rate is doubled on incomes from self-employment, and in a few cases it is fixed even higher. Another direct levy is the tax imposed on collective farmers in respect of the area of their private plots. The rate of this tax is fairly low but is increased by a significant margin if the farmer fails to reach a stated work-day quota in the collective. This regulation is designed to ensure that farmers do not neglect their work in the collective in favour of their private activities.

Contemporary policies and the economic outlook. Since the mid-1960s, the Soviet Union has been engaged in the gradual implementation of a far-reaching program of economic reform. Seen from the Western standpoint, the aim of the program—the need for which was widely recognized for several years—appears to be to mitigate the rigidity, the inconsistency, the complexity, and the inefficiency of detailed central planning by placing greater reliance on incentives, individual initiative, and a rather restricted use of some of the economic instruments that are operating in non-centrally planned economies. Soviet authorities view the flexible methods of the reforms as stemming from the growth and complexity of the economy—a situation that has rendered the historically necessary highly centralized system obsolete.

Economic reform

The most important change concerns the control, the responsibilities, and the independence of the state enterprises. Before the introduction of the reforms, these were given a large number of centrally determined targets, including physical output figures, which left practically nothing to the management's discretion and provided no incentives for producing goods of the right type and quality and for using resources of labour and capital efficiently. Under the reform program, however, the number of targets or indicators was reduced from 40 to eight, the most important of which are concerned with sales and profits. Although these changes gave enterprises only a limited freedom of action, they nevertheless had the effect of encouraging the enterprises to pay greater attention to improving quality, increasing efficiency, and thereby raising profits.

Coupled with these changes were some reforms that related to the system of employee remuneration and the

financial responsibilities and independence of state enterprises. The principal development in this field was the creation of three new incentive funds: for material incentives, cultural and social welfare and housing, and the development of production. The size of the material incentive and cultural and social funds determines the amounts available for bonus payments and social expenditure and is dependent on the fulfillment of sales and profit indicators. The production development fund is a function of profits, depreciation allowances, and proceeds from the sale of existing capital assets and is designed to allow a much greater degree of self-financing than previously for those enterprises that reach their profit targets and are efficiently managed. The aim of these changes, combined with the strong link between profits and the size of the incentive funds, is to encourage a more efficient use of capital resources.

Industrial prices revision

An equally important part of the reform program was a general revision of industrial prices, which, of course, are not determined by market forces as in a capitalist economy. In 1966 and 1967 new prices were fixed on the basis of production costs, although, in the case of coal, electricity, and oil, allowance was made for the wide regional cost differences. Some basic raw materials prices saw a very large jump—coal went up by 78 percent, gas by 51 percent, electricity by 22 percent, zinc by 64 percent, and heavy chemicals by 61 percent. Since the 1967 price revision, much thought has been given to the problem of increasing the flexibility of the price-fixing mechanism.

Although not part of the economic reforms announced in 1965, the slow trend toward the creation of associations of enterprises also deserves mention. The aim is to create larger units in order to take advantage of the economies of scale, to reduce the duplication of certain facilities, to encourage centralized supply of equipment and materials and of such services as research, which smaller units would not be able to finance without substantial assistance. Although there have been some ill-conceived amalgamations, there is evidence that, on the whole, the program has had a positive effect, and current policy is to encourage further mergers, especially among smaller enterprises.

Further reform

The evidence available to Western specialists at the time of the introduction of the 1976–80 plan indicated that the economic reforms have, on the whole, had a positive effect on industrial efficiency. Although there is much discussion in Soviet circles, it is generally recognized that reforms need to be deepened and developed, and it is thought likely that further changes in the system of economic management will evolve. Among the problems that are receiving considerable attention are some reforms to the procedure for fixing prices; the form and content of planning; and a further increase in the independence of the state enterprises. It is often admitted that the current system is inflexible, which makes it difficult to ensure that costs and market conditions are fully taken into account and that scarce resources are allocated in the most efficient way. There are plans for more frequent price revisions and for the development of techniques that will allow economic factors to play a greater role than hitherto. Some tentative steps have already been taken in this direction, but there is still a long way to go. In the field of planning, the emphasis is on reducing the number of binding output targets. Much progress has already been made in this field—the number of mandatory output targets was lower in the 1971–75 plan than ever before and was reduced still further in the plan for 1976–80—but further changes are likely. The discussion relating to the degree of independence of the state enterprises concerns their right to choose their suppliers and to take a more active part in procurement of raw materials. Although some steps have been taken along these lines, contact between most producing and consuming enterprises is still fairly weak.

In sum, it appears that the enormous development of the Soviet economy since World War II has generated a whole new set of problems and challenges, which—not least in the realm of technology—will occupy Soviet economists, planners, and politicians for several decades ahead. (E.I.U./J.C.D.)

IV. Transportation

THE SOVIET SYSTEM IN ITS NATIONAL CONTEXT

In a country the size of the Soviet Union—where great distances separate industrial resources from manufacturing centres and manufacturing centres from markets—transport costs, both capital and recurrent, assume a much greater scale and significance than they do in smaller geographic entities. This is particularly so where heavy industry has received priority and requires the movement of enormous quantities of bulky materials over long distances; in addition, long-distance movement of foodstuffs, particularly grain, takes place on a very large scale. It is not surprising, therefore, that the development of the Soviet economy has involved a vast increase in the volume of freight movement and, to a lesser extent, of passenger travel. The various transport media have been developed as a single coordinated system, avoiding the competition between various modes of transport that occurs in the capitalist world.

The transportation system employs about one in 10 Soviet workers and receives about 10 percent of all capital investment.

The expansion of the economy is reflected in the growth of transport operations, with railways accounting for about 60 percent of freight turnover in the mid-1970s. The growth rates of freight handling by other forms of transport—pipeline, sea, and road—are even higher, and the railways' share is slowly diminishing. The 1976–80 five-year plan provided for a 30 percent increase in total freight turnover and continued to emphasize sea, pipeline, and road transport.

For historical, economic, and geographical reasons, the transport network is very unevenly distributed. The densest road and rail networks are in the Baltic republics, the Ukraine, and Moldavia; intermediate densities occur in the Central Industrial region and Belorussia; and very low densities are characteristic of the Asiatic regions.

THE COMPONENT SYSTEMS

Rail transport. Though it is declining, first place in volume of operations is still taken by rail transport; the explanation for this is the country's immense overland expanses and the spread of the principal production zone over several thousand miles from west to east. Only rail transport can effect long-distance mass carriage of freight and passengers all the year and this despite the severe climatic conditions of the eastern and northern regions.

Railway freight movement

The railways account for around 60 percent of the country's internal freight turnover, which amounts to more than half the freight turnover of all the world's railways. Rail transport handles most of such regular mass deliveries as coal, ore, timber, metal, and grain.

The total length of the rail network has roughly doubled during the Soviet period. Railway construction in the European sector has been mainly confined to rebuilding old lines and building shortcuts to reduce the length of many of the more circuitous routes between the cities of the European plain; the only completely new route is the Vorkuta–Kotlas line, constructed in the 1940s to provide an outlet for the Pechora coalfield. Most of the new routes built since the Revolution are to be found in the economically expanding middle zone of the country, embracing the Volga, Urals, West Siberian Plain, Kazakhstan, and Central Asian regions. In this zone, the Novosibirsk–Novokuznetsk main line linked the main Trans-Siberian line with the Kuzbass; the Artyshta–Barnaul–Tselinograd–Magnitogorsk line linked the Kuzbass with the Urals via north Kazakhstan; the Petropavlovsk–Karaganda–Chu line linked the Trans-Siberian with the Central Asian system, while the latter was drawn together by the completion of the Turkistan–Siberia (Turksib) route from Krasnovodsk on the Caspian via Tashkent, Alma-Ata, and Semipalatinsk to Barnaul. The busiest and most important main lines have been reconstructed, notably the Moscow–Donbass, Moscow–Sverdlovsk, Yaroslavl–Arkhangelsk, and Che-

lyabinsk–Vladivostok sections. Until the early 1970s there was little new building east of Lake Baikal. The largest single project under construction in the middle of the decade, however, was the Baikal–Amur Mainline, linking existing railheads at Bratsk and Komsomolsk-na-Amure. Running about 250 miles (400 kilometres) north of the Trans-Siberian line for a distance of some 2,000 miles (3,200 kilometres), this line, due for completion by 1985, is designed to tap a variety of new industrial resources in Eastern Siberia and the Far East. It will also lessen reliance on the Trans-Siberian line, which, over much of its length, runs very close to the border with China.

Railway modernization

Traffic on many sections of the Soviet railway system is the heaviest in the world, and the growth of traffic during the 1960s necessitated the replacement of steam by diesel and electric traction, along with other technological improvements. Several major routes have been electrified, notably those from Moscow to Leningrad, the Donbass, northern Ukraine, Caucasus, and Urals, together with the Trans-Siberian as far as Irkutsk. Virtually all other lines are now operated by diesel locomotives.

Sea transport. Sea transport plays an increasingly important role, accounting for about 14 percent of total freight traffic (but only about 3 percent of passenger movement), and has a high growth rate. In addition to its dominance in Soviet foreign trade, sea traffic is of considerable importance in freight movement between the various regions, particularly where, as in northern Siberia and much of the Far East, no alternative form of communication is available. The great length (30,000 miles) of the Soviet coastline favours this, though the winter freeze in northern regions limits its growth potential.

First place in internal traffic volume is taken by the Caspian and Black Sea–Sea of Azov basins, followed by the Far Eastern, Baltic, and Arctic basins. The northern sea route, through the seas of the Arctic and Pacific oceans, has, since 1935, played an important role in ensuring regular communications between regions of the European North and Siberia and the Far East.

Foreign traffic nevertheless predominates; the well-equipped Soviet merchant fleet holds sixth place in the world and carries about 7.5 percent of world sea freight traffic. The main emphasis in technological development in the Soviet fleet is toward containerization of freight and expansion of regular freight and passenger line services.

Under the 1971–75 plan, seagoing freight more than doubled in volume, and the merchant fleet is being augmented by ships with automatic control systems.

Transport by inland waterways. Inland waterways are an important component part of the unified system, accounting for 4.3 percent of total freight traffic. The majority of waterways (of which about a fifth are artificial) lie off the main lines of communication, a factor that helps to explain the relatively small percentage of river freight traffic. But wherever the river networks coincide with the main transport routes, or other forms of transport are insufficiently developed, the proportion of inland waterways traffic increases. The Volga area, the northwest of Soviet Europe, the Far East, and Eastern Siberia are such regions.

Some 70 percent of the freight carried is oil, timber, and bulk materials such as stone, gravel, ores, coal, and grain. More than half the freight is carried on the rivers of the Volga–Kama Basin. Other important areas are the Northern Dvina, Dnepr, and Don in the European part of the country; the Ob and Irtysh, Yenisey and Angara, Lena, and Amur and its tributaries in Siberia and the Far East, and the Amu Darya in Central Asia.

The most important navigation canals are the Moskovsky Kanal (a short but vital link between the Moskva and Volga rivers), the Volgo–Donskoy Kanal, the White Sea–Baltic Waterway, and the Volga–Baltic Waterway (a system of canals linking the Baltic Basin and the Volga).

Pipeline transport. One of the most economic forms of transport, pipelines, became of major importance only after World War II. By the mid-1970s pipelines accounted for about 13 percent of total freight turnover, and pipeline traffic more than tripled between 1965 and 1975.

Oil pipelines

The oil pipeline network is most highly developed in the European Soviet Union and the Caucasus. Oil is carried from the Volga–Urals field to the major cities of the European plain, to exporting ports on the Baltic and Black seas, and, via the Druzhba (Friendship) pipeline, to the Communist countries of eastern Europe as well as to West Germany, France, and Italy. A major oil pipeline from the Volga–Urals region (Tuymazy–Omsk–Angarsk) traverses Siberia; in the 1970s lines from the Tyumen oil field of Western Siberia were connected to the trans-Siberian pipeline, which then began pumping Siberian oil westward as well as eastward. Oil pipelines also connect the Mangyshlak field of eastern Kazakhstan with the European system and carry oil from Sakhalin to Komsomolsk-na-Amure.

The first big gas pipeline, running 524 miles (843 kilometres) from Saratov on the Volga to Moscow, began to function in 1947. Since then an extensive system has been laid in the western part of the Soviet Union carrying gas from the Ukraine, Caucasus, and Central Asia to the Central Industrial region, the Urals, and other industrial zones.

Road transport. Road transport, second only to that of the railways, accounts for about 6.5 percent of the total freight traffic and 40.6 percent of the total passenger traffic by public transport. Most of this movement is over relatively short distances.

The densest network of improved roads is in the Baltic region, the western Ukraine, the Crimea, Transcaucasia, and the region around Moscow. The oil field areas of Tatariya and Bashkiriya are also well provided with hard-surfaced roads. On the whole, the European part of the country is better provided with roads than the east. In Siberia and the Far East, roads are the main means of communication in a number of areas remote from railways.

Bus services

About three-quarters of road freight is carried in the European part of the Soviet Union, and four-fifths of the urban and suburban bus traffic is found there, with around 2,000 communities served by bus routes. In the towns, buses account for about 40 percent of all passenger transport. The greatest use of buses is in the suburban zones of Moscow and Leningrad. The freight turnover of public road transport more than doubled between 1965 and 1975.

Air transport. Air transport is most widely used for passenger movement. By the mid-1970s, internal airline routes reached more than 3,500 towns and inhabited localities. Air travel is very popular as the fastest form of transport over the country's immense distances, and more than 40 percent of all passenger flights are over distances of more than 300 miles. The high proportion (about 60 percent) of flights of less than 300 miles is to some extent due to the growth of air communication in regions in which it is the sole form of transport.

Aviation is used on a lesser scale for freight transport. Sixty percent of the freight traffic of local airlines is in the trackless northern and eastern regions.

Soviet international air routes link the nation with some 60 states in Europe, Asia, Africa, and America; Soviet air transport, by volume of traffic, accounts for more than 25 percent of the world total.

Urban electric passenger transport. Electric systems for the conveyance of passengers in urban areas are of major importance. By the middle of the 1970s there were tram services in well over 100 towns and trolleybus services in about 100. Trams and trolleybuses together carried 16,200,000,000 passengers a year. Five great cities—Moscow, Leningrad, Kiev, Tbilisi, and Baku—have underground railway systems. (V.N.D./J.C.D.)

V. Administration and social conditions

Differing national and international perspectives on the evolution of the Soviet political system may be found in the articles RUSSIA AND THE SOVIET UNION, HISTORY OF; COMMUNISM; and SOCIALISM. The following description

of the Soviet system can be supplemented by appropriate sections of the articles on the Soviet constituent republics.

STATE AND SOCIETY

The constitutional framework. In view of the domination exercised over political, economic, social, and cultural life in the Soviet Union by the ruling, monopolistic Communist Party, constitutions and other legal documents are subordinate to the policies of the party and its leadership, centred in the Politburo. The party, however, desires to project an aura of constitutional legitimacy. Hence, it is important that finally, after many years of a somewhat shadowy existence, the Constitutional Commission, appointed in 1962, produced in May 1977 a draft of a new constitution, which, after several months of well-organized discussion in the controlled press, was enacted unanimously into law by the Supreme Soviet in October 1977.

Comparison of the Stalin Constitution—as the "fundamental law" enacted in 1936 was usually called by the Soviet communications media—and the constitution of 1977 reveals no essential changes. The new document is far longer than the old one and in some ways more candid. It much more openly proclaims the ruling and all-penetrating role of the Communist Party of the Soviet Union (CPSU; Kommunisticheskaya Partiya Sovetskogo Soyuza), for example, describing it (Article 6) as "the nucleus of [the] political system." It describes more fully the various legal rights of citizens than did its predecessor. In emphasizing that all rights are accorded by the authorities in return for exemplary performance of duties, however, it in effect limits the rights of citizens more severely than did even the Stalin Constitution.

The fundamental point, however, is that, to a far greater extent than is the case in Western democracies, the Soviet constitution of 1977, like that of 1936, will have meaning only to the extent that the leadership desires and allows it.

The economic foundations. The economic life of the Soviet Union and all other aspects of social activity are dominated by the CPSU. Its Politburo determines economic policy and sees to it that its decisions are enforced by the Soviet government. The CPSU penetrates and guides the government by a system of interlocking membership of party and governmental executive bodies and by many other means. The Soviet economy is held to be "socialist" because its basic principle is state ownership, established by Lenin in 1918 insofar as the main elements (industry, transport, finance, etc.) were concerned and extended by Stalin to agriculture, which Lenin had for pragmatic reasons left largely in the hands of smallholder peasants. Despite state domination of the Soviet economy, not all property is directly administered by state agencies. The state owns and administers land, minerals, water resources, forests, industrial enterprises, banks, large state agricultural enterprises (known as *sovkhozy,* or state farms), means of transport and communication, municipal enterprises, and basic housing funds. In addition to state property, there are two other kinds of property in the Soviet Union, namely cooperative property, predominant in agriculture, and personal property. Most farmers are members of cooperative, or collective, farms, known as *kolkhozy*. Nominally, the land farmed by the collective farmers is assigned to them in perpetuity, but, as is indicated by government-ordered consolidations of collective farms into larger units since the late 1950s, the collective farms in fact exercise whatever control they have over their land within a framework of policies determined by the political authorities. See further above *The national economy: Management of the economy*.

Ownership of property

As for personal property, it includes articles of clothing, household objects, electronic goods, automobiles (in the case of citizens who can afford them, at prices several times higher than in the West, and only after waiting several years), and many other types of goods, including (again for upper-income persons) country cottages, or dachas. Soviet inheritance law with respect to personal property is fairly liberal. (See further SOVIET AND SOCIALIST LEGAL SYSTEMS.)

Social and political aspects. According to official doctrine, there are two "friendly" classes in Soviet society, the working class and the collective farm peasantry. There is also a rapidly growing "stratum," the intelligentsia, which, Western observers agree, is by far the most privileged of the three. This does not mean that all members of the intelligentsia are better off than the workers or even, perhaps, than the peasants. Highly skilled workers earn considerably more than the more poorly paid members of the intelligentsia, which roughly corresponds to the professional classes in Great Britain or America. However, the highest ranking writers, scientists, and other professionals are, next to highly placed party and state bureaucrats, the most privileged of Soviet citizens.

As implied by the official slogan, "From each according to his ability, to each according to his work," Soviet society is not egalitarian. However, there is a considerable degree of equality of opportunity. Generally, life is very competitive, and the unofficial guiding principle of distribution of rewards might be described as that of the career open to talents. This is modified by the requirement that to achieve success a person must be not only professionally competent but also politically reliable.

The CPSU, which as of late 1977 had about 16,200,000 members, dominates the political life of the country. The party rules through the formal structures of government, which include the pyramid of soviets ("councils") extending down from the Supreme Soviet of the U.S.S.R. to village soviets, and the system of ministries, headed by the Moscow-centred Council of Ministers and extending down to local representatives of ministries. Nominally, the deputies and presiding officers of the soviets at all levels are elected, but there is only one candidate for any office in those elections, and the selection of candidates is controlled by the CPSU. Party, soviet, and ministerial posts—as well as other positions of influence and power—are allocated in accordance with a "nomenclature" (*nomenklatura*) system, which is operated by the central authorities of the CPSU.

Soviets

The federal structure. The Soviet Union, in fact a highly centralized state, is nominally a federal republic. Its federal aspect derives from the fact that it consists of 15 theoretically equal republics, each of which has equal representation in the Soviet of Nationalities, the upper chamber of the bicameral Supreme Soviet, the national legislature. The republics, as well as lower ranking autonomous republics, autonomous regions, and autonomous (formerly national) districts, constitute a means by which symbolic political representation is granted to the numerous ethnic groups of which the Soviet Union is composed. All of the people of the Soviet Union, however, share common federal citizenship. Federal laws have equal force in the territories of the union republics. The federal legislature, under the guidance of the CPSU and the State Planning Committee (Gosplan), determines the federal budget, which includes the budgets of the union republics. (See also above *The national economy: Management of the economy.*) Thus the central government in Moscow dominates the economics of the union republics.

This dominance by Moscow is fortified by features of the Soviet system more fundamental than constitutional provisions. Perhaps the most important instrument of central control is the hierarchical, centralized structure of the CPSU. Party headquarters in Moscow, through its centralized control over personnel assignments, can appoint, promote, demote, or transfer even the heads of the party organizations of republics.

The tight control exercised by the Moscow leadership over local economic, political, and cultural life engenders keen resentment among many members of the non-Russian minority nationalities residing in the republics that are named for them. Resistance to Moscow's control has been manifested in a variety of ways; at its mildest it takes the form of anti-Russian jokes and other innocuous verbal behaviour (though even these have at

Resistance to Moscow

times been punished harshly) or—as after Soviet conquest of the western Ukraine as a consequence of the Nazi–Soviet pact of 1939 and Soviet victory over Germany in 1945—it may be manifested in sanguinary guerrilla warfare. Needless to say, the right of secession from the Soviet Union granted by the constitution (both that of 1936 and that of 1977) has never been exercised. Indeed, any suggestion of dissatisfaction with Moscow's rule on the part of members of non-Russian nationalities is usually regarded as grounds for the imposition of draconian penalties. Of all of the many forms of dissent or opposition to Soviet rule with which the Soviet authorities have had to contend over the years, nationality dissent has aroused the deepest anger and has been punished most severely.

Organs of state power and government. *The Supreme Soviet.* Formally, the Supreme Soviet is the highest organ of state power in the Soviet Union. In keeping with this status, affirmed in the constitution, only the Supreme Soviet can make laws for the entire country. In practice, however, the Supreme Soviet enacts only such legislation as has already been initiated by the CPSU; this is indicated by the fact that the CPSU Central Committee meets before the Supreme Soviet does and lays down the policies that are subsequently endorsed by the latter. It is also clear that the Council of Ministers, nominally appointed by the Supreme Soviet, is in fact a more powerful body than the latter. Despite its limitations, however, the Supreme Soviet is an important body. Its legislative commissions, each of which specializes in a given field of legislation, have been playing an increasingly larger role in bringing to bear on the legislative process the expert knowledge of lawyers, economists, and other kinds of specialists. Thus the competence of the Supreme Soviet to perform its legislative functions has increased. Certainly the Supreme Soviet is not the mere rubber stamp it was in the Stalin era, as is indicated by the increased regularity with which it has met since Stalin's death and by the increased length of its sessions, which, however, are still extremely short by Western standards.

Role of the CPSU

Nonetheless, the Supreme Soviet is far from being a parliament of the Western democratic type. Among the many striking differences, for example, is the fact that in the Supreme Soviet all legislation, including crucially important budgetary legislation, passes unanimously. Quite apart from this, and from the unfailing correspondence between CPSU directives and Supreme Soviet legislation, much law is made, despite the apparent lack of constitutionality, by governmental bodies other than the Supreme Soviet, such as the Council of Ministers. There are even many secret laws, not passed by and unknown to the Supreme Soviet.

In form, the Supreme Soviet is bicameral. Members of both chambers are elected (in one-party, one-candidate elections) for four-year terms. One chamber, called the Soviet of the Union, represents, with one deputy for every 300,000 inhabitants, constituents regardless of national origin. The other chamber, the Soviet of Nationalities, is so organized as to take at least symbolic account of the multinational composition of the Soviet population; it has 32 deputies from each of the 15 union republics and for the lower units, designated in descending order as autonomous republics, autonomous regions (*oblasti*), and autonomous districts (*okruga*), respectively 11, five, and one each. The Supreme Soviet normally assembles for regular sessions twice a year. It passes all legislation introduced by its leading members (who are also leading members of the CPSU) unanimously and expeditiously. It sometimes holds special sessions.

The two chambers

The Presidium of the Supreme Soviet. The Presidium of the Supreme Soviet has 37 members, elected by the Supreme Soviet itself and, like the parent body, serving for four years. Between sessions of the Supreme Soviet, which last for only about a week, the Presidium acts for it, performing such functions as convening and dissolving the Supreme Soviet, receiving ambassadors of foreign states, selecting and recalling Soviet diplomats, removing and appointing ministers, instituting and awarding honorary titles, etc. It may also order mobilization of the armed forces and may declare war.

It is obvious that the Presidium is a more powerful body than the Supreme Soviet, although probably somewhat less powerful than the Council of Ministers, which it appoints, and vastly less powerful than the Politburo and Central Committee of the CPSU. Of the 37 members of the Presidium, only its chairman is normally a major party leader and a significant national political figure. Fifteen vice chairmen, one from each of the union republics, serve under the chairman of the Presidium, symbolizing the nominally federal structure of the Soviet state. There are 20 additional members and a secretary.

Both the Supreme Soviet and its Presidium symbolize but do not really exercise power. They legitimize the power of the real ruling force in the Soviet polity, the CPSU. The Presidium also provides an instrument through which the CPSU can deal with foreign states in a manner more understandable and acceptable to the latter than direct transactions with the party would be. With regard to very important business, however—such as negotiations in the area of arms control—the general secretary of the CPSU has dealt with presidents of the United States, dramatically displaying the superordination of party to state in the Soviet Union.

The Council of Ministers. Nominally subordinate to the Supreme Soviet and, between Supreme Soviet sessions, to its Presidium, the Council of Ministers is in fact the most important governmental and administrative agency in the country; in fact, it is designated in the constitution as the "government" of the Soviet Union. According to the constitution, the decrees and ordinances issued by the Council of Ministers must conform to law as promulgated by the Supreme Soviet, but the Supreme Soviet has never overruled any decree of the Council of Ministers. This is not really surprising, since both bodies are dominated by the most powerful leaders of the CPSU.

Working under the direction of the council are ministries of two kinds, the extremely centralized all-union ministries (somewhat equivalent to departments of the U.S. federal government) and slightly less centralized union republican ministries, so-called because in addition to their Moscow headquarters they have local headquarters in the capitals of the union republics. In addition, there are under the council various state committees and special agencies, some of which wield enormous power; one of these is the Committee of State Security (KGB), which controls the Soviet political police and is probably, with the exception of the CPSU itself and perhaps the regular armed forces, the most powerful organization in the country. Both the state committees and the special agencies have ministerial status.

Ministries

Republican and local state organs. The political structure of republics and autonomous republics is almost identical to that of the country. To a somewhat lesser degree this is also true in the local administrative subdivisions, which are called, respectively, *kraya* (territories, very large in area), *oblasti* (variously translated into English as regions, provinces, etc.), *okruga,* and *rayony;* the last two subdivisions (both translated as districts) may be either rural or urban in character.

The essential element in this system of units is its hierarchical, pyramidal character; the whole constitutes a unified structure. There is no counterpart in the Soviet system to the independent unit of local government such as the American state, county, or city government, each with lawmaking and taxing powers.

The union republics have their own supreme soviets, but they are unicameral rather than bicameral like the Supreme Soviet. Like the federal Supreme Soviet, the supreme soviets of the republics (both union and autonomous) are elected for five-year terms; *oblast* and *rayon* soviets are elected for two and a half years. The soviets elect executive committees, in whose hands most of their power is concentrated. As at all levels of the system, CPSU groups in the local soviets see to it that these bodies carry out party policy.

It is possible that the highly developed "mass participa-

tion" in the working of this vast machine affords some satisfaction to some citizens. Official sources, such as speeches by party leaders, and a steady stream of articles on socialist democracy in the controlled press assert that the participation of citizens in local government constitutes a special advantage of the Soviet political system. It is impossible, because of official secrecy, either to confirm or refute such claims by direct investigation, but enough is known about the reality of political life in the Soviet Union to cast doubt on them. It is clear that participation is so hedged about with controls and limits that for most of its alleged beneficiaries it must be more a burden than a privilege.

The electoral system. Little needs to be said about Soviet one-party, one-candidate elections. They are so carefully managed that the official candidates—who are said to represent the bloc of party and nonparty voters—invariably receive virtually 100 percent of all possible votes. All citizens 18 years of age or older may vote, except for the certified insane and criminals confined by court sentence. In fact, voting, though not formally compulsory, is obviously regarded by the authorities as the duty of all citizens; enormous pressure and great fanfare are resorted to in getting people to vote.

THE JUDICIARY, ARMED FORCES, AND PUBLIC ORDER

The judicial system. The Soviet judicial system is highly centralized, hierarchical, and subject to supervision and control by the CPSU. Although far less arbitrary and coercive than it was under Stalin (whose police officials arrested, deported, and did to death millions of persons without anything remotely resembling due process of law), today's system is, by Western and especially by Anglo-American standards, exceedingly defective. There is no equivalent of habeas corpus, no trial by jury, no presumption of innocence, no real independence of the judiciary from ideological and political pressures. Indeed, judges, prosecutors, attorneys, and all others involved in dispensing justice are required to propagate CPSU policy in the judicial field. (See further SOVIET AND SOCIALIST LEGAL SYSTEMS.)

Political crimes

Moreover, although the Soviet authorities deny the existence of political crimes in the Soviet Union, they obviously do exist and are dealt with by very special, severe methods. The criminal code of the Russian Soviet Federated Socialist Republic, largest of the 15 union republics, contains provisions for prosecuting "especially dangerous state crimes" that are clearly political in nature; there are parallel provisions in the codes of the other republics. Loosely drafted, catchall statutes in effect make any publicly expressed criticism of official policy or ideology a crime. Thus, circulating or even having in one's possession written statements that "slander" the state or social system is punishable by long terms in concentration camps, euphemistically described as corrective labour camps. In several important categories of cases, judicial investigation is carried out not by the regular prosecutors but by officials of the security police (KGB). Outside the political sphere there is a system of courts, prosecutors, and attorneys that functions fairly well, subject to the limitations referred to above. In evaluating this system, however, it is necessary to bear in mind that it is subject to CPSU policies, which can change arbitrarily from time to time.

At the apex of the system is the Supreme Court of the U.S.S.R., whose judges are appointed by the Supreme Soviet. Below the Supreme Court (the only national court, apart from military tribunals) stand supreme courts of the union and autonomous republics, provincial, territorial, and city courts, courts of autonomous regions and national areas, and, at the lowest level, the people's courts, the last of which try the overwhelming bulk of all cases. The members of people's courts include a professional judge, who is elected, and two lay "assessors," who theoretically represent the people in the administration of justice at this level. Another so-called popular feature of the people's courts is the openness of their proceedings, though this is restricted in cases involving state security or intimate personal or family matters. Judges at the levels higher than the people's courts are legally trained professionals, appointed rather than elected. There is no institution of judicial review in the Soviet Union, though there is a well-developed system of appealing the verdicts of lower to higher courts.

An important role is played in Soviet justice by the institution of the procuracy. The procurator general of the U.S.S.R. is appointed for a seven-year, renewable term. He in turn appoints lower level procurators for five-year terms. The system of procurators has two main functions. The most unusual of these, in terms of U.S. practice, is supervision of the actions of government officials, with a view to preventing and if necessary prosecuting illegal actions taken by officials. The procurators also conduct criminal investigation and prosecute accused thieves, embezzlers, counterfeiters, and other violators of law. To be sure, prosecution of violators of certain Soviet laws (for example, that of "speculators" for engaging in private trade or other types of prohibited economic activity) is unfamiliar to Westerners.

The armed forces. In the Soviet Union military service is compulsory for all male citizens: two years for the army, three for the navy. There are a very large professional Soviet Army, Soviet Navy, Long Range Air Force, Air Defense Command (established 1955), and Strategic Rocket Force (1960). Most of the armed forces are under the direction of the Ministry of Defense, but the KGB also has several hundred thousand men under arms in such services as the Border Troops, which are equipped with patrol boats, helicopters, aircraft, trained dogs, etc. There is also a vast paramilitary organization known as DOSAAF (Vsesoyuznoye Dobrovolnoye Obshchestvo Sodeystviya Armii, Aviatsii i Flotu S.S.R., or All-Union Voluntary Society for Assistance to the Army, Air Force and Navy of the U.S.S.R.), which, among other functions, provides military training and patriotic indoctrination for schoolchildren. Of great importance also is the GlavPU SA i VMF (Glavnoye Politicheskoye Upravleniye Sovetskoy Armii i Voyenno-Morskogo Flota, or Main Political Directorate of the Soviet Army and Navy), an agency of the CPSU Central Committee, whose task is political supervision and indoctrination of the military forces.

DOSAAF

In the late 1970s it was estimated that the Soviet Union was spending 11–13 percent of its gross national product annually on its military forces, compared with 6 percent in the U.S., 5.1 in the U.K., and 3.7 in France. Its army strength was estimated at 1,825,000, its navy at 450,000 (including marines), and its air force at 1,025,000.

Public order. As in the administration of justice, two administrative systems maintain public order. The rough Soviet equivalent of Western police systems is the Ministry of Internal Affairs (Ministerstvo Vnutrennikh Del, or MVD), which controls the militia, or ordinary police, as well as prisons, labour camps, fire prevention services, road safety services, registration of births, marriages, and deaths, etc. The more political aspects of order maintenance are administered by the KGB (Komitet Gosudarstvennoy Bezopasnosti, or Committee of State Security), which performs espionage and counterespionage, guards high officials and state frontiers, and also exercises surveillance over foreign diplomats and press correspondents and in general safeguards the power of the regime.

MVD and KGB

There is an auxiliary police system known as the *druzhina,* whose more than 5,000,000 members, called *druzhinniki,* normally serve a few hours once a month. Wearing red arm bands, they patrol streets and other public places; they may apprehend violators of public order and turn them over to the militia.

THE SYSTEM OF PUBLIC ORGANIZATIONS

The Communist Party of the Soviet Union. The CPSU is the driving and integrating force and the main control mechanism of state and society in the Soviet Union. This fact is registered in article 6 of the constitution (1977), which states that the party is the "leading and guiding force of Soviet society . . . , of all state organizations and public organizations." Article 100 of the constitution also mentions the CPSU first in a list of organiza-

tions having the right to nominate candidates for election to the soviets. Of course these articles suggest only feebly the overwhelming initiatory, mobilizing, and penetrating capacities of the CPSU. Although official doctrine characterizes the party as a public organization, it is in fact the collective ruler of the country, containing within its ranks the men and women who make and enforce national policy. It, not the state through which it rules, dominates all Soviet life.

As of late 1977 its membership numbered more than 16,000,000. Perhaps 100,000 of those members enjoyed some measure of executive authority, but only a few hundred had much power. These were the people who headed the republic and main *oblast* party organizations and the party organizations of major economic and other enterprises and institutions, those who edited leading newspapers, and those who held high military or police positions or headed Central Committee departments and on the basis of such status were full members of the Central Committee or full or alternate members of the party's Politburo (Political Bureau).

Nomenclature system

The dominance of the party's central headquarters over the rank and file is effected by many means and methods. One of the most important is the nomenclature (*nomenklatura*) system of centralized personnel selection, according to which only party members who pass the screening tests imposed from above by appropriate levels of the party leadership can be promoted to a given post. There are lists of positions for which only Moscow headquarters can clear aspirants; lower levels of the party hierarchy screen candidates for positions at their levels, and so on down the line. Another important principle of party organization that enhances centralized control is known as democratic centralism, the most important components of which are strict party discipline and subordination of the minority to the majority (in practice, of lower to higher ranking organizations) and the obligatory nature of the decisions of higher organs for lower. In addition, Moscow headquarters dominates the Soviet mass media of communication. The CPSU Central Committee's newspaper, *Pravda,* and other Central Committee organs such as the magazine *Kommunist,* set the tone of the national press. Central control is also strengthened by the ladderlike system of political schools (in their role as hurdles for promotion somewhat similar to military academies in the United States armed forces), through which party leaders must pass in order to get to the top.

Komsomol and Pioneers. The All-Union Leninist Communist League of Youth (Vsesoyuzny Leninsky Kommunistichesky Soyuz Molodyozhi, or VLKSM), better known as the Komsomol, according to Soviet sources had nearly 35,000,000 members, between the ages of 14 and 28, in 1976. The Komsomol has grown very rapidly, a growth that reflects in large part the continuing urbanization and industrialization of the country, since Komsomol recruitment is far more successful in the cities than in rural areas. The Komsomol's organizational structure is modelled on that of the CPSU, and many former Komsomol leaders have risen to high posts as executives of the party. The Komsomol is regarded by the party as its reserve and helper, terms that convey some sense of the organization's role as an agency of indoctrination in the official ideology and as a channel for dissemination of CPSU propaganda on domestic and international issues. The Komsomol issues a number of publications, the best known of which is the mass circulation newspaper *Komsomolskaya Pravda* ("Komsomol Truth").

There is general agreement among Western experts on the Komsomol that it has by no means achieved its officially imposed objective of creating a "new Soviet man," imbued with Leninist zeal and self-sacrificing devotion to the ruling party's goals. This verdict is based, in part, on frequent Soviet press reports indicating political apathy, cynicism, and even widespread delinquency among Komsomol members. On the other hand, the Komsomol is a useful device for mobilizing Soviet youth for such drives as the Virgin and Idle Lands Campaign of 1954–56, to settle virgin lands especially in Siberia and Central Asia.

Pioneers

The Komsomol also performs other functions for the party, such as assistance in the work of the Pioneers (All-Union Lenin Pioneer Organization, or Vsesoyuznaya Pionerskaya Organizatsiya Imeni V.I. Lenina), an organization the membership of which embraces Soviet schoolchildren between the ages of nine and 14. Vastly larger in membership than the Boy Scouts and Girl Scouts in the U.S., and, unlike them, an organ of the state, the Pioneers assist the schools in ideological and patriotic indoctrination of children. Pioneers wear a three-cornered red scarf, which according to Soviet sources symbolizes the unity of the Communist, Komsomol, and Pioneer generations. Under the guidance of the CPSU and the Komsomol, the Pioneers seek to instill in children disciplined work habits, love of learning, and identification with Soviet ideology and the social and political system. Children's summer camps and educational and recreational centres aid in the pursuit of these goals.

The trade unions. Soviet trade unions are organized according to broad branches of the economy, covering employees of all ranks within a particular branch, such as medicine or engineering. Collective farmers are not covered. Membership is not compulsory but appears highly advisable on both economic and political grounds. By the later 1970s, total union membership exceeded 113,500,000, accounting for all but a small percentage of the total number of eligible employees. Among the largest unions are those for state farm and allied employees and educational institutions.

The basic units are the enterprise and factory unions, most of which are run by small committees. The activities of the enterprise and factory unions are coordinated by town, provincial, or republican union councils, which are elected at the conferences of the basic unit representatives. At the top is the All-Union Trade Union Congress, which is supposed to be convened every four years. The principal task of the congress is to elect the All-Union Central Council of Trade Unions (or VTsSPS; Vsesoyuzny Tsentralny Sovet Professionalnykh Soyuzov), the Presidium of which is responsible for broad policy and activities. Detailed administrative work is handled by the Secretariat. Individual unions, the structure of which follows this pattern, are closely controlled by the central union authorities.

Union function

Unlike the situation in Western countries with market economics, where the function of trade unions is to advance the economic and other interest of workers in relation to employers, labour unions in the Soviet Union are agents of the state. Strikes are forbidden, if not in law certainly in practice. Worker protests and even brief spontaneous labour stoppages occur now and then, but the outside world learns of these only by word of mouth reports that reach foreigners; they are not reported in the controlled Soviet press. From such reports as have reached the West, it appears that military force has been used to put down labour disturbances and that very severe measures were taken against some of their participants.

In 1977 a number of workers founded a trade union independent of the state. The authorities reacted with repressive measures, but this effort to break the official monopoly of organization of labour was thought by many to be a harbinger of highly significant future developments.

The main function of the trade unions is to stimulate workers to increase their productivity. Among the methods used to achieve this, perhaps the best known is "socialist emulation," in which factory is pitted against factory, industry against industry, in nationwide production contests; winning enterprises are awarded banners and receive much praise in the press. Another important function of the unions is the administration of the social insurance system. They also operate an extensive system of rest homes and sanatoriums and generally play an important role in providing services to workers who are in good standing. Like the other Soviet mass organizations, the unions operate an extensive publications net-

work, which of course functions within the framework of overall directives laid down by the CPSU Central Committee but focusses on matters of special interest to employees, including teachers and other professionals as well as industrial workers. The most important trade union publication is the newspaper *Trud* ("Labour").

SOCIAL CONDITIONS AND SERVICES

The basic principle of distribution of goods and services in the Soviet Union is stated in the constitution as "From each according to his ability, to each according to his work." The interpretation and application of this slogan is of course determined by the supreme leadership of the CPSU and is not subject to public discussion. There is abundant evidence, even in Soviet official sources—guarded though they are on such matters as the existence of pronounced inequality and privilege—that this constitutional principle is interpreted so as to permit very wide differences in income, perquisites, opportunities for domestic and foreign travel, access to quality education, and status and power. The existence of such differences is emphasized in statements by dissidents and by many former Soviet citizens now émigrés. In terms especially of power, but also in access to the best vacation and medical facilities, foreign travel, and many other values, high-ranking party and police executives, military officers, and government officials are at the upper end of the scale, while office workers, unskilled labourers, and collective farmers are at the bottom. Also exceptionally well off, in comparison with ordinary citizens, are writers and performing artists in good standing with the political authorities. Favoured novelists and playwrights receive enormous royalties. Scientists who rise to the coveted status of membership in the Academy of Sciences of the U.S.S.R. are perhaps the most honoured citizens of the Soviet Union. Not only do they enjoy, by Soviet standards, very comfortable incomes, superior recreational and medical facilities, and preferential access for their children to the best institutions of higher education, but they also receive substantial additional emoluments by virtue of their academy membership, in addition to their salaries as directors of scientific laboratories. Individual party officials who abuse their positions—for example, by building for their personal use excessively luxurious private homes, using materials illegally obtained at bargain prices—are from time to time sharply criticized or even severely punished, but, according to statements by leading Soviet dissidents, the traditional practice of regularly distributing special pay envelopes to party officials persists.

Privileged persons

Soviet society is thus far from being egalitarian. The people are also, by the standards of advanced Western industrial nations, rather poor. The average wage of industrial workers, even at the most generous estimate, is about $200 a month U.S. Though rents are low, and housing, though still of very poor quality, is increasingly abundant, the prices of food and clothing are far higher, in proportion to income, than they are in the West. Both husband and wife usually must work in order to make ends meet, though it should also be said that the overwhelming majority of women apparently want to work at paying jobs. In certain respects, women have achieved much in the Soviet economy, as is exemplified by the dominant role of women in the medical profession. Most of the best jobs in all fields, however, and particularly in industry, politics, and public administration, are occupied by men.

Standard of living

Since Stalin's death, the Soviet authorities have made great efforts to raise the standard of living above the very low level of the Stalin era. Perhaps even more important is the program, begun under N.S. Khrushchev, to raise the incomes of the lowest paid categories of workers and in general to improve the economic situation of the workers and collective farmers. This has involved, among other measures, the extension of social security coverage to the collective farm peasantry, who until the mid-1960s remained outside the social security system. Since the mid-1950s there has been a fairly steady rise in Soviet living standards, though this has not necessarily increased the contentment of the mass of the population and has been interrupted by bad harvests, such as those of 1972 and 1975. There has been substantial improvement in a previously dire housing and very poor food situation. There has also been rapid progress in availability of basic consumer goods—excellent in terms of refrigerators, quite good in terms of washing machines and even of television sets, but still very poor in terms of private automobiles. The Soviet population by the later 1970s, however, was still at a standard of life far below that of all of the developed capitalist countries and of several of the eastern European Communist nations.

Health services. Public health and education are perhaps the areas in which Soviet achievements have been greatest, especially in quantitative terms. The official Soviet claim is that medical services are "free and accessible." It is also asserted in Soviet sources that the Soviet Union has about a quarter of all the physicians in the world and that it has more than 3,000,000 hospital beds. Even if these claims are somewhat exaggerated, and in any case say nothing about the quality of medical care, there is no doubt that the Soviet Union's record in this field has been impressive, particularly in view of the backward conditions that prevailed in the tsarist period and the difficulties that had to be overcome along the way. It is also apparent, however, that Soviet medical services—which are administered centrally by the Ministry of Health—suffer from the evils common to overcentralized, bureaucratically administered programs: excessive time is spent by both doctor and patient in record keeping and patients often experience long delays in getting treatment. Nor can it be said that the Soviet Union has a record of innovation in medical research and development even remotely comparable to that of the United States and other advanced Western nations, and the quality of care available to Soviet citizens (so far as it can be judged on the limited information to which foreign students have access) is not equal to the best offered by Western medicine. There is no doubt, however, that death rates have been substantially reduced (though beginning in the 1970s the infant mortality rate rose substantially and life expectancy fell slightly), and the health of the population compares well with that of citizens of other industrial countries.

Education. An objective Western observer cannot fail to be impressed by the achievements of the Soviet Union in the field of education, especially quantitatively. At the same time, he will also be struck by the peculiarities of the vast Soviet program as it is viewed from the perspective of Western traditions. According to Soviet statistics, 46,468,000 pupils were enrolled in "general educational" schools of all kinds in 1976–77 (a drop from six years previously, when the figure exceeded 49,000,000). At the level of higher education, with a reported 4,950,000 students in 1976–77, the Soviet educational effort was surpassed quantitatively only by that of the United States, which had about 8,000,000 students in four-year colleges and universities (and another 3,000,000 in two-year institutions).

Politics and education

The Soviet educational system is characterized by extreme centralization and uniformity in structure and curriculum and by heavy-handed political control and political indoctrination. Especially (but not only) in the humanities and social sciences, textbooks and classroom instruction are permeated by Marxist–Leninist propaganda. Applicants for higher education must pass examinations in political subjects and must continue to devote several hours a week to the study of political subjects even after admission to medical or other professional schools.

Even in quantitative terms, however, Soviet education has thus far by no means achieved the goals set by the CPSU leadership. Universal secondary education (10 years of schooling) has not been attained, despite many years of predictions in the official press that it was imminent. Certainly the country is far from having made education equally available to all of its youth. Rural, as

contrasted to urban, youth are still at a great disadvantage in access to secondary and still more to higher education. As for access to the best universities, such as the state universities of Moscow and Leningrad, or to the best specialized scientific, economic, and other institutes, children of workers and peasants are at an almost hopeless disadvantage in comparison with children of professional people, administrators, and officials. The advantages of the intelligentsia in this respect are apparently so great that many students of the Soviet Union believe that a hereditary class, or caste, system may be forming, if indeed it has not already taken shape. Great efforts are made to provide opportunities to work toward higher education, by evening or correspondence study, etc., for young people who have chosen or have been forced to take jobs in industry upon graduation from primary or secondary school. It is generally agreed, however, that, useful though such efforts may be to the economy, they come nowhere near to equalizing opportunities for higher education.

Education in the Soviet Union is a formidable instrument of national policy. It provides the national economy, the armed forces, science and engineering, and the communications media with an abundance of reasonably well-trained man- and womanpower. These young people have acquired basic skills in a highly organized setting, in which discipline and conformity to authority are stressed. Also, though almost half of the Soviet population belongs to non-Russian national minorities, an apparently successful effort is made to teach all graduates (at least of the secondary schools and of course of higher educational institutions) to read and speak Russian effectively. Indeed, for non-Russians, mastering Russian—in addition to a "sound" political outlook—is a prerequisite for career advancement. Certainly the educational system is one of the sinews of the Soviet state. It is possible, however, that its successes, by providing for some individuals at least a basis for critical thought, may cause trouble for Soviet rulers in the future.

(F.C.B.)

VI. Cultural life and institutions

THE CULTURAL MILIEU

The roots of Soviet literature and arts lie deep in folk traditions. The vast extent, multinational nature, and varied history of the area that was to become the Soviet Union produced an exceptionally rich and diverse folk culture that continues to manifest itself in oral poetic forms, songs, dances, folk instrumental music, and various forms of decorative and applied art. The Russian Revolution and the establishment of the Soviet Union introduced fundamental changes into the cultural milieu. The process of change has been helped by the reduction of illiteracy and by expansion of education; by an immense growth in the output of printed matter; and by increasing the number of cultural institutions of all kinds, especially in the rural areas.

Multinational culture

One of the distinguishing features of creative life in the Soviet Union is its multinational character. Russian literature, music, and art had achieved international significance in the 19th century and were not without influence on the cultural life of all the peoples of the Russian Empire. Despite their somewhat lower overall level of economic and cultural evolution, many of the minority peoples—from the Baltic to Armenia, from the Ukraine to Central Asia—also had highly distinctive cultures and art forms. During the Soviet period, these have, in the majority of cases, been preserved and developed, within the official Soviet policy that all art should be "national in form and Socialist in content." The annual festivals of the arts and literature of the union republics (held in Moscow), together with republican festivals of the arts of other republics, help to assist the understanding of the cultures of the many different peoples of the Soviet Union. Annual competitions in the various arts further the same aim. All these events enjoy great popularity, as attested by large attendance figures.

Specifically Soviet culture, as distinct from Russian or any other national culture, is characterized by what is known as "Socialist Realism," and all national cultures are expected to conform to this approved artistic style. It consists of an emphasis on a Communist approach; "revolutionary humanism"; a popular spirit and appeal; an optimistic vision of the future; and a close connection with what is viewed as actual life. The characters portrayed are to embody what is conceived as the reality of contemporary Soviet life and the ideals of Soviet society, both nationally and internationally.

Socialist Realism

Soviet culture has also become known internationally (often through exhibitions, film distribution, and performances by leading individuals and cultural bodies) and has had a considerable influence, particularly in the other Communist countries. Internally, the highest recognition of creative work is indicated by award of the Lenin and State prizes and by republican prizes. Soviet cultural work is under the direction of the all-union Ministry of Culture; related union and autonomous republican bodies; and the cultural organs of the local soviets. Many mass organizations are also involved in cultural life.

(The following descriptive account of the main aspects of Soviet cultural life may be supplemented by reference to the articles on the union republics. Additional perspectives are offered in the articles LITERATURE, WESTERN; VISUAL ARTS, WESTERN; MOTION PICTURES, HISTORY OF; MUSIC, WESTERN; DANCE, WESTERN; and ARCHITECTURE, ART OF. See also ARCTIC PEOPLES AND CULTURES; CENTRAL ASIAN PEOPLES, ARTS OF. There are also biographies of the major figures of Russian and Soviet culture.)

THE INDIVIDUAL ARTS

Literature. Belles lettres occupy a leading place in Soviet cultural life, drawing on a multinational tradition that has produced acknowledged world masterpieces. The great poets of Central Asia—Neẓāmī, Fuzûlî, and ʿAlī Shīr Navāʾī—were writing in the Middle Ages. Shota Rustaveli's 12th-century Georgian poem *The Knight in the Tiger's Skin* is another landmark in world literature, and the major Transcaucasian poets of the 18th century, Sayat-Nova and Vagif, together with their contemporary the Ukrainian poet and philosopher Hryhoriy Skovoroda, also contributed much to this tradition.

But it was 19th-century Russian literature that acquired the greatest world standing. Giants of this period include Aleksandr Pushkin, the "father of Russian literature," whose masterpiece, *Eugene Onegin,* is a verse novel that achieved an encyclopaedic portrayal of Russian life; Mikhail Lermontov, whose *Hero of Our Time* is a brilliant novel in a Caucasian setting; Nikolay Gogol, whose unfinished masterpiece, *Dead Souls,* was a remarkable portrayal of the ruling class of the time; Aleksandr Herzen, whose memoirs, *My Past and Thoughts,* offered a penetrating commentary on Russian society; Ivan Turgenev, whose major novel *Fathers and Sons* produced much discussion among Russian youth and their parents alike; Ivan Goncharov, whose *Oblomov* is an acknowledged triumph of psychological portrayal; and, perhaps most famous of all, the great novelists Leo Tolstoy (*War and Peace*) and Fyodor Dostoyevsky (*Crime and Punishment*). There were also many other figures of major status active at the time, ranging from the dramatist Anton Chekhov to the poet and novelist Ivan Bunin, who in 1933 became the first Russian recipient of the Nobel Prize.

19th-century Russian achievements

Their ranks also included figures in what became the union republics. Among them were Taras Shevchenko and Ivan Franko (the Ukraine); Yanka Kupala and Yakub Kolas (Belorussia); Akaki Tsereteli and Ilia Chavchavadze (Georgia); Jānis Rainis and Andreijs Upīts (Latvia); Julija Žymantienė-Žemaitė (Lithuania); Eduard Vilde (Estonia); Khachatur Abovyan and Ovanes Tumanyan (Armenia); and many others.

The foundations of Soviet literature were laid by the great working class writer Maksim Gorky, whose *Mother* (1906) initiated a series of the first classics of Socialist Realism; and by the major early Soviet poet Vladimir Mayakovsky. Contemporary Soviet literature is marked by a wealth of forms and genres, which, although multi-

20th-century achievements

lingual and multinational, are integrated by the common Communist ideological content. In prose fiction, Soviet opinion ascribes major significance to the work of Aleksey N. Tolstoy (*The Road to Calvary, Peter the Great,* and other novels), whose attitude to life may be indicated by the statement in his autobiography that he loved life and opposed every kind of abstraction with all the power of his temperament. Mikhail Sholokhov, a Nobel Prize winner, created in his novels *The Silent Don* and *Virgin Soil Upturned* epic accounts of the struggle to construct the new society. Aleksandr Fadeyev was the author of *The Rout* and *The Young Guard,* novels that deal realistically with partisan and guerrilla activities in the Civil War and World War II; he was for three decades head of the Union of Soviet Writers. Aleksandr Serafimovich's *Iron Flood,* completed in 1923 when the author was 60, was one of the first significant novels about the Revolution. Dmitry Furmanov's novels *Chapayev* and *Mutiny* were among the first and best on the Civil War. Nikolay Ostrovsky, who died in 1936 at the age of 32, yet produced, in *How the Steel Was Tempered,* a classic autobiographical statement of the revolutionary generation. There are many other major figures, including Aleksandr Blok, the great poet of the Revolution; Ilya Ehrenburg; Boris Pasternak, a Nobel Prize winner, lyric poet, and translator of Georgian, German, and English poets; Konstantin Fedin; and Leonid Leonov. The great poet Anna Akhmatova incurred the displeasure of the authorities but was "rehabilitated" in 1959. Another major poet, Osip Mandelstam, perished in the Great Purge of the 1930s. The Tadzhik author Sadriddin Ayni; the Uzbek poet and dramatist Niyazy; the Tatar poet and writer Musa Dzhalil; the Kazakh author Mukhtar Auezov; and writers from the Baltic republics also have achieved national and in some cases international fame.

Oral folk traditions are continued in the work of such outstanding bards as Jambil Jabayev (called Jambil) in Kazakhstan, Suleyman Stalsky in Dagestan, and Toktogul in Kirgiziya.

Contemporary literature

In contemporary Soviet literature, Soviet opinion ascribes significance to such authors as Galina Nikolayeva, Mikhail Bubennov, and Sergey Sartakov, whose novels have described post-World War II agricultural and industrial life; Konstantin Simonov (*The Living and the Dead, They Were Not Born as Soldiers, The Last Summer,* and other works), Aleksandr Chakovsky (*Blockade*), and Vadim Kozhevnikov (*Shield and Sword*), who, with many others, have written of the war years; the Kirgiz author Chingiz Aytmatov, whose novel *Dzhamilya* is about a young woman who breaks away from traditional ways of life; and a host of others. The Nobel Prize-winning novelist Aleksandr Solzhenitsyn was exiled in 1974. The poets Yevgeny Yevtushenko, Andrey Voznesensky, Stepan Shchipachov, Nikolay Tikhonov, Rasul Gamzatov (of Dagestan), Vsevolod Rozhdestvensky, Leonid Martynov, Bella Akhmadulina, among many others, have achieved great contemporary popularity. Soviet children's literature—prominent writers of which include Korney Chukovsky, Agniya Barto, and Sergey Mikhalkov—is well-known at home and abroad.

Visual arts. Paleolithic and Neolithic art has been discovered in the Caucasus, in Central Asia, in Siberia, and on the northern Black Sea coast. In the Middle Ages the most notable work was done by the great icon painters of the Novgorod and Moscow schools, among whom the finest were Theophanes the Greek, Andrey Rublyov, and Dionisy. In the 19th century the modern realistic schools of Ivan Kramskoy, Ilya Repin, and Isaak Levitan were founded.

Major works of the revolutionary art of the Soviet period include the scenes from Revolutionary history associated with Isaak Brodsky, Valentin Serov, and Kuzma Petrov-Vodkin; the genre painting of Nikolay Kasatkin and Yury Pimenov; and the portraits of Mikhail Nesterov and Aleksandr Gerasimov. Artists from minority nationalities, including Ural Tansykbayev of Uzbekistan, Martiros Saryan of Armenia, and others, have recorded the varying scenery of the country. Yevgeny Lansere and Aleksandr Deyneka have produced monumental works, while the graphic arts, represented by Viktor Deni, Dmitry Moor, and "Kukryniksy" (a nickname of Mikhail Kupriyanov, Porfiry Krylov, and Nikolay Sokolov) are marked by great variety. Vera Mukhina's group "Worker and Collective Farm Woman" and the memorial ensembles created under the direction of Yevgeny Vuchetich in East Berlin and at Volgograd, as well as works by Sergey Konenkov and Sergey Merkurov, are regarded as masterpieces of monumental sculpture.

Monumental and decorative art

In the decorative and applied arts field, there is great demand for the products of factory and folk industries, which include pottery, glassware, and carpets; carving on wood, stone, bone, and horn; silverware; and delicate lacquer miniatures.

Persons working in the fine arts, as well as theatre designers, film cartoonists, and art scholars, are members of the Union of Artists and the Academy of Arts.

Music. The works of the classics—including those of Mikhail Glinka, Modest Mussorgsky, Aleksandr Borodin, Nikolay Rimsky-Korsakov, Peter Ilich Tchaikovsky, and Sergey Rachmaninoff—remain very popular. Soviet music is regarded as embodying important modern themes as well as those of the past: an example is provided by the symphonies of Dmitry Shostakovich, particularly his *Seventh* (*Leningrad*), *11th* (dedicated to the Revolution of 1905), and *12th* (dedicated to the October Revolution of 1917). Soviet opera reflects both national epics (as in Uzeir Hajjibekov's *Ker-Ogly*) and historical and modern themes, respectively illustrated by Sergey Prokofiev's *War and Peace* and *Semyon Kotko.* Soviet ballet music ranges from Prokofiev's *Romeo and Juliet* to Aram Khachaturian's *Gayane.* There is a union of Soviet composers.

Soviet performing artists have traditionally won international recognition, and to such well-known names as those of the pianist Emil Gilels and the violinist David Oistrakh have been added those of a brilliant new group including, among many others, the pianist Svyatoslav Richter, the cellist Mstislav Rostropovich (who, however, while serving as conductor of the National Symphony Orchestra in Washington, D.C., was deprived of Soviet citizenship), and the conductor Kiril Kondrashin.

Large folk-song choirs from differing national backgrounds have also won fame at home and abroad, and professional musical activities generally blend into and enrich amateur and folk music of all sorts, from balalaika orchestras to brass bands. Each summer, song festivals are held in public parks and squares, with individual performers and choirs participating. The international P.I. Tchaikovsky Competition for pianists, violinists, cellists, singers, and ballet dancers is held every four years in Moscow. At the first competition, in 1958, the young American pianist Van Cliburn won first prize.

Music festivals

Theatre and ballet. Soviet theatre, embodying the principles of Socialist Realism, has inherited much from the past, notably the tradition of Russian Realist theatre. By the second half of the 19th century and the turn of the 20th, this tradition included the Maly Theatre of Moscow (founded in 1824) and the renowned Moscow Art Theatre, founded in 1898 by the theatrical innovator Konstantin Stanislavsky, often characterized as the founder of modern acting, and his colleague V.I. Nemirovich-Danchenko.

Soviet theatrical growth was marked by the interaction of the old-established Bolshoi ("Large"), Maly ("Small"), and Art theatres in Moscow and the Aleksandrinsky, Mariinsky, and Mikhaylovsky theatres in St. Petersburg (now Leningrad), with the new Soviet theatres. Major Soviet drama producers have used a variety of forms and methods, often based on those of Stanislavsky. In similar fashion, the great traditional Russian actors passed on their skill to the newer generations of the Soviet period.

Of particular interest are the developments outside the leading centres, Moscow and Leningrad; such peoples as the Uzbeks, Kazakhs, Turkmens, and Kirgiz, who had no professional theatre before the Revolution, have

created vigorous national theatre based on traditional performing arts forms.

In addition to Soviet drama, the country's theatres put on many plays from the world and Russian literary traditions, from Chekhov to Shakespeare, Molière to Schiller. Modern foreign playwrights—Bertolt Brecht, Jean Anouilh, and Arthur Miller among them—are also often performed. These contemporary developments have been associated with the emergence of a young school of producers and actors.

Ballet stars

Contemporary Soviet ballet theatres strive to preserve all that is regarded as the best of the great pre-Revolutionary era while creating a repertory that conforms to Soviet reality and the interests of new audiences. Thus, new, heroic qualities appeared in the dancing of such stars as Yekaterina Geltser, A.N. Yermolayev, and S.M. Messerer, and the 1930s saw the development of many-act ballets with highly developed plots. That decade also saw the flowering of the talents of Galina Ulanova, the first Soviet prima ballerina *assoluta*. Ballet companies were also organized with striking results in the constituent republics. In later decades a new generation of performers—including Maya Plisetskaya, Yekaterina Maksimova, Vladimir Vasilyev, and many more—has emerged. Ballet in the West has been enriched by the emigration of several Soviet-trained dancers, among them the great Rudolf Nureyev and Mikhail Baryshnikov. Russian ballet has of course had a major influence on world dance, and Soviet companies continue to give many well-received performances abroad. There are almost 20 ballet schools, the best known being located in Moscow and Leningrad. Amateur ballet is also well developed.

The major Soviet theatres include the Gorky Moscow Art Theatre (MKhAT), whose curtain bears a sea-gull emblem as a symbol of the theatre's close links with Chekhov; the Maly Theatre in Moscow, one of the country's oldest; the Pushkin Academic Drama Theatre in Leningrad (the Aleksandrinsky Theatre from 1832 to 1920), after the Maly the great centre of Russian Realism; and the Academic Gorky Bolshoy State Drama Theatre in Leningrad, founded in 1919. In Kiev are found the Leslya Ukrainka and Ivan Franko theatres.

Opera and ballet theatres, in addition to the renowned Bolshoi Theatre (formally, the State Academic Bolshoy Theatre of the U.S.S.R., or GABT; founded in 1772), include the Kirov State Academic Theatre of Opera and Ballet in Leningrad (formerly the Mariinsky Theatre, founded 1836) and the same city's Maly Theatre (formally, Leningrad Academic Maly Theatre of Opera and Ballet), as well as the Academic Musical Theatre in Moscow, named after Stanislavsky and Nemirovich-Danchenko. The Shevchenko Theatre of Opera and Ballet is in Kiev. The Moscow Operetta Theatre is the major musical comedy stage.

Children's theatre

Almost all the country's theatres present special performances for children, and there are more than 150 theatres for children and young people, including puppet theatres. The most important include the Moscow Central Children's Theatre, the Moscow Young Spectators Theatre and its equivalent in Leningrad, and youth theatres in the major cities. The unique State Central Puppet Theatre in Moscow, founded and directed by Sergey Obraztsov, has delighted patrons of all ages in many countries.

Actors and producers are trained at the Lunacharsky Theatre Art Institute (GITIS), the Shchepkin and Shchukin theatre schools, the Moscow Art Theatre School, and many other institutions. Actors belong to the All-Russian Theatrical Society and analogous republic organizations.

Variety and circus. Both the variety stage and the circus are extremely popular in the Soviet Union, and the Soviet Folk Dance Ensemble, the Red Army Song and Dance Ensemble, and national groups have achieved world fame. Illusionists, acrobats, jugglers, and others appear in variety performances and also in circuses, which are well known for the variety of their performances and their high level of technique. Entertainments range from performances by the animals trained by V.L. Durov, noted for his scientific approach, to the clowning of Oleg K. Popov and others, much of which has a satirical, topical character.

Motion pictures. The cinema is undoubtedly the most popular form of Soviet art, the first Russian film having appeared in 1907. Historical themes and the developing Soviet experience provided the major subject matter of earlier films, and landmark productions of international renown include *The Battleship Potemkin* (1925; director Sergey Eisenstein), regarded as one of the world's great films; *Mother* (1926; director Vsevolod I. Pudovkin), from Gorky's novel, with Nikolay Batalov in the main role; *Earth* (1930; director Aleksandr P. Dovzhenko); and *Peter the Great* (1937–38; director Vladimir M. Petrov), from A.N. Tolstoy's novel.

Postwar films

Later years were characterized by a change of artistic emphasis to contemporary themes, notably the upheavals of World War II. Films include *The Cranes Are Flying* (1957; director Mikhail K. Kalatozov), *Ballad of a Soldier* (1961; director Grigory N. Chukhray), and *The Fate of a Man* (1959; director and leading actor Sergey F. Bondarchuk). There has also been a considerable output of biographical, documentary, and semidocumentary films. Films based on Russian and world classics include the *Hamlet* and *King Lear* of Grigory Kozintsev, several dramatizations of the novels of Dostoyevsky, and epics in several parts based on Tolstoy's *War and Peace* and Sholokhov's *Silent Don*. Other works popular with Soviet audiences include *The Living and the Dead, Soldier's Father,* the five-part epic *Liberation, The Belorussian Terminal,* and many others. International film festivals, held biennially in Moscow and Tashkent, are also very popular. There is a union of film workers (founded 1957).

Amateur activities. Folk art and mass nonprofessional art are widely developed. They include the production and performance of artistic works by dancers, ballad singers, singers of *chastushki* (rhymed couplets), and instrumentalists. There are many crafts workers. Clubs, palaces of culture, houses of folk art, and similar centres are the main base for a host of amateur organizations. Professionals may take part in amateur activities, and vice versa. Regular festivals, reviews, and competitions help raise the level of skill, and many professional groups and performers were originally amateur.

CULTURAL AND EDUCATIONAL INSTITUTIONS

Cultural and educational institutions include clubs, the institutions known as people's universities, libraries, lecture halls, museums, and parks of rest and culture. The stated basic object of these bodies is to spread political and scientific knowledge; to propagate the achievements of science and technology, of the arts, and of literature; and to organize the cultural leisure of the population. The Soviet authorities favour communal rather than individual leisure activity.

Clubs. There are some 135,000 clubs, all but 20,700 of them in rural areas. These are attached to and organized by individual state economic enterprises and collective farms, which are obliged to devote part of their income to the cultural and recreational functions carried out by the clubs. Membership is voluntary, but there are often strong social pressures on the individual to join, given the Soviet ethos. The clubs organize recreational evenings, concerts, lectures, and so on and provide facilities for amateur organizations such as dramatic, choral, dance, and musical societies.

Lecture halls and people's universities of culture. Lecture halls are to be found at many factories, farms, and other enterprises. They are important centres for the propagation of the state ideology as well as for the dissemination of cultural and scientific knowledge. A leading role is played by the all-union Znaniye (Knowledge) society, which organizes lectures, often providing travelling lecturers on particular topics in science and culture. The voluntary people's universities of culture have the aim of broadening people's political and cultural horizons through part-time courses of one to three years'

duration on such topics as Soviet and foreign literature, aesthetics, and the sciences.

Libraries. A considerable portion of the output of books, magazines, and newspapers enters the state library system. The Lenin State Library in Moscow is the country's largest, housing about 27,000,000 volumes in 173 languages. Other leading libraries are found in Leningrad at the Saltykov-Shchedrin Public Library (17,000,000 volumes) and at the U.S.S.R. Academy of Sciences; in Kiev at the Ukrainian S.S.R. Academy of Sciences; and in Moscow at the Moscow State University (MGU), whose Gorky Scientific Library houses 6,000,000 volumes.

Museums and galleries. Museum attendance in the Soviet Union is high. Most museums and galleries of art are state institutions, administered by the Ministry of Culture, but there are also museums established by volunteer enthusiasts, often found in small towns and rural areas. Famous museums include (in Moscow) the Central Lenin Museum, the Central Museum of the Revolution, the Armoury Museum in the Kremlin, the State Historical Museum, the Polytechnic Museum, the State Pushkin Museum of Fine Arts, and the State Tretyakov Gallery. Leningrad has the Hermitage and the State Russian Museum, while Kiev is the home of a historical museum, the Kiev State Museum of Ukrainian Art, and the Kiev State Museum of Russian Art. A unique place is occupied by the permanent Exhibition of National Economic Achievements, in Moscow; founded in 1959, it attracts 10,000,000 visitors annually.

Parks of rest and culture. Parks are administered by the state and by trade unions. The largest include the Gorky Central Park of Culture and the Sokolniki and Izmaylovsky parks in Moscow and the Kirov Central Park of Culture and Rest in Leningrad. Parks are the scene of many cultural events and lectures, including open-air Chess matches in the spring and summer.

COMMUNICATIONS MEDIA

The press. Soviet citizens are noted for their devotion to reading, and the Soviet Union produces more printed material than any other country. It takes first place internationally both in the number of books published and in the number of translations, producing works in 89 Soviet and more than 50 foreign languages. About a quarter of the total output deals with political, economic, and social matters, and one-tenth of the titles (and one-third of the copies) are belles lettres. About 30 percent of the books printed are educational literature, and books for children have become increasingly important. There are more than 200 book and some 15 newspaper and magazine publishing houses attached to a variety of state bodies.

The CPSU and the press

Newspapers include universally available Russian-language dailies, regional dailies in the languages of the constituent republics, and papers of smaller circulation published by individual ministries, economic enterprises, collective farms, and other organizations. Newspapers are owned and published by the Communist Party organization of the area or enterprise within which they circulate. The main all-union dailies are *Pravda* ("Truth"), the organ of the Central Committee of the CPSU; *Izvestiya* ("News"), the government newspaper, published by the Presidium of the Supreme Soviet; *Komsomolskaya Pravda,* published by the Central Committee of the Komsomol; and *Trud* ("Labour"), the organ of the All-Union Central Council of Trade Unions. *Pionerskaya Pravda,* the children's newspaper, is published thrice weekly, also by the Central Committee of the Komsomol. *Krasnaya Zvezda* ("Red Star"), the organ of the Ministry of Defense, is published six times a week. Other papers published two or three times a week are owned by the appropriate industrial ministry or trade union or by both jointly; these include, among others, special newspapers for workers in the railway, timber, and construction industries and for cultural workers, teachers, the medical profession, and athletes.

Content of all newspapers is closely scrutinized by party officials, and deviations from official policy are either eliminated by the editorial board or vigorously criticized after publication. On the other hand, newspapers often publish and investigate complaints by individual citizens and in this sense do provide a medium for the expression of public opinion. Criticism of the political and economic system as such, however, does not appear in the Soviet press. While content of individual newspapers varies according to the special interests of their readers, there is much common material supplied by official press agencies, of which by far the largest are TASS (the Soviet Telegraph Agency, founded in 1925) and APN (the Novosti Press Agency, founded in 1961).

Magazines

The number and circulation of weekly and monthly magazines increased greatly during the 1960s, though all are subject to the same restraints as the newspapers and are controlled by party organizations. Among the weeklies, two of the most important are *Literaturnaya Gazeta* ("Literary Gazette"), published by the Union of Soviet Writers, and *Ekonomicheskaya Gazeta* ("Economic Gazette"), published by the Central Committee of the CPSU, which deals particularly with problems of industrial organization. There are about 20 monthly publications for teachers, a dozen literary monthlies, and a variety of scholarly journals.

Radio and television. Soviet radio and television, like the press, are closely controlled by the state and party, and transmissions on ideological, political, economic, and social affairs, to which much time is devoted, reflect official policies. In addition to such aspects there is an increasing volume of light entertainment and special programs directed at, for example, preschool children, collective farmers, and the armed forces. Internal broadcasts are conducted in 70 languages, external broadcasts in 57 languages.

Although the first Soviet television transmissions began in 1931, it was not until the 1960s that a nation-wide system was developed. By the mid-1970s, television was accessible to about 80 percent of the total population. By use of communications satellites, the peoples of Central Asia, Siberia, the Far North, and the Far East receive programs originating in Moscow as well as the transmissions of their local stations. The Moscow television centre is the nerve centre of the Soviet system, from which the great majority of programs, whether live or filmed, originate, and the 1,758-foot (536-metre) tower at Ostankino is a striking feature of the city skyline, transmitting one colour and five black and white channels.

(I.V.Ko./J.C.D.)

BIBLIOGRAPHY

The land and people of the Soviet Union: Important works in Russian include AKADEMIYA NAUK SSSR, BOTANICHESKY INSTITUT, *Растительный покров СССР,* 2 vol. (1956), a detailed description of vegetation types; AKADEMIYA NAUK SSSR, ZOOLOGICHESKY INSTITUT, *Животный мир СССР; обзор фауны территории Союза и отчасти прилежащих стран на эколого-фаунистической и зоогеографической основе,* 5 vol. (1937–58), an excellent account of animal life; N.A. BOBRINSKOY, *Животный мир и природа СССР,* 2nd ed. (1960), a semipopular work providing a description of animal life; V.V. ALEKHIN, *Растительность СССР в основных зонах* (1951), a detailed description of the zonal vegetation types written by a major Soviet geobotanist; B.P. ALISOV, *Климат СССР* (1956), a comprehensive study of climate; A.A. BORISOV, *Климаты СССР,* 3rd ed. (1967; Eng. trans., *Climates of the U.S.S.R.,* 1965), dealing in detail with regional climates; L.S. BERG, *Географические зоны Советского Союза,* 2nd–3rd ed., 2 vol. (1947–52), detailed information on the natural landscapes according to geographical zones; L.K. DAVYDOV, *Гидрография СССР (воды и суши),* 2 vol. (1953–55), an authoritative work on hydrology; S.V. KALESNIK *et al.* (eds.), *Советский Союз, географическое описание в 22 томах* (1966–72), a 22-vol. geography of the republics and regions; N.I. MIKHAYLOV and N.A. GVOZDETSKY, *Физическая география СССР: Азиатская часть,* 2nd ed. (1970), a comprehensive study of the physical geography of the Asiatic regions, a continuation of F.N. MILKOV and N.A. GOVOZDETSKY, *Физическая география СССР,* 3rd ed. (1969), a general survey of the physical geography of the Soviet Union as a whole; F.N. MILKOV, *Природные зоны СССР* (1964), a detailed yet readable account of the natural regions; V.V. POKSHISHEVSKY, *География Населения СССР*

(1971), the standard work on population geography. English-language studies of the land and people include L.S. BERG, *Natural Regions of the U.S.S.R.* (1950; Eng. trans. from 2nd Russian ed., 1938); VIKTOR BUNAK *et al.*, *Contributions to the Physical Anthropology of the Soviet Union* (1960); JOHN P. COLE and FRANK C. GERMAN, *A Geography of the U.S.S.R.*, 2nd. ed. (1970); JOHN C. DEWDNEY, *A Geography of the Soviet Union*, 2nd ed. (1971); INNOKENTY P. GERASIMOV *et al.*, *Natural Resources of the Soviet Union, Their Use and Renewal* (1971); JAMES S. GREGORY, *Russian Land, Soviet People: A Geographical Approach to the U.S.S.R.* (1968); DAVID J.M. HOOSON, *The Soviet Union: People and Regions* (1966); PAUL E. LYDOLPH, *Geography of the U.S.S.R.*, 3rd ed. (1977); ROY E.H. MELLOR, *Geography of the U.S.S.R.* (1964); WILLIAM H. PARKER, *An Historical Geography of Russia* (1968) and *The Soviet Union* (1969); V.V. POKSHISHEVSKY, *Geography of the Soviet Union: Physical Background, Population, Economy* (1974); PHILIP R. PRYDE, *Conservation in the Soviet Union* (1972); SERGEI P. SUSLOV, *Physical Geography of Asiatic Russia*, ed. by JOSEPH E. WILLIAMS (1961; orig. pub. in Russian, 1947).

The economy: Useful introductory works include MORRIS BORNSTEIN and DANIEL FUSFELD, *The Soviet Economy: A Book of Readings*, 3rd ed. (1970); GEORGE R. FEIWEL, *New Currents in Soviet-Type Economies: A Reader* (1968); MARSHALL I. GOLDMAN, *The Soviet Economy: Myth and Reality* (1968); F.D. HOLZMAN (ed.), *Readings on the Soviet Economy* (1962); ALEC NOVE, *The Soviet Economy: An Introduction*, 3rd rev.ed. (1968), and *An Economic History of the U.S.S.R.* (1969); HARRY SCHWARTZ, *The Soviet Economy Since Stalin* (1965). Studies of the growth and development of the Soviet economy include ABRAM BERGSON, *The Real National Income of Soviet Russia Since 1928* (1961); ROBERT W. CAMPBELL, *The Soviet-Type Economies: Performance and Evolution*, 3rd ed. (1973; orig. pub. 1960 as *Soviet Economic Power: Its Organization, Growth, and Challenge);* MAURICE DOBB, *Soviet Economic Development Since 1917*, 6th ed. (1966); DONALD R. HODGMAN, *Soviet Industrial Production, 1928–1951* (1954); RAYMOND HUTCHINGS, *Soviet Economic Development* (1971); G. WARREN NUTTER *et al.*, *Growth of Industrial Production in the Soviet Union* (1962); VLADIMIR G. TREML and ROBERT FARRELL (eds.), *The Development of the Soviet Economy: Plan and Performance* (1968). Economic resources and their development are treated in GEORGE B. CRESSEY, *Soviet Potentials: A Geographic Appraisal* (1962); JOHN C. DEWDNEY, *The USSR* (1976); A.N. LAVRISHCHEV, *Economic Geography of the USSR* (1969); R.S. MATHIESON, *The Soviet Union: An Economic Geography* (1975); N.P. NIKITIN *et al.* (eds.), *Экономическая География СССР* (1973); THEODORE SHABAD, *Basic Industrial Resources of the U.S.S.R.* (1969). Organization and output in various branches of the economy are treated in ROBERT W. CAMPBELL, *The Economics of Soviet Oil and Gas* (1968); CONGRESS OF THE UNITED STATES, JOINT ECONOMIC COMMITTEE, SUBCOMMITTEE ON FOREIGN ECONOMIC POLICY, *New Directions in the Soviet Economy* (1966); LESLIE DIENES, *Locational Factors and Locational Developments in the Soviet Chemical Industry* (1969); IAIN F. ELLIOT, *The Soviet Energy Balance: Natural Gas, Ohter Fossil Fuels, and Alternative Power Sources* (1974); DAVID GRANICK, *Soviet Metal-Fabricating and Economic Development: Practice Versus Policy* (1967); G.W. HEMY, *The Soviet Chemical Industry* (1971); JORDAN A. HODGKINS, *Soviet Power: Energy Resources, Production and Potentials* (1961); DMITRI B. SHIMKIN, *Minerals: A Key to Soviet Power* (1953); ERICH STRAUSS, *Soviet Agriculture in Perspective: A Study of Its Successes and Failures* (1969); LESLIE SYMONS, *Russian Agriculture: A Geographic Survey* (1972); LESLIE SYMONS and COLIN WHITE, *Russian Transport: An Historical and Geographical Survey* (1975). A valuable compendium of economic statistics for the Soviet period is ROGER A. CLARKE, *Soviet Economic Facts 1917–1970* (1972). Regional economic studies include VIOLET CONOLLY, *Beyond the Urals: Economic Developments in Soviet Asia* (1967) and *Siberia Today and Tomorrow: A Study of Economic Resources, Problems, and Achievements* (1976); E. STUART KIRBY, *The Soviet Far East* (1971); ALEC NOVE and J.A. NEWTH, *The Soviet Middle East: A Model for Development?* (1967). Soviet economic planning is the subject of a large literature, among which may be recommended EDWARD AMES, *Soviet Economic Processes* (1965); ABRAM BERGSON, *The Economics of Soviet Planning* (1964) and *Planning and Productivity Under Soviet Socialism* (1968); PHILIPPE J. BERNARD, *Planning in the Soviet Union* (1966); HENRI CHAMBRE, *L'Aménagement du territoire en U.R.S.S.* (1959); JANE DEGRAS and ALEC NOVE (eds.), *Soviet Planning: Essays in Honour of Naum Jasny* (1964); MICHAEL ELLMAN, *Economic Reform in the Soviet Union* (1969) and *Soviet Planning Today: Proposals for an Optimally Functioning Economic System* (1971); JERE L. FELKER, *Soviet Economic Controversies: The Emerging Marketing Concept and Changes in Planning, 1960–1965* (1966); JOHN P. HARDT *et al.* (eds.), *Mathematics and Computers in Soviet Economic Planning* (1967); ALEC NOVE, *Economic Rationality and Soviet Politics; or, Was Stalin Really Necessary?* (1964; British title *Was Stalin Really Necessary? Some Problems of Soviet Political Economy);* HOWARD J. SHERMAN, *The Soviet Economy* (1969); NICHOLAS SPULBER, *Soviet Strategy for Economic Growth* (1964); J. WILCZYNSKI, *Socialist Economic Development and Reforms* (1972); EUGENE ZALESKI, *Planning Reforms in the Soviet Union, 1962–66* (1967); ALFRED ZAUBERMAN *et al.*, *Aspects of Planometrics* (1967). Foreign economic relations are discussed in ALAN A. BROWN and EGON NEUBERGER (eds.), *International Trade and Central Planning: An Analysis of Economic Interactions* (1968); MARSHALL I. GOLDMAN, *Soviet Foreign Aid* (1967); MICHAEL KASER, *Comecon: Integration Problems of the Planned Economies* (1965); FRANK O'BRIEN, *Crisis in World Communism: Marxism in Search of Efficiency* (1965); P.J.D. WILES, *Communist International Economics* (1968).

Administration and social conditions: JOHN A. ARMSTRONG, *The Politics of Totalitarianism: The Communist Party of the Soviet Union from 1934 to the Present* (1961); FREDERICK C. BARGHOORN, *Politics in the U.S.S.R.*, 2nd ed. (1972); HAROLD J. BERMAN, *Justice in the U.S.S.R.: An Interpretation of Soviet Law*, rev. ed. (1963); H.J. BERMAN and JOHN B. QUIGLY (eds. and trans.), *Basic Laws on the Structure of the Soviet State* (1969); EMILY C. BROWN, *Soviet Trade Unions and Labor Relations* (1966); ARVID BRODERSEN, *The Soviet Worker* (1966); LEIF BJÖRK, *Wages, Prices and Social Legislation in the Soviet Union* (Eng. trans., 1953); ANTONY BUZEK, *How the Communist Press Works* (1964); L.G. CHURCHWARD, *Contemporary Soviet Government* (1968); ROBERT CONQUEST, *Power and Policy in the U.S.S.R.* (1961); MARGARET DEWAR, *Labour Policy in the U.S.S.R., 1917–1928* (1956; reprinted 1977); MERLE FAINSOD, *How Russia Is Ruled*, rev. ed. (1963); NIGEL GRANT, *Soviet Education* (1964); PHILIP HANSON, *The Consumer in the Soviet Economy* (1968); JOHN N. HAZARD, *Settling Disputes in Soviet Society* (1960) and *The Soviet System of Government*, 4th rev. ed. (1968); INTERNATIONAL LABOR OFFICE, *Trade Union Rights in the U.S.S.R.* (1959) and *The Trade Union Situation in the U.S.S.R.* (1960); A. INKELES, *Social Change in Soviet Russia* (1968); A. INKELES and R. BAUER, *The Soviet Citizen* (1968); EDWARD L. JOHNSON, *An Introduction to the Soviet Legal System* (1969); ROBERT G. KAISER, *Russia: The People and the Power* (1976); WOLFGANG LEONHARD, *Kreml ohne Stalin*, 3rd ed. (1962; Eng. trans., *The Kremlin Since Stalin*, 1962); M. LEWIN, *Russian Peasants and Soviet Power* (1968); DAVID MACE and VERA MACE, *The Soviet Family* (1964); B.Q. MADISON, *Social Welfare in the Soviet Union* (1968); MERVYN MATTHEWS, *Class and Society in Soviet Russia* (1972); KLAUS MEHNERT, *Der Sowjetmensch* (1958; Eng. trans., *The Anatomy of Soviet Man*, 1961); JAMES MILLAR (ed.), *The Soviet Rural Community: A Symposium* (1971); JACK MILLER, *Life in Russia Today* (1969); BARRINGTON MOORE, *Soviet Politics: The Dilemma of Power* (1950, reprinted 1976); J.P. NETTL, *The Soviet Achievement* (1967); W.H. PARKER, *The Russians* (1973); T.H. RIGBY, *Communist Party Membership in the U.S.S.R., 1919–1967* (1968); S.M. ROSEN, *Significant Aspects of Soviet Education* (1965); WALT W. ROSTOW and ALFRED LEVIN *et al.*, *The Dynamics of Soviet Society*, rev. ed. (1967); HARRISON E. SALISBURY (ed.), *Anatomy of the Soviet Union* (1968); LEONARD B. SCHAPIRO, *The Communist Party of the Soviet Union*, 2nd rev. ed. (1970), and *The Government and Politics of the Soviet Union*, 2nd rev. ed. (1967); SOLOMON SCHWARZ, *Labor in the Soviet Union* (1952); HEDRICK SMITH, *The Russians* (1976); S.V. UTECHIN, *Russian Political Thought: A Concise History* (1964); NICHOLAS P. VAKAR, *The Taproot of Soviet Society* (1962); LEONID VLADIMIROV, *The Russians* (1968). A valuable set of social statistics appears in ELLEN MICKIEWICZ (ed.), *Handbook of Soviet Social Science Data* (1973).

Cultural life and institutions: B.A. VVEDENSKY *et al.* (eds), *Союз Советских Социалистических Республик, 1917–1967; энциклопедический справочник* (1967), a reference book containing detailed information for the period 1917–67; M. LEBEDEVA and A. SMIRNOVA (eds.), *Культура, наука, искусство СССР; словарь-справочник* (1965), a reference work on cultural institutions; WALTER Z. LAQUEUR and GEORGE LICHTHEIM (eds.), *The Soviet Cultural Scene, 1956–1957* (1958); ARTHUR VOYCE, *Moscow and the Roots of Russian Culture* (1964); BERNHARD J. STERN and SAMUEL SMITH (eds.), *Understanding the Russians: A Study of Soviet Life and Culture* (1947, reprinted 1971); THOMAS P. WHITNEY (ed. and trans.), *The New Writing in Russia* (1964).

Space Exploration

Space exploration may be defined as the investigation, by means of spacecraft, of all the reaches of the universe beyond the atmosphere of the Earth. Spacecraft, vehicles that operate above the Earth's atmosphere, include sounding rockets, Earth satellites, and lunar, planetary, and deep space probes.

Covered in this article are the basic elements of space flight, space programs, both manned and unmanned, space launch vehicles, the kinds of knowledge gained from space exploration, and proposed future missions (see also ROCKETS AND MISSILE SYSTEMS and LIFE-SUPPORT SYSTEMS).

On October 4, 1957, the Soviet Union launched the world's first artificial satellite, Sputnik 1, and set in motion a series of programs of space exploration by the United States and the Soviet Union. The first U.S. satellite, Explorer 1, was launched on January 31, 1958, not quite four months after Sputnik 1. Both nations participated during the next two decades in a space race, with nearly 1,700 successful launches of satellites and space probes of all varieties: scientific research, communications, meteorological, photographic reconnaissance, and navigation satellites, lunar and planetary probes, and manned space flights. The Soviet Union launched the first man into orbit around the Earth on April 12, 1961. On July 20, 1969, the U.S. landed two men on the surface of the Moon.

The cost

The cost of such programs was high. During the first 20 years of the space age, the United States spent more than $90,000,000,000 on its combined civilian and military space programs. More than half a million people have been employed directly in U.S. space endeavours. Corresponding data are not available from the Soviet Union, but from the physical evidence it is judged that Soviet expenditures and resources committed have been comparable, if not greater.

Many of the spacecraft, such as manned and reconnaissance vehicles, are designed for recovery. Most operational satellites become inert after a few months or years of operation. The North American Air Defense Command (Norad) keeps a constant watch on the thousands of objects of human origin circling the Earth in a variety of orbital paths. Both radar and optical telescopes are used. In addition to satellite payloads (active or inert) is a much larger number of objects classified as debris—spent upper stages of launch vehicles, bolts, tether cables, etc., that go into orbit along with functioning satellites, as well as fragments that result from in-space explosions. Eventually low-altitude debris re-enters the Earth's atmosphere and usually burns up. By mid-1977 Norad had tracked the decay of 5,700 payloads and debris. At that time there were still 888 payloads and 3,359 items of debris catalogued and being tracked in orbit. In size they ranged from the U.S. Skylab space station to bits of debris as small as a golf ball.

It is not yet possible to judge the value of the new knowledge of the Earth, planets, and solar system that has been gained through space endeavours. A wealth of new understanding and of new scientific puzzles has resulted. Many questions concerning the nature and origin of the cosmos have been answered, and many more have been raised. Until 1957, human beings were passive observers of the solar system. Now they have entered the realm of space and gathered firsthand knowledge of their planetary environment. Twelve men have been landed on the Moon, set up data-gathering experiments on the lunar surface, and brought back lunar rocks and soil for study and analysis. The hidden side of the Moon and the planets Mercury, Venus, Mars, and Jupiter have been photographed and studied, and probes have been sent on six-year journeys to Saturn. From the vantage point of Earth orbit, the Earth and its weather have been studied and communications relays and navigation aids have been established. Space stations have been placed in orbit with crews to observe and perform experiments for nearly three months.

Doing these things required new materials and a major expansion of technology and manufacturing techniques. Reliability goals far greater than ever before were set and exceeded. The functional reliability demanded in launch vehicles having 6,000,000 component parts—a reliability greater than 99.9999+—had never before been required. Whole new industries based on these new technologies have become established; *e.g.*, high-speed digital computers and inertial guidance systems. A new standard of technical ability has resulted from space exploration and space utilization. The sociological and long-term effects of space technology are difficult to assess from a limited perspective. Historical analogy suggests that the effects may well be as far-reaching as those of the Industrial Revolution.

Functional reliability of craft

The article is outlined as follows:

I. Prehistory to Sputnik 1
II. Elements of space flight
  The environment of space
  Basic considerations in spacecraft design
  Launching into space
  Flight trajectories
  Navigation, docking, recovery
III. Space programs
  Space launch vehicles
  Unmanned programs
  Manned programs
  Applications satellites
IV. Contributions to the physical sciences
V. Proposed future missions

I. Prehistory to Sputnik 1

The dream of flight into space is as old as astronomy. Once people learned that the lights in the night sky were actual bodies, a long way off, they felt an urge to travel to them. Tales of fanciful flight to the Earth's nearest celestial neighbour, the Moon, may be traced back to the 2nd century AD, when the Greek rhetorician Lucian wrote a satirical account of such a fantastic journey.

But hundreds of years were to pass before the motions of the planets about the Sun and the immense distances to be travelled were appreciated. Awareness of the universe evolved slowly until tools of observation were developed. In the 17th century Galileo's use of the telescope to study the Moon added greatly to contemporary knowledge. Within a few weeks he had mapped the major visible mountains and valleys of the Moon and concluded that it was a solid world. Turning his telescope to the planet Jupiter, he discovered four tiny specks that passed slowly across its face. Galileo deduced that these moving dots were moons of the planet. He postulated further that their small apparent size in comparison to that of the Earth's Moon was explained by the great distance to Jupiter. Thus for the first time a scale of magnitude was established. Further observation with the telescope confirmed the pattern of the solar system described by Copernicus a hundred years earlier. Johannes Kepler calculated the elliptical orbits of the planets. Later in the 17th century, Isaac Newton formulated his "laws of motion," which at last placed physics and astronomy on a firm theoretical foundation. Not until nearly a hundred years later were the first free balloon ascents made and the journey upward and away from Earth began (see also FLIGHT, HISTORY OF).

Fictional accounts of flight to the Moon appeared sporadically in the 17th and 18th centuries but were mostly based upon fantasy rather than upon any extrapolation of technology. In 1865 Jules Verne published *From the Earth to the Moon,* which made use of many still correct assumptions concerning space flight. A few years later, Edward Everett Hale's "Brick Moon" appeared, an account of an artificial satellite that was also designed to be used by seamen for navigational purposes. By the turn of the century the science fiction writings of H.G. Wells (*The First Men in the Moon* [1901] and *The War of the Worlds* [1898]) attracted much attention. The engineering achievements of the 19th and early 20th centuries led many to believe that flight to the Moon could be possible.

Fictional accounts of space flight

Three brilliant men, destined to be called rocket pioneers, were the first scientists to conceive pragmatically of space flight: the Russian Konstantin Eduardovich

Tsiolkovsky, the American Robert Hutchings Goddard, and the German Hermann Oberth.

Technology in the early 20th century was a long way from the level required for rocket-powered flight, but the theory and dynamics of such flights had been rigorously studied. By the end of World War II, the German development of rocket propulsion for aircraft and guided missiles (notably the V-2) had reached a high level.

Thus, with the German surrender in 1945, the United States as well as its Allies—Great Britain, France, and the Soviet Union—fell heir to the technical knowledge of rocket power developed by the Germans. The technical director of the German missile effort, Wernher von Braun, and some 150 of his top aides surrendered to U.S. troops. Most emigrated to the U.S., where they assembled and launched V-2 missiles that had been captured and shipped to the U.S. The Soviet Union carried out an unpublicized but extensive program that must have been very similar; Great Britain and France conducted smaller programs.

Missile technology and satellites

In both the United States and the Soviet Union the development of military missile technology was essential to the achievement of satellite flight. Preparations for the International Geophysical Year (an 18-month period devoted to a worldwide program of geophysical research) stimulated discussion of the possibility of launching artificial Earth satellites for scientific investigations. As a result, the planning committee for the IGY in 1954 passed a formal resolution calling attention to the desirability of using artificial satellites in the IGY program. The IGY extended from July 1957 through December 1958. Both the United States and the Soviet Union responded with announcements that they would prepare scientific satellites for launching during the IGY. While the United States was still developing a satellite launch vehicle the Soviet Union startled the world by placing Sputnik 1 in orbit.

The launching of Sputnik 1, followed a month later by Sputnik 2 carrying a live dog, caught millions by surprise. The failure by the U.S. to launch its small (two kilograms) payload on December 6, 1957, heightened that nation's political discomfiture in view of its supposed advanced status in science. Following debates in Congress on the necessity to achieve parity, the U.S. established the National Aeronautics and Space Administration (NASA) in October 1958. NASA conducts the U.S. national space program, making public all plans and scientific data obtained, with the exception of the military space program. The Soviet Union space program is, with few exceptions, clothed in secrecy.

II. Elements of space flight

THE ENVIRONMENT OF SPACE

Space, as considered here, may be defined as all the reaches of the universe beyond the atmosphere of the Earth. There is no definitive boundary of the atmosphere of the Earth. For convenience it may be considered to extend to an altitude of 100 miles (160 kilometres) above the Earth's surface (see also ATMOSPHERE). Such a distance is infinitesimal in comparison with the immensity of the universe. Even within the solar system, distances between planets are measured in tens of millions of miles. The distance to Pluto, the Sun's outermost planet, is more than 3,000,000,000 miles. While the planets and their moons all travel at different speeds about the Sun, the solar system itself is travelling through our galaxy, the Milky Way. The Earth's nearest neighbouring stars lie more than four light-years (about 25,000,000,000,000 miles) away.

Contents of space

This apparent unimaginable void of space, however, is not empty. Throughout these vast reaches, matter (largely hydrogen) is scattered at the extremely low density of perhaps 100 particles per cubic centimetre in interplanetary space and 10 particles per cubic centimetre in interstellar space. This is, however, a much greater vacuum than has been achieved on Earth. Additionally, space is permeated by gravitational fields and a wide spectrum of electromagnetic radiation, by cosmic rays and magnetic fields of unknown intensities and distributions. Until 1946, all deductions about space had been made from observations through the distorting atmosphere of the Earth. With the advent of high-altitude sounding rockets after World War II and the instrumented satellites and probes of the space age, it has been possible to discover at first hand the great complexities of space phenomena.

BASIC CONSIDERATIONS IN SPACECRAFT DESIGN

Types of spacecraft

Spacecraft is a general term that includes sounding rockets, artificial satellites, and space probes. They are considered separately from the rocket-powered space launch vehicle, which gives satellite or escape velocity to the craft. A space probe is a spacecraft that is launched at higher than Earth orbital velocity and escapes the Earth's gravitational attraction.

Space probes may be classed as lunar, planetary, or deep space. Other classifications of spacecraft are manned or unmanned, active or passive. A passive satellite transmits no radio signals. It may be tracked optically or with radar, and radio communications signals may be "bounced" off its surface. Active satellites send out radio signals to make tracking easier and to transmit data from their instruments to ground stations or other craft.

One other general differentiation of satellites is by function: scientific or applications. A scientific satellite carries instruments to obtain scientific data on magnetic fields, space radiation, the Sun or other stars, etc. Applications satellites have utilitarian tasks; examples are Earth survey, communications, and navigation satellites.

Spacecraft thus differ greatly in size, shape, complexity, and purpose. Since some 2,500 successful launchings of satellites and space probes were made between 1957 and 1978, it is convenient to group them into program families; for example, the Soviet Sputnik, Vostok, Molniya, Venera, etc., and the U.S. Explorer, Mercury, Intelsat, Apollo, Mariner, etc.

Lightness of weight and functional reliability are primary features of spacecraft design. Depending upon their mission, spacecraft may spend minutes, days, months, or years in the environment of space. Mission functions must be performed while exposed to high vacuum, extreme variations in temperature, and radiation.

There are nine general categories of subsystems found on most spacecraft: power supply, on-board propulsion, communication, attitude control, environmental control, guidance, computer, structure, and engineering instrument.

Power-supply subsystem. The power supply furnishes energy to all other subsystems. All active satellites need electrical power. It may be provided by batteries or by solar cells, fuel cells, or nuclear isotope devices. The amount of power supply required varies widely.

On-board propulsion subsystems. Many spacecraft missions require a source of rocket power to change spacecraft velocity for a variety of reasons. The rocket motors may use liquid, storable bipropellants, a monopropellant, or a solid propellant. Such on-board propulsion systems are used to change satellite orbits, to reduce velocity (retrorocket) for re-entry, or to slow down a spacecraft.

Communication subsystem. Radio communications link spacecraft with receivers on the Earth. Spacecraft transmitting and receiving antennas must be precisely pointed toward the Earth-based antenna "dishes" for maximum signal strength. Very high frequencies are used for spacecraft communication to penetrate the Earth's ionosphere. Frequencies in the range of 100 to 3,000 megahertz (millions of cycles per second) are common. Spacecraft do not transmit continuously, although tracking beacons often do to ensure against loss of contact, particularly in deep space (planetary) probes.

Attitude-control subsystem. There are two main functions of a spacecraft's attitude-control subsystem: stabilization and pointing. Stabilization is the act of maintaining a spacecraft oriented in a specific direction despite the contrary effects of various natural forces. "Pointing" refers to a requirement to point precisely at selected targets. Balanced in a weightless condition in satellite orbit or in free-fall on an Earth-escape trajectory, a spacecraft is subject to rotation about any, or all, of its three axes

(pitch, yaw, and roll). Various stabilization techniques have been studied and tried. The pressure of sunlight itself can cause a spacecraft to twist into undesired orientation. The pressure of sunlight acting on paddles has been used for attitude control on planetary space probes.

Pointing requires most satellites and space probes to aim their sensors at a target, be it the Sun, portions of the Earth, a planet, or a star. Similarly, the transmitter must be aimed at the Earth during transmission.

Environment-control subsystem. Anticipating and understanding the environmental conditions of space in a planned mission, a spacecraft designer provides for control of these conditions within desired limitations. The high vacuum, extreme temperatures of the Sun, ultraviolet, cosmic, and X-radiation, and other space conditions can affect instrumentation and communications equipment. Lest one side of a sunlit spacecraft be always in shadow, rotation may be employed to distribute solar temperature effects evenly. On the Moon, surface experiments must survive the very low temperature during the two-week-long lunar nights. Heating is sometimes provided and reflective insulation used. The interior of a spacecraft may become hot because of electric power thermal effects. Some spacecraft have louvres that open on the shaded side of the spacecraft to cool the interior.

To be habitable, manned spacecraft require a reasonable approximation of Earthly conditions. Fresh oxygen to breathe is necessary. Air-conditioning systems circulate the weightless air, remove toxic substances, and control temperature and pressure (see LIFE-SUPPORT SYSTEMS).

Guidance and control subsystems. The function of the guidance and control subsystem is to steer a spacecraft to its target and control its velocity. In the case of a spacecraft whose mission is to land softly on the Moon, on approach trajectory the attitude-control subsystem will rotate the spacecraft so that its retrorocket and landing radar are pointed toward the lunar surface; radar signals will tell the spacecraft guidance system the rate of descent and the distance to the surface, and this information will be compared continuously with on-board data that tell the spacecraft when to fire its retrorockets, which fire until radar data indicate that touchdown will occur at near-zero velocity and then shut off. Concurrently, the attitude control rocket thrusters will maintain precisely vertical trajectory to the surface.

Use of radar

Computer subsystem. Early satellites transmitted orientation data to the ground control, where digital computers calculated corrective changes. Increasingly, miniature computers are placed in the spacecraft.

Structure subsystem. The structure subsystem is the skeleton framework of the spacecraft, which physically supports and protects all other subsystems. Many spacecraft have a central core and deployable appendages such as antennas, solar cell paddles, and gravity-stabilization booms. The structure is usually made of aluminum or magnesium for lightness. Spacecraft shape is always symmetrical about at least one axis, to prevent wobbling while rotating; but since most spacecraft are not designed for recovery but operate in space only, their appearance is often unusual and angular, without concern for aerodynamic streamlining. Structural stiffness is required to withstand the stresses of launch accelerations. Thermal protective shields are provided for spacecraft that reenter the Earth's atmosphere and are recovered.

Engineering instrument subsystem. This subsystem is composed of all the "housekeeping" sensors. Included are voltmeters, ammeters, thermometers, and other instruments whose measurements provide data on the status of the spacecraft. In a manned spacecraft these status sensors telemeter to Earth as well as to display panels for monitoring by the astronauts. Also, the astronauts themselves are monitored by bio-instruments to keep watch on their physical well-being.

LAUNCHING INTO SPACE

Gravity. The gravitational attraction of the Earth was one of the major obstacles to space flight. Rocket pioneers understood the principles of space flight based upon the observations and calculations of Copernicus, Galileo, and Kepler, and on Newton's universal laws of motion; but the use of these principles had to await the development of rocket power to project a spacecraft to the required altitude above the Earth and to give it the velocity necessary to accomplish its mission.

The spacecraft at the forward end of a multistaged launch vehicle is projected upward by reaction to the high-speed jet of combustion gases produced in the rocket motor. If the lifting force, or thrust, of a rocket-powered launch vehicle is twice the weight of the vehicle at lift-off, then the spacecraft assembly will rise at an initial acceleration of one *g*, or 32.2 feet (9.8 metres) per second, each second. Because the propellant mass is being consumed and ejected from the rocket motor, the vehicle is continually lightened and acceleration increases.

The gravitational pull of the Earth upon the upward (and outward) travelling spacecraft subsides slowly. At an altitude of 100 miles (160 kilometres), the gravitational attraction is just 1 percent less than at the Earth's surface. At 1,700 miles (2,700 kilometres) it is halved (16.1 feet per second per second), and at 60,000 miles (97,000 kilometres) it is one-twentieth. For the purposes of space flight, the gravitational pull of the Earth becomes negligible only at distances of several million miles, except when a spacecraft approaches the Moon and lunar gravity (one-sixth that of Earth) exerts effect.

Decrease of gravitational pull

Staging. This term is used to describe a technique by which several rocket propulsion systems are mounted one on top of the other. The lowest, or first, stage ignites and lifts the vehicle at increasing velocity until its propellants are exhausted. At that point the stage drops off and a second stage is ignited. In this technique the empty propellant tankage and vehicle structure are cast off, thus lightening the weight of the launch vehicle. The second stage of lower thrust then begins to accelerate the launch vehicle and spacecraft, commencing at the velocity reached under first-stage power. Most space launch vehicles have three stages. Tsiolkovsky, Goddard, and Oberth all investigated this technique independently.

Acceleration rates. In general, the longer it takes a space vehicle to leave the Earth's atmosphere and achieve required velocity, the less economical the procedure becomes. At low accelerations the launch vehicle wastes great quantities of rocket propellant because it is losing, in effect, about 32 feet per second of velocity each second of travel. An upper limit of acceleration is governed by the maximums of accelerative stress permissible upon the vehicle structure and the spacecraft payload. In manned flight a gravity pull of six *g* is considered the maximum tolerable when the human body is positioned transversely with respect to the acceleration force (*g*-loading); *i.e.*, with the head and heart at the same level.

FLIGHT TRAJECTORIES

There are four general types of trajectories: sounding rocket, Earth orbit, Earth escape, and planetary.

Sounding rockets. The sounding rockets first launched for upper atmosphere studies in 1945 were fired nearly vertically. Generally single-staged, they reached speeds of 3,000–5,000 miles (4,800–8,000 kilometres) per hour. Burnout, the completion of rocket motor firing, usually occurred at altitudes of 10 to 20 miles (16 to 32 kilometres), at which point the rocket coasted upward, slowly losing speed because of gravity. Velocity upward dropped to zero at peak altitude, and the rocket began its fall, picking up speed until it crashed into desert or ocean. Instrumentation carried aboard for data gathering was sometimes ejected to descend by parachute, but increasingly the data were radioed (telemetered) to a tracking station at the launch site. Altitudes of 100 miles were reached with single-stage rockets.

Earth orbit. Earth orbital flight is achieved by launching vertically and then tilting the trajectory so that flight is parallel to the Earth's surface at the time that orbital velocity at the desired altitude is reached. At this precise point, the rocket engine is cut off. A spacecraft attached to the final-stage rocket is then in free fall about the Earth, the centrifugal pull on the spacecraft being equal to the Earth's pull of gravity. At an altitude of 125 miles

(200 kilometres), Earth orbital velocity is about 18,000 miles (29,000 kilometres) per hour. Since this 125-mile altitude is above most of the atmosphere, aerodynamic drag is not great, and the spacecraft will continue to orbit for an extended time.

The length of time required for the satellite to make one complete revolution is known as the period of orbit. At 125 miles this is about 90 minutes. At higher altitudes than this above the Earth, the velocity of a satellite decreases and the orbital period increases.

For example, at an altitude of 1,075 miles (1,730 kilometres) the orbital velocity is 15,800 miles (25,400 kilometres) per hour and the period is two hours. At 22,300 miles the velocity is 7,000 miles (11,300 kilometres) per hour and the period 24 hours. Since this particular period is equal to the time the Earth rotates once, such a satellite travels at the same angular velocity as the surface of the Earth and appears to be stationary in the sky. This particular orbit, called geostationary (or geosynchronous), has special value to communications and meteorological satellites. Finally, the Moon travels in an orbit of about 240,000 miles (386,000 kilometres). Lunar velocity about the Earth is about 2,300 miles (3,700 kilometres) per hour, with a period of about 28 days.

The geostationary orbit

The above applies only to circular orbit, often ideal, but difficult to achieve. Usually a satellite's orbit is an ellipse with a perigee altitude and an apogee altitude. Perigee and apogee are defined, respectively, as the points of the orbit

By courtesy of National Aeronautics and Space Administration

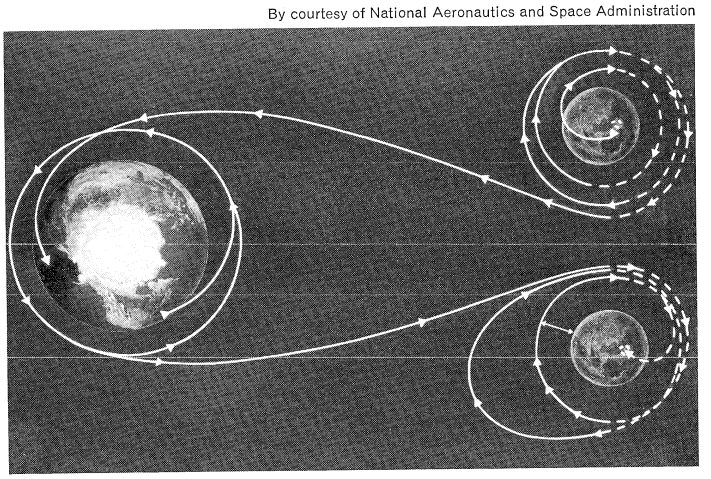


Figure 1: Apollo 11 lunar landing mission profile.

nearest and farthest from the body orbited. Orbits may be circularized if thrust is available by reducing the velocity at perigee (apogee is lowered) or by increasing the velocity at apogee (perigee is raised). Rocket power in such instances is applied along the axis of flight paths.

In projecting a satellite into Earth orbit, the launch vehicle is tilted after liftoff in an easterly direction. Launching to the east is done to take advantage of the Earth's eastward surface velocity. This rotational surface velocity is about 1,500 feet (450 metres) per second at the equator and 1,370 feet (400 metres) per second at the latitude of Cape Canaveral. It would be perfectly possible to launch a satellite on a westerly orbit, but an additional 2,000 feet (600 metres) per second would be required for an orbit of the same altitude compared with an easterly orbit.

If the satellite is launched in a northerly or southerly direction, a polar orbit is obtained. The easterly surface velocity launch advantage is lost, but there are other advantages. As the Earth turns on its polar axis, the satellite travels over all parts of the globe every few revolutions. This ground track, the path around the Earth directly under the satellite, thus varies according to the orbit, which is chosen according to the desired characteristics of a particular mission. Ground-tracking and data-receiving stations must be on or near the ground track.

Earth escape. In order to escape completely from the Earth's gravity, a spacecraft launch velocity of about 25,000 miles (40,000 kilometres) per hour is required. At that speed the spacecraft is able to reach distances beyond the Earth's gravitational field. If the spacecraft does not come under the gravitational influence of other celestial bodies, it will go into an orbit about the Sun like a tiny planetoid. With precise timing it is possible for the spacecraft to enter a trajectory that will carry it near the Moon. In the case of eight Apollo flights, the spacecraft was placed in a trajectory calculated to pass ahead of the Moon and, under the influence of lunar gravity, to swing around the far side. If no manoeuvre was made the spacecraft would loop around the Moon and return on a trajectory toward the Earth. By reducing flight speed on the far side of the Moon, the spacecraft was placed in lunar orbit held by lunar gravity.

Consideration of the actual dynamics of the Apollo trajectory will illuminate the statements made above (see Figure 1). The Apollo 11 spacecraft had been in an Earth orbit of 115–117 miles altitude, travelling at 17,427 miles per hour. By firing the rocket motor at the exact moment when the spacecraft was precisely aligned along the proper trajectory, the velocity was increased to 24,200 miles per hour. Because the Earth's gravitational pull continued to act upon the spacecraft during its 64-hour journey toward the Moon, the spacecraft velocity, with respect to the Earth, dwindled to 2,040 miles per hour at a distance of 39,000 miles from the Moon. At this point lunar gravitational attraction became greater than the Earth's, and the spacecraft began accelerating as it swung toward and around the far side of the Moon, reaching a speed of 5,600 miles per hour. By firing the rocket propulsion system, the velocity was reduced to 3,680 miles per hour and the spacecraft entered an elliptical orbit about the Moon. This orbit was later circularized preparatory to the actual lunar landing operations.

Apollo 11

The so-called three-body problem of mechanics (the relative motions of the Earth, the spacecraft, and the Moon) is extremely complex (see MECHANICS, CELESTIAL). Although equations expressing the relative motions can be written, no general solution was possible before the development of the high-speed digital computer, for ballistics calculations for long-range missile trajectories. The computer integrates the complicated equations of motion numerically upon command, provides direct readout of the complete translational motions of the spacecraft, and compares the actual to the preplanned flight path at any point in time.

Planetary flights. Because of the elliptical nature of the planetary orbits, distances to Earth's nearest neighbouring planets, Venus and Mars, vary. A so-called favourable launch opportunity occurs about every two years. Flights can be made at other times, but the velocity required is greater and length of time is longer, or, for the same launch vehicle, a lighter payload is necessary. Figures 2 and 3 show flight paths to Venus and Mars, respectively.

The trajectory from the Earth to Venus or Mars may be planned to take advantage of the changing orbital relationships of the planets for the most economical flight. Such advantageous paths were calculated in the late 1920s. Although these transfer trajectories require the least velocity, they are of long duration, as long as 260 days to Mars. Thus a compromise trajectory is used, as in the case of Mariners 6 and 7 in 1969. Launched on February 25, 1969, Mariner 6 passed within 2,131 miles of the planet 157 days later, when Mars was 58,000,000 miles distant from the Earth.

Changing orbital relationships of planets

NAVIGATION, DOCKING, RECOVERY

Navigation. Travelling from point A to point B in space is almost never in a straight line or at constant velocity because of the many influences on the body in motion.

The basis for space navigation is inertial guidance; that is, guidance based on the inertia of a spinning gyroscope, irrespective of any external forces and without reference to the Sun or stars. By the use of three gyroscopes and accelerometers, precise measurements may be made of any change of velocity (acceleration), either positive or

negative, along any or all of the three axes. By use of a memory system in the computer it is possible to determine where the vehicle is at any time along a trajectory.

Trajectory correction

By changing attitude (rotation about one or more axes) and firing a small jet or rocket motor, corrections may be made to the trajectory. Preprogrammed digital computers, both on the ground and in larger spacecraft, automatically or on command "keep an eye on" where the spacecraft is, where it was, and where it is supposed to be going. Direct readouts of such data from the computers are displayed to the crew in manned spacecraft.

During the launch phase, corrections to deviations in planned flight path are usually made at once by special small thrust motors, by deflection of the rocket exhaust jet, or by swinging one or more of the rocket engines in a gimbal mount.

In the case of rendezvous between two spacecraft, radar sightings fed to the computer tell the crew the corrections required along each of the three axes in feet per second (plus or minus). In the case of a lunar mission, deviations encountered are stored in a computer memory system until requested. One or more midcourse manoeuvres may be made. During Apollo lunar flights an optical sighting on three stars by means of a sextant was used as a backup system and check on the spacecraft position.

The above description of navigation techniques is oversimplified; the systems used are complex, powerful tools without which manned or unmanned spacecraft would not be feasible.

Rendezvous and docking. Rendezvous is the action of one spacecraft approaching another spacecraft. Rendezvous in space operations refers to a match of orbital trajectories and the action of one spacecraft arriving in proximity to another. Ideally, the orbit planes should be the same. The term is relative but, generally speaking, rendezvous distance should be within 100 yards or metres. To achieve rendezvous, the launch of the second spacecraft is timed within a fraction of a second. Since the orbiting spacecraft already has a high velocity and the second spacecraft is at rest (zero velocity) before launch, the second spacecraft is launched well before the first spacecraft passes overhead. Aim is for a co-planar orbit just below the first spacecraft. When this is done, the first spacecraft is travelling at a slower speed, being at a higher orbit; thus the second spacecraft will overtake the first. When slightly ahead of the first spacecraft, application of rocket thrust causes the second spacecraft to rise in orbit and thus to slow down, matching the first spacecraft's orbital altitude and velocity. Radar systems and on-board computers are necessary for such operations.

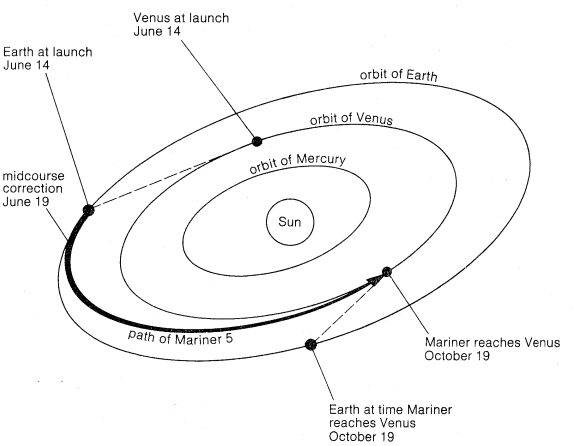


Figure 2: Flight path of Mariner 5 from Earth to Venus, 1967.

First spacecraft rendezvous

Gemini 6 and 7 were the first spacecraft to rendezvous with each other. In Apollo lunar landing flights the ascent stage of the Lunar Module rose from the surface of the Moon to rendezvous and dock with the orbiting Command Module.

Docking, the term for meeting and mating of two spacecraft, was an essential element of the Apollo lunar landing (lunar-orbital rendezvous) technique and is essential for many plans for future manned space exploration, such as a permanent Earth-orbiting space station. Significant in docking manoeuvres, from the starting point of rendezvous, are precise radar data (distance and closing velocity rate), attitude control, and computer subsystems. Whereas the United States has relied upon on-board crew guidance for close rendezvous and docking, the Soviet Union has demonstrated the ability to perform this task automatically (1967).

By courtesy of the National Aeronautics and Space Administration

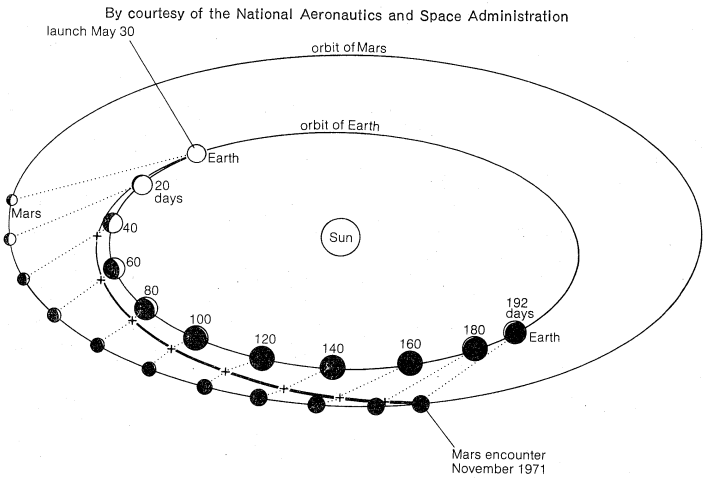


Figure 3: Flight path of Mariner 9 from Earth to Mars, 1971.

The concept of building a large space station out of component parts brought to orbit in successive stages or, similarly, assembling a deep space mission as for a manned trip to Mars, requires a reliable rendezvous and docking technique. The entire vehicle and propellants cannot be launched into Earth orbit at one time because of limitations on space launch vehicle payload. Further, rotation of space crews and emergency rescue missions require rendezvous and docking capability.

Re-entry and recovery. Re-entry means the return of a spacecraft into the Earth's atmosphere. This blanket of relatively dense gas surrounding Earth is useful as a braking, or retarding, force resulting from aerodynamic drag. A concomitant effect, however, is rapid and severe frictional heating caused by rapid flow of nitrogen and oxygen molecules along the blunt forward profile of the spacecraft. By use of an ablative heat shield, specially formulated plastic materials melt, vaporize, and carry away boundary layer heat. Insulation behind the heat shield keeps the spacecraft interior at a constant tolerable temperature.

Angle of re-entry

Inherent in safe re-entry of a spacecraft is precise control of the angle of re-entry. This angle of trajectory with respect to the Earth's horizon is —6.2° and is held within limits of ±1°. A returning Apollo Command Module approached the Earth at nearly 25,000 miles (40,000 kilometres) per hour. If the angle is too shallow, the spacecraft will skip or bounce off the atmosphere out into space probe trajectory. If the angle of re-entry is too great, the heat shield will not survive the extreme heating rates nor the crew and spacecraft the high *g*-forces. Even so, the heat shield of the Apollo Command Module was subjected to temperatures approaching 5,000° F (3,000° C).

The Apollo spacecraft under atmospheric drag slowed to the speed of sound (Mach 1) at 30,000 feet (9,000 metres). Shortly thereafter, three closed parachutes were deployed, then opened fully at about 10,000 feet (3,000 metres). The Command Module struck the ocean at about 22 miles (35 kilometres) per hour. Recovery helicopters and surface ships reached the spacecraft usually within minutes.

The Soviet Union recovers manned spacecraft on the plains of Siberia. Soviet cosmonauts, at least on early flights, could elect to parachute separately from the spacecraft in final stages of descent.

The United States Air Force recovers small spacecraft by aircraft while the spacecraft is still descending by parachute. This air-snatch technique eliminates the problem of search and recovery of spacecraft after landing and minimizes the time before study can be made of data obtained.

III. Space programs

Considering the two categories of space programs, manned and unmanned, a few generalizations may be made. Spacecraft without a human being aboard have invariably pioneered explorations. They are smaller, can operate for months or years, and offer no hazard to human life. Experiments and measurements, however, are limited by the need for preplanning. In manned flights, the range of experiments is greater because judgment can be exercised in observations, instrumentation can be adjusted, and, perhaps, repairs can be made and equipment maintained.

Manned space flight is much more expensive because of the added weight of vehicle and equipment required to provide a habitable atmosphere and controls. This extra weight requires a much larger space launch vehicle. Often, redundant or backup systems are provided in manned spacecraft to provide for high reliability and safety of personnel.

Thus, both categories of spacecraft have advantages and disadvantages.

SPACE LAUNCH VEHICLES

Although sounding rockets may reach altitudes above the atmosphere of the Earth, the term space launch vehicles is applied usually to those rocket boosters designed to place satellites in orbit or to impart Earth-escape velocity to spacecraft.

About 1950 the technology of rocket propulsion had reached a level at which consideration of a project to launch an artificial Earth satellite became feasible. Although space travel—flight to the Moon and planets—had been dreamed about for centuries, it was only now that the critical element, a rocket-powered space launch vehicle, could be constructed. Worldwide scientific studies during the International Geophysical Year (IGY), 1957–58, provided the basis for funding. In 1955 both the United States and the Soviet Union announced satellite programs as part of their national effort in the IGY.

First satellite programs

When Sputnik 1 and 2 were launched in 1957 the Soviet Union released no details of their launch vehicles. In May 1958 Sputnik 3, weighing nearly 3,000 pounds (1,360 kilograms), was launched.

It was not until 1967 that the basic Soviet launch vehicle was displayed. It was a 2½-stage vehicle: two stages with four drop-away booster pods. Each booster pod contained four rocket engines (totalling 16) with propellant tankage, and the central core had four engines. Propellants were conventional liquid oxygen and kerosine. Lift-off thrust of this vehicle was estimated at about 1,000,000 pounds (450,000 kilograms); it was a conversion of the Soviet intercontinental ballistic missile Sapwood.

The United States launched its early small satellites with two different vehicles, the Jupiter-C and Vanguard. Jupiter-C was a modified Redstone liquid-propellant ballistic weapon of medium range to which were added additional tankage length and three upper stages of clustered solid-propellant rockets. The modification was originally designed to achieve a velocity of 20,000 feet (6 kilometres) per second to test a nose cone (re-entry vehicle). The desired velocity was obtained with two upper stages, one a cluster of four solid-propellant rockets and the other a single rocket. Each rocket developed 1,800 pounds (800 kilograms) of thrust for 6.5 seconds. It was obvious that by increasing the final velocity 5,000 feet per second to the required 25,000 feet per second, Earth satellite velocity could be obtained for a small scientific payload. The additional velocity was obtained by adding another stage with a cluster of solid-propellant rockets so that the upper stages consisted of 11, three, and finally one rocket carrying a payload weighing 18.1 pounds (8.2 kilograms). Late in 1954 the Army Ballistic Missile Agency and the Office of Naval Research jointly proposed this scheme, known as Project Orbiter, but a newly designed Vanguard launch vehicle was selected. Failures in early attempts to launch Vanguard, however, resulted (two and one-half years later) in eventual approval of the Project Orbiter approach. Thus the first U.S. satellite was launched by a Jupiter-C on January 31, 1958.

The Vanguard launch vehicle was a three-stage booster approximately equal in length (about 72 feet) to the Jupiter-C but much lighter in takeoff weight (22,600 pounds compared to 64,000 pounds). Vanguard launched its first satellite (3 pounds) into high orbit on March 17, 1958. After a few more flights, the Jupiter-C was retired in 1958 and the Vanguard in 1959.

Vanguard

During the 1960s the United States developed a series of standard launch vehicles. The Air Force modified a Titan II intercontinental ballistic missile (ICBM) for space launch purposes by strapping two solid-propellant booster rockets, 10 feet in diameter, to the liquid-propellant core vehicle. The Titan IIIC is used for large military satellites. Then NASA increased performance of the obsolete Thor intermediate-range ballistic missile (IRBM) by adding solid-propellant boosters. A liquid oxygen/liquid hydrogen upper stage, Centaur, is used on obsolete Atlas ICBM's and Titan III ICBM's to launch large planetary space probes.

The Saturn series of NASA launch vehicles was developed specifically for the Apollo lunar mission program. The two operational Saturn models were the two-stage Saturn IB and three-stage Saturn V. The Saturn IB was used for Earth orbital developmental missions of Apollo, while the Saturn V, largest of all known space boosters, was used for lunar missions. Saturn V stood 363 feet (110.6 metres) high and weighed more than 6,000,000 pounds (2,700,000 kilograms) at launch. It was capable of placing 230,000 pounds in orbit and sending 95,000 pounds (4,300 kilograms) to escape velocity.

At the conclusion of the Apollo program in 1972, excess Apollo spacecraft and Saturn launch vehicles were used for the Skylab missions and the Apollo–Soyuz Test Project. The development of these great space boosters with the safety and reliability standards required for launch of manned missions to the Moon was a major challenge to the aerospace industry.

In the Soviet Union all space launch vehicle developments, as well as launch operations, are controlled by the military forces. Knowledge of details of design and performance of all vehicles is limited.

For some years the launching of spacecraft was limited to the U.S. and Soviet Union. The reason was that the rocket-powered launch vehicles were based upon long-range ballistic missiles, which only these countries had developed. France was the third nation to launch a satellite (1965), followed by Japan (1970), the People's Republic of China (1970), and the United Kingdom (1971). Both the United States and the Soviet Union have launched satellites or carried scientific payloads for other nations. The United States regularly launches communications satellites for U.S. industrial organizations and other nations, the cost being reimbursed. Eleven of a total of 15 NASA launches in 1977, for example, were for other nations, the North Atlantic Treaty Organization (NATO), the International Telecommunications Satellite Organization (Intelsat), and the European Space Agency (ESA).

Cooperative launchings

UNMANNED PROGRAMS

The unmanned space exploration program includes sounding rockets, scientific satellites and probes, and lunar and planetary programs (Table 2).

Sounding rockets. At the conclusion of World War II the Allies fell heir to the great advances in rocket technology made by the Germans in connection with guided missile development (see also ROCKETS AND MISSILE SYSTEMS).

The first program to use sounding rockets in exploration of the upper atmosphere was carried out from 1946 to 1951, when the United States Army fired about 60 V-2s in

near-vertical trajectories in New Mexico. The one-ton warhead was replaced by a payload of scientific instruments. The highest altitude obtained in these firings was 133 miles (214 kilometres).

The Aerobee and Viking sounding rocket programs were similar projects conducted at White Sands by the United States Navy, reaching altitudes of about 200 miles (320 kilometres).

During the decade following the war, rocket engine technology, supersonic design, flight controls, tracking, instrumentation, and other essential elements of sounding rockets were under development while high-altitude data were being obtained. Typical instrumentation carried aboard Viking rockets included Geiger (particle) counters, ionization chambers, photomultiplier tubes, and cameras.

The development of the Nike series of surface-to-air guided missiles provided a new family of relatively low-cost sounding rockets. With the exception of those sounding rockets based on the liquid-propellant Aerobee design, all U.S. sounding rockets by 1971 used solid propellants. During the 18-month International Geophysical Year (1957–58), there were many cooperative programs around the world in which sounding rockets were fired for scientific purposes. The U.S. launched 210 rockets from widely separated locations, including the Arctic, California, and Guam in the Pacific Ocean.

IGY programs

Meanwhile the Soviet Union and many other countries developed and launched sounding rockets. The Soviet Union launched 125 rockets during the IGY from locations in the Arctic, Antarctica, and Central Asia. It also has a variety of sounding rockets called simply Geophysical Rockets. In 1957 one of these carried a large payload of 4,850 pounds (2,200 kilograms) to an altitude of 132 miles (212 kilometres). In the payload was a capsule containing two dogs, which were parachuted safely to Earth. Later in the IGY a large Soviet payload soared to 294 miles (473 kilometres). Measurements were made of the Earth's magnetic field, physical and chemical nature of the Earth's upper atmosphere, cosmic and solar radiation, X-rays and ultraviolet radiation, and micrometeoroids.

Other countries having sounding rocket programs include Argentina, Australia, Canada, France, Germany, Great Britain, India, Italy, Japan, Norway, Poland, and Sweden. Often these programs are cooperative.

Less glamorous than their larger cousins that launch satellites and space probes, sounding rockets are workhorses valuable to continuing scientific studies. Although the period of their measurements is brief, usually five to 10 minutes, they provide the only feasible way of taking such measurements between 25 and 100 miles (40 and 160 kilometres) altitude above the Earth—above the range of aircraft and balloons and below minimum continuous satellite orbital height. Additionally, by using miniaturized payloads of sophisticated design, the U.S. has employed sounding rockets to perform experiments in materials processing under brief weightless conditions and conducted observations of stars emitting gamma rays.

Scientific satellites and space probes. This category of programs includes those spacecraft whose purpose is to make measurements of natural phenomena in space environment. Generally, these measurements are of solar and other cosmic radiation, magnetic fields and observation of stars. Such data are important in developing and verifying theories of the evolution of the solar system and the fundamental natural processes and forces that affect Earth and people. These natural processes directly affect living conditions, agriculture, and economy. This is the underlying logic behind space research.

The United States has a number of programs in this category. The Orbiting Solar Observatory (OSO), first launched in 1962, was the first of a series of these observatories that point toward the Sun and record and transmit a variety of data such as the frequency and energy of solar electromagnetic radiation in ultraviolet, X-ray, and gamma ray regions of the spectrum. An OSO carries instruments for eight to 10 experiments. The spacecraft consists of two major parts, a spinning base and a fan-shaped panel of solar cells that points continuously at the Sun. Altitude above the Earth is about 350 miles (560 kilometres).

Orbiting Solar Observatory

A Pioneer series of small spacecraft operates in deep space in solar orbit, typically carrying out experiments involving measurement of magnetic fields, plasma analysis, and cosmic-ray and solar-wind studies. Data are stored on recording tape and played back rapidly upon command from Earth.

An Explorer family of spacecraft collects a widely variant series of data. Explorer 1, the first United States satellite, was launched by the U.S. Army. The program was taken over by NASA after its organization in 1958. Explorer satellites have made measurements of the magnetosphere, Van Allen radiation belts, micrometeoroids, energetic particles, cosmic radiation, solar wind, and ionosphere. Some Explorer satellites, called Interplanetary Monitoring Platforms, are placed in eccentric orbits to take measurements at maximum altitudes. Another subseries of Explorer satellites is the Small Astronomy

Table 1: Some Space Launch Vehicles

vehicle	stages	propellants	thrust (000 lb)	diameter (ft)	height (ft)	launch weight (000 lb)	Earth orbit payload (lb)	escape payload (lb)
U.S.								
Scout D	4*	solid, all stages	140; 60; 21; 5.6	3.7	73	47	400	85
Delta 2914	3	LOX†, kerosine +3, 6, or 9 strap-on solid‡; nitrogen tetroxide/hydrazine mixture	205+162–468; 10; 15	8	116	293	3,900	1,050
Atlas–Centaur	2½	LOX/kerosine; LOX/liquid hydrogen	503; 30	10	131	326	10,300	2,500
Titan IIIE–Centaur	4	2 strap-on solid§, nitrogen tetroxide/hydrazine mixture first and second stages; LOX/liquid hydrogen third stage	520+2, 400; 101; 30	30	160	1,400	38,000	11,300
Space Shuttle‖	2	2 strap-on solid; main engines, LOX/liquid hydrogen	booster 5,200; sustainer 1,250	78 (wingspan)	184	4,400	230,000	—
Soviet Union								
A-2 (Soyuz)	2½	LOX/kerosine; liquid	1,091; 66	34	167	720	16,550	3,640
D-1 (Proton)	3–4	liquid; upper stages unknown	3,000	15	150–200	—	37,600	12,000
Japan								
Mu-3C	3	all stages solid	440.7; 80.9; 16.5	4.6	67	93	150	—
N	3	LOX/kerosine; nitrogen tetroxide/hydrazine mixture; solid	170+356; 12; 8.7	8	106.9	198	2,200	300
European Space Agency								
L-III S Ariane¶	3	nitrogen tetroxide/hydrazine mixture, first two stages; LOX/liquid hydrogen	533; 157; 13.5	12.5	156	443	12,000	—
China								
Satellite launcher		—	—	—	—	—	500	—

*Five-stage version also. †LOX=liquid oxygen. ‡Strap-on solid-propellant units, 54,000 lb thrust each. §Strap-on solid-propellant units, 1,300,000 lb thrust each. ‖First launch planned 1979. ¶First launch planned 1978.

Table 2: Space Exploration by Selected Unmanned Artificial Satellites and Space Probes

name	nation	launched	launch vehicle	perigee–apogee (mi)	spacecraft weight (lb)	remarks
Sputnik 1	U.S.S.R.	Oct. 4, 1957	*	143.5–584	185	launching of first artificial satellite
Sputnik 2	U.S.S.R.	Nov. 3, 1957	*	128–1,056	1,120	first inhabited space capsule (dog, Laika)
Explorer 1	U.S.	Jan. 31, 1958	Jupiter-C	224–1,585	18	discovery of inner (Van Allen) radiation belt
Vanguard 1	U.S.	March 17, 1958	Vanguard	405–2,462	3	first proof that Earth is pear-shaped; first use of solar cell power
Sputnik 3	U.S.S.R.	May 15, 1958	*	133–1,178	2,926	first instrumented multipurpose space laboratory (atmospheric pressure, composition; positive ion concentration, primary cosmic ray data, etc.)
SCORE	U.S.	Dec. 18, 1958	Atlas	115–915	8,800	first transmission of voice messages from space
Luna 1	U.S.S.R.	Jan. 2, 1959	*	91.5–214.7	796	first space probe to achieve Earth-escape velocity; radio contact to 371,000 mi
Vanguard 2	U.S.	Feb. 17, 1959	Vanguard	347–2,064	22	first cloud-cover photos transmitted by television
Luna 2	U.S.S.R.	Sept. 12, 1959	*	(Moon impact)	860	first spacecraft on Moon
Luna 3	U.S.S.R.	Oct. 4, 1959	*	25,500–294,250	959	first circumnavigation of Moon; first pictures of far side of Moon
TIROS 1	U.S.	April 1, 1960	Thor-Able	428.7–465.9	280	first weather satellite; television view of cloud cover; 22,952 photos transmitted
Transit 1B	U.S.	April 13, 1960	Thor-Able-Star	430–468	265	first navigation satellite
Discoverer 13	U.S.	Aug. 10, 1960	Thor-Agena A	161–436	300	first successful recovery of capsule from orbit
Echo 1	U.S.	Aug. 12, 1960	Thor-Delta	945–1,049	132	passive communication satellite; first successful voice and picture transmission by reflection
Sputnik 5	U.S.S.R.	Aug. 19, 1960	*	190–211	10,143	recovery of orbited animals (two dogs, six mice)
Sputnik 8–Venera 1	U.S.S.R.	Feb. 12, 1961	*	123–198	1,420	first hyperbolic escape of probe to measure interplanetary magnetic fields and solar radiation in space and near Venus; first launch from Earth orbit
Midas 2	U.S.	May 22, 1961	Atlas-Agena 2	298–301	5,000	first missile-launch detection satellite
Transit 4A	U.S.	June 29, 1961	Thor-Able-Star	543–623	175	all-weather global navigation satellite; first spacecraft to use nuclear isotope power system
OSO 1	U.S.	March 7, 1962	Delta	344–370	458	first Orbiting Solar Observatory, transmitting solar flare, radiation data
Ariel 1	U.K.–U.S.	April 26, 1962	Delta	242–754	132	first international cooperative satellite (U.S.–U.K.); ionospheric, X-ray, and cosmic ray data
Telstar 1	U.S.	July 10, 1962	Delta	593.4–3,501.8	170	first transatlantic relay of television signals; first colour television relay
Mariner 2	U.S.	Aug. 27, 1962	Atlas-Agena B		448	Venus flyby (21,594 mi); measurements of atmospheric and surface temperatures, magnetic field; radio contact to 54,300,000 mi
ANNA 1B	U.S.	Oct. 31, 1962	Thor-Able-Star	670–728	355	first geodetic satellite; transmitter, flashing lights
Mars 1	U.S.S.R.	Nov. 1, 1962	*	(solar orbit)	1,971	Mars flyby; transmissions ceased after 66,000,000 mi
Syncom 2	U.S.	July 26, 1963	Delta	22,132–22,823	79	first communications satellite in synchronous orbit
Vela	U.S.	Oct. 17, 1963	Atlas-Agena D	63,000–70,000	485 each	first nuclear explosion detection satellites; two orbited
Ranger 7	U.S.	July 28, 1964	Atlas-Agena B	(Moon impact)	807	transmitted 4,316 television photos of Moon's surface in last 13 min before impact
OGO 1	U.S.	Sept. 5, 1964	Atlas-Agena B	173–92,271	1,000	carries 20 coordinated geophysical experiments
Pegasus 1	U.S.	Feb. 16, 1965	Saturn I	308–462	3,080	micrometeoroid data; 2,300-sq-ft detection panels
Intelsat 1 (Early Bird)	U.S.	April 6, 1965	Thrust-augmented Delta	22,251–22,270	86	first communications satellite launched for commercial use
Venera 3	U.S.S.R.	Nov. 16, 1965	*	(Venus impact)	2,117	first spacecraft to land on another planet (Venus); no data
A-1 Asterix	France	Nov. 26, 1965	Diamant	328–1,099	93	first French satellite; third nation to achieve space launch
Luna 9	U.S.S.R.	Jan. 31, 1966	*	(Moon landing)	3,491	first successful lunar "soft landing"; photos of lunar surface televised
Luna 10	U.S.S.R.	March 31, 1966	*	(Moon orbit)	3,528	first spacecraft placed in lunar orbit
Surveyor 1	U.S.	May 30, 1966	Atlas-Centaur	(Moon landing)	2,194	first U.S. soft landing on Moon; televised photos of itself and lunar surface
Lunar Orbiter 1	U.S.	Aug. 10, 1966	Atlas-Agena D	(Moon orbit)	853	first lunar orbiting surface survey spacecraft; studied potential Apollo landing sites
Surveyor 3	U.S.	April 17, 1967	Atlas-Centaur	(Moon landing)	2,282	soft landing; surface sampler claw dug trenches; surface bearing strength data

Sources: Jane's *All the World's Aircraft* (annual); U.S. Senate, *Soviet Space Programs, 1971–75* (1976); U.S. House of Representatives, *World-Wide Space Activities* (1977); Congressional Research Service, *United States and Soviet Programs in Space* (1977). *Not announced.

Satellite (SAS), the first of which was launched in late 1970 and was designed to scan the celestial sphere and catalog stellar X-ray sources.

The Orbiting Astronomical Observatory is a large satellite equipped with a precisely stabilized guidance platform. Its purpose is to map the entire electromagnetic spectrum that Earthbound astronomers cannot see because of the absorptive and distorting features of the atmosphere.

One such device, launched in 1968, carried 11 star telescopes and studied ultraviolet, gamma ray, X-ray, and infrared radiation emissions from precise locations in the celestial sphere. Another, launched in 1972, carried a single large (32-inch) telescope. In 1982, according to plan, a 10-ton Space Telescope will be launched aboard a Space Shuttle. With a 95-inch aperture, the Space Telescope is expected to detect objects that are 50 times fainter than anything visible from Earth.

Pegasus was the name of a series of three satellites flown as a secondary experiment in 1965. The purpose was micrometeoroid detection. Fitted to its service module were wide, extendable wings carrying three layers of aluminum foil to measure penetration by small, medium, and large meteoritic particles.

Orbiting Geophysical Observatory

Orbiting Geophysical observatories are a class of large Earth-orbiting satellites weighing more than 1,000 pounds (450 kilograms) and carrying instrumentation for 20 to 30 experiments designed to study the relationship between Sun, Earth, and space environment. Areas of investigation include cosmic rays, energetic particles, magnetic fields, solar radiation, solar plasma, very-low-frequency radiation, aurora, airglow, micrometeoroids, atmospheric composition, and solar flares. Some of these satellites are placed into a low polar orbit, others into very eccentric orbits ranging out to nearly 100,000 miles.

Biosatellites

Biosatellite was the name of a family of recoverable scientific satellites flown to test the effects of weightlessness, radiation, and absence of the Earth's 24-hour day–night rhythm on biological specimens. The period of test was three days. Experiment specimens included bacteria, mold, seedlings, frog eggs, and amoebas. Retrieval in mid-air was by aircraft. In 1969 a pigtailed monkey was instrumented, orbited, and recovered. Data were inconclusive, however, and the program ceased.

Although the Soviet Union has also conducted a large number of scientific experiments in space, its programs have not been widely publicized. The Cosmos (Kosmos) series of satellites is said to be conducted for research purposes.

The United Kingdom built two Ariel satellites, which were launched by NASA to carry out experiments focussed on ionosphere and atmosphere research. In 1971 was launched Prospero, a technology satellite, using the British Black Arrow. The European Space Research Organization (now European Space Agency), a cooperative enterprise, in 1968 had two satellites called Iris and Heos launched by the U.S. Iris carried seven experiments to study solar and cosmic radiation in the lower Van Allen radiation belt. Heos carried experiments to measure radiation and magnetic fields.

Table 2: Space Exploration by Selected Unmanned Artificial Satellites and Space Probes (cont'd)

name	nation	launched	launch vehicle	perigee–apogee (mi)	spacecraft weight (lb)	remarks
Venera 4	U.S.S.R.	June 12, 1967	*	(Venus landing)	2,439	transmitted data on temperature and pressure of Venusian atmosphere; ceased transmitting before reaching surface
DODGE	U.S.	July 1, 1967	Titan III C	(lunar orbit) 20,661–20,884	430	took first colour photo of full Earth surface
Bios 2	U.S.	Sept. 7, 1967	Thor-Delta	180–188	440	studied effects of weightlessness on various plants and organisms during 47 orbits; capsule recovered
Surveyor 5	U.S.	Sept. 8, 1967	Atlas-Centaur	(Moon landing)	2,216	soft landed on Moon; transmitted 19,000 photos of lunar surface; chemical analysis of lunar soil
Cosmos 186	U.S.S.R.	Oct. 27, 1967	*	130–146	14,398(?)	first automatic rendezvous and docking; ejected capsules recovered
Cosmos 188		Oct. 30, 1967	*	119–158	14,398(?)	
Zond 5	U.S.S.R.	Sept. 14, 1968	*	(Moon orbiting)	12,789(?)	first recovery of circumlunar payload; carried biological experiments, turtles
OAO 2	U.S.	Dec. 7, 1968	Atlas-Centaur	479–485	4,400	first point-stabilized Orbiting Astronomical Observatory; hundreds of stars mapped and measured daily
Mariner 6	U.S.	Feb. 24, 1969	Atlas-Centaur	(Mars probe)	838	photos of Martian surface reveal Moonlike craters
Osumi	Japan	Feb. 11, 1970	Lambda 4S	218–3,188	53	first Japanese satellite; fourth nation to achieve space launch
China 1	China	April 24, 1970	*	237–1,287	382	first Chinese satellite; fifth nation to achieve space launch
Venera 7	U.S.S.R.	Aug. 17, 1970	*	(Venus landing)	2,602	first transmission of data from surface of Venus (Dec. 15)
Luna 16	U.S.S.R.	Sept. 12, 1970	*	(Moon landing)	12,789(?)	soft landed; remotely controlled vehicle bored into lunar surface, placed 101 grams soil into sealed container, launched payload to Earth and successful recovery
Luna 17	U.S.S.R.	Nov. 10, 1970	*	(Moon landing)	12,789(?)	soft landed; remotely controlled vehicle travelled 6.5 mi during 10 lunar days, relaying television pictures and obtaining scientific data
Mars 2	U.S.S.R.	May 19, 1971	Proton	(Mars probe)	10,253	orbited Mars, relayed data, ejected Mars lander (Nov. 27), brief transmission from surface of planet
Mars 3	U.S.S.R.	May 28, 1971	Proton	(Mars probe)	10,253	orbited Mars, relayed data, ejected Mars lander (Dec. 2), brief transmission from surface of planet
Mariner 9	U.S.	May 30, 1971	Atlas-Centaur	(Mars probe)	2,271	inserted into Mars orbit Nov. 13, relayed hundreds of photos and obtained extensive data
Apollo subsatellite	U.S.	Aug. 4, 1971	Apollo 15 service module	62–87 (lunar orbit)	80	released during Apollo 15 postlanding orbit of Moon; lunar magnetometer and charged particle measurements
Prospero	U.K.	Oct. 28, 1971	Black Arrow	915–2,640	146	first all-British satellite; micrometeorite and technology experiments
Pioneer 10	U.S.	March 2, 1972	Atlas-Centaur	(Jupiter probe)	569	first spacecraft to explore asteroid belt, fly by Jupiter, and leave solar system
ERTS 1 (Landsat 1)	U.S.	July 23, 1972	Delta	569–559	1,965	continuous global, multispectral Earth survey
Anik 1 (Telesat A)	Canada	Nov. 9, 1972	Delta	2,262–117	1,250	first domestic Comsat
Pioneer 11 (Pioneer Saturn)	U.S.	April 5, 1973	Atlas-Centaur	(Saturn probe)	571	will be first spacecraft to obtain close photos of Saturn
Mariner 10	U.S.	Nov. 3, 1973	Atlas-Centaur	(Mercury probe)	1,111	first flyby of Mercury
SMS 1	U.S.	May 17, 1974	Delta	22,026–19,331	1,379	first synchronous meteorological satellite
ATS 6	U.S.	May 30, 1974	Titan IIIC	synchronous	3,090	demonstrates new potential for satellite communication broadcast
Helios 1	U.S.–W.Ger.	Dec. 10, 1974	Titan IIIE-Centaur	(Sun probe)	816	closest approach to the Sun
Venera 9	U.S.S.R.	June 8, 1975	D-1	(Venus probe)	10,884	photos from surface of Venus
Viking 1	U.S.	Aug. 20, 1975	Titan III-Centaur	(Mars probe)	7,497	first photos from Mars surface
Aryabhata	India–U.S.S.R.	April 19, 1975	*	*	*	first Indian satellite
Lageos	U.S.	May 4, 1976	Delta	3,694–3,626	906	laser geodetic satellite

Sources: Jane's *All the World's Aircraft* (annual); U.S. Senate, *Soviet Space Programs, 1971–75* (1976); U.S. House of Representatives, *World-Wide Space Activities* (1977); Congressional Research Service, *United States and Soviet Programs in Space* (1977). *Not announced.

Other countries that have developed scientific satellites include France, whose spacecraft have been launched both with a French space booster and by the United States; Canada, whose International Satellite for Ionospheric Studies was launched in 1969 by NASA; and Italy, using a United States launch vehicle to launch the San Marco satellite. Japan and the People's Republic of China have launched experimental satellites using their own launch vehicles, as indicated in Tables 1 and 2.

Lunar programs. An early target for space exploration was the Moon, man's nearest extraterrestrial neighbour. By 1978 the Soviet Union and the United States were still the only countries with space launch vehicles capable of sending payloads to the Moon.

Luna series

The Soviet Luna 3 first photographed the far side of the Moon in 1959. Other Luna-series vehicles orbited the Moon or soft landed, transmitted television photographs of the surface, and tested the load-bearing characteristics of the lunar soil. Luna 16, in September 1970, soft landed, obtained a core sample of soil, and returned it to Earth in a sealed capsule.

A significant development was Luna 17, which in November 1970 landed an automated roving vehicle. Remotely controlled from the Earth, this television-equipped craft ranged several miles and survived nine lunar nights in robot exploration of the lunar surface.

Another Soviet series of lunar probes used the so-called Zond vehicles. A number of these made circumlunar flights carrying biological specimens and returned photographs to Earth.

Three U.S. programs studied the Moon with spacecraft: Ranger, Surveyor, and Lunar Orbiter. Ranger spacecraft transmitted television pictures of the Moon from as close as 10,000 feet before crashing on the surface. The Surveyor spacecraft was a soft lander, using retrorockets to reduce trajectory velocity to zero. Safely landed, Surveyor transmitted to Earth thousands of television pictures of the lunar surface and data on lunar soil characteristics and chemical composition. Lunar Orbiter was designed to obtain high-quality television photos of the Moon's surface in selected areas in preparation for Project Apollo manned landings.

Planetary programs. Both the United States and the Soviet Union have sent spacecraft to Mars and Venus, the Soviet Union concentrating most effort on Venus and the United States on Mars. In addition, the U.S. has sent spacecraft to Jupiter, Mercury, and Saturn.

Venera series

The Soviet Union made numerous attempts to launch spacecraft to Mars and Venus and achieved success, beginning with Venera 4 in 1967, and subsequently. In each of these flights an instrumented capsule was dropped by parachute into the hot, dense Venusian atmosphere. In each case, data were incomplete, though Venera 7 (1970) and 8 (1972) survived the landing and returned significant data from the surface. Venera 9 and 10 (1975) provided further surface data and the first photographs taken on the surface of another planet.

In May 1971 the Soviet Union launched two spacecraft to Mars. Arriving six months later, both spacecraft ejected landing capsules. The capsules had been

Table 3: Manned Space Flights

spacecraft	launch date	crew	duration (hr:min)	remarks
Vostok 1	April 12, 1961	Y.A. Gagarin	1:48	first manned flight
Mercury–Redstone 3	May 5, 1961	A.B. Shepard, Jr.	0:15	first U.S. flight; suborbital
Mercury–Redstone 4	July 21, 1961	V.I. Grissom	0:16	suborbital; capsule sank after landing
Vostok 2	Aug. 6, 1961	G.S. Titov	25:18	first flight exceeding 24 hr
Mercury–Atlas 6	Feb. 20, 1962	J.H. Glenn, Jr.	4:55	first American to orbit
Mercury–Atlas 7	May 24, 1962	M.S. Carpenter	4:56	landed 400 km from target
Vostok 3	Aug. 11, 1962	A.G. Nikolayev	94:22	first dual mission (with Vostok 4)
Vostok 4	Aug. 12, 1962	P.R. Popovich	70:57	came within 6 km of Vostok 3
Mercury–Atlas 8	Oct. 3, 1962	W.M. Schirra, Jr.	9:13	landed 8 km from target
Mercury–Atlas 9	May 15, 1963	L.G. Cooper, Jr.	34:20	first long U.S. flight
Vostok 5	June 14, 1963	V.F. Bykovsky	119: 6	second dual mission (with Vostok 6)
Vostok 6	June 16, 1963	V.V. Tereshkova	70:50	first woman in space
Voskhod 1	Oct. 12, 1964	V.M. Komarov K.P. Feoktistov B.G. Yegorov	24:17	first 3-man crew
Voskhod 2	March 18, 1965	A.A. Leonov P.I. Belyayev	26: 2	first extravehicular activity (Leonov, 10 min)
Gemini 3	March 23, 1965	V.I. Grissom J.W. Young	4:53	first U.S. 2-man flight; first manual manoeuvres in orbit
Gemini 4	June 3, 1965	J.A. McDivitt E.H. White II	97:56	21-min extravehicular activity (White)
Gemini 5	Aug. 21, 1965	L.G. Cooper, Jr. C. Conrad, Jr.	190:55	duration record
Gemini 7	Dec. 4, 1965	F. Borman J.A. Lovell, Jr.	330:35	duration record
Gemini 6-A	Dec. 15, 1965	W.M. Schirra, Jr. T.P. Stafford	25:51	rendezvous within 3 m of Gemini 7
Gemini 8	March 16, 1966	N.A. Armstrong D.R. Scott	10:41	first docking of 2 orbiting spacecraft
Gemini 9-A	June 3, 1966	T.P. Stafford E.A. Cernan	72:21	extravehicular activity; rendezvous
Gemini 10	July 18, 1966	J.W. Young M. Collins	70:47	first dual rendezvous (Gemini 10 with Agena 10, then Agena 8)
Gemini 11	Sept. 12, 1966	C. Conrad, Jr. R.F. Gordon, Jr.	71:17	highest Earth-orbit altitude (1,372 km)
Gemini 12	Nov. 11, 1966	J.A. Lovell, Jr. E.E. Aldrin, Jr.	94:35	longest extravehicular activity to date (Aldrin, 5:37)
Soyuz 1	April 23, 1967	V.M. Komarov	26:37	cosmonaut killed in re-entry accident
Apollo 7	Oct. 11, 1968	W.M. Schirra, Jr. D.F. Eisele R.W. Cunningham	260: 9	first U.S. 3-man mission
Soyuz 3	Oct. 26, 1968	G. Beregovoy	94:51	manoeuvred near unmanned Soyuz 2
Apollo 8	Dec. 21, 1968	F. Borman J.A. Lovell, Jr. W.A. Anders	147: 1	first manned orbit(s) of Moon; first manned departure from Earth's sphere of influence; highest speed ever attained in manned flight
Soyuz 4	Jan. 14, 1969	V. Shatalov	71:23	Soyuz 4 and 5 docked and transferred 2 cosmonauts from Soyuz 5 to Soyuz 4
Soyuz 5	Jan. 15, 1969	B. Volynov A. Yeliseyev Y. Khrunov	72:56	
Apollo 9	March 3, 1969	J.A. McDivitt D.R. Scott R.L. Schweickart	241: 1	simulation of lunar operations in Earth orbit
Apollo 10	May 18, 1969	T.P. Stafford J.W. Young E.A. Cernan	192: 3	successfully demonstrated Lunar Module descent to 14,300 m from the lunar surface
Apollo 11	July 16, 1969	N.A. Armstrong M. Collins E.E. Aldrin, Jr.	195:19	first landing of human being on lunar surface and safe return to Earth; first return of rock and soil samples to Earth, and manned deployment of experiments on lunar surface
Soyuz 6	Oct. 11, 1969	G. Shonin V. Kubasov	118:42	Soyuz 6, 7, and 8 operated as group flight without actually docking; each conducted certain experiments, including welding and Earth and celestial observations
Soyuz 7	Oct. 12, 1969	A. Filipchenko V. Volkov V. Gorbatko	118:41	
Soyuz 8	Oct. 13, 1969	V. Shatalov A. Yeliseyev	118:50	
Apollo 12	Nov. 14, 1969	C. Conrad, Jr. R.F. Gordon, Jr. A.L. Bean	244:36	second manned lunar landing; retrieved parts of Surveyor 3 spacecraft

Source: Through 1976, National Aeronautics and Space Administration, *Aeronautics and Space Report of the President, 1976 Activities.*

previously sterilized to prevent contamination by Earth organisms. Retrorocket and parachute braking were used. The capsules were equipped with television transmission, but the signals ceased shortly after landing, possibly because of spacecraft damage from high winds or from sinking into dust. The Mars 2 and 3 spacecraft continued to orbit Mars in highly elliptical orbits, transmitting a variety of scientific data.

Mariner probes

The United States Mariner program encompasses those interplanetary probes that are designed to pass close to Mars, Venus, and Mercury. Mariners 2 in 1962 and 5 in 1967 passed Venus within 22,000 and 2,500 miles (about 35,000 and 7,600 kilometres) and made temperature and atmospheric density measurements. Mariners 4 in 1964 and 6 and 7 in 1969 obtained photographs of the Martian surface and made significant analyses of the atmosphere. Mariners 6 and 7 made thermal maps of Mars by means of infrared radiometers. Mariners 1 and 3 were failures.

In May 1971 the U.S. launched Mariner 9 toward Mars. It arrived and was placed in eccentric orbit in November, while the planet was shrouded in a global dust storm with winds perhaps of 100 miles per hour. After several weeks the planet's thin atmosphere cleared and the orbiter's two television cameras transmitted clear photographs of the surface, revealing many new details and formations. Other sensors measured atmospheric pressure (about 3–8.3 millibars compared with 1,013 for Earth), temperature, the ionosphere, and the gravitational field. Large areas of the planet were mapped, and the two moons of Mars, Phobos and Deimos, were photographed.

Mariner 10 was launched in 1973 to explore Mercury. Swinging by Venus, utilizing Venusian gravity to increase velocity, the spacecraft flew by Mercury, obtaining the first close-up photographs of this planet.

One of the most exciting planetary missions in the his-

Table 3: Manned Space Flights (cont'd)

spacecraft	launch date	crew	duration (hr:min)	remarks
Apollo 13	April 11, 1970	J.A. Lovell, Jr. F.W. Haise, Jr. J.L. Swigert, Jr.	142:55	mission aborted because of explosion in Service Module; ship circled Moon, with crew using Lunar Module as "lifeboat" until just prior to re-entry
Soyuz 9	June 1, 1970	A.G. Nikolayev V.I. Sevastianov	424:59	duration record
Apollo 14	Jan. 31, 1971	A.B. Shepard, Jr. S.A. Roosa E.D. Mitchell	216: 2	third manned lunar landing
Soyuz 10	April 22, 1971	V. Shatalov A. Yeliseyev N. Rukavishnikov	47:46	docked with Salyut 1
Soyuz 11	June 6, 1971	G.T. Dobrovolskiy V.N. Volkov V.I. Patsayev	570:22	docked with Salyut 1; crew perished during return phase of mission
Apollo 15	July 26, 1971	D.R. Scott A.M. Worden J.B. Irwin	295:12	fourth manned lunar landing and first use of Lunar Roving Vehicle
Apollo 16	April 16, 1972	J.W. Young C.M. Duke, Jr. T.K. Mattingly II	265:51	fifth manned lunar landing
Apollo 17	Dec. 7, 1972	E.A. Cernan H.H. Schmitt R.E. Evans	301:52	sixth and final Apollo manned lunar landing
Skylab 2	May 25, 1973	C. Conrad, Jr. J.P. Kerwin P.J. Weitz	627:50	docked with Skylab; repaired damaged station; duration record
Skylab 3	July 28, 1973	A.L. Bean J.R. Lousma O.K. Garriott	1,427: 9	docked with Skylab; duration record
Soyuz 12	Sept. 27, 1973	V. Lazarev O. Makarov	47:16	checkout of improved Soyuz
Skylab 4	Nov. 16, 1973	G.P. Carr E.G. Gibson W.R. Pogue	2,017:16	docked with Skylab; duration record
Soyuz 13	Dec. 18, 1973	P. Klimuk V. Lebedev	188:55	astrophysical, biological, and Earth resources experiments
Soyuz 14	July 3, 1974	P. Popovich Y. Artyukhin	377:30	docked with Salyut 3
Soyuz 15	Aug. 26, 1974	G. Sarafanov L. Demin	48:12	failed to dock with Salyut 3
Soyuz 16	Dec. 2, 1974	A. Filipchenko N. Rukavishnikov	142:24	test of ASTP configuration
Soyuz 17	Jan. 10, 1975	A. Gubarev G. Grechko	709:20	docked with Salyut 4
Anomaly	April 5, 1975	V. Lazarev O. Makarov	0:20	Soyuz stages failed to separate; crew recovered after abort
Soyuz 18	May 24, 1975	P. Klimuk V. Kubasov	1,511:20	docked with Salyut 4
Soyuz 19	July 15, 1975	A. Leonov V. Kubasov	142:31	target for Apollo in docking and joint experiments in ASTP mission
Apollo	July 15, 1975	T.P. Stafford D.K. Slayton V.D. Brand	217:28	docked with Soyuz 19 in joint experiments of ASTP mission
Soyuz 21	July 6, 1976	B. Volynov V. Zholobov	1,182:24	docked with Salyut 5
Soyuz 22	Sept. 15, 1976	V. Bykovskiy V. Aksenov	189:54	Earth resources study with multispectral camera system
Soyuz 23	Oct. 14, 1976	V. Zudov V. Rozhdestvenskiy	48: 6	failed to dock with Salyut 5
Soyuz 24	Feb. 7, 1977	V.V. Gorbatko Y. Glazkov	425:23	docked with Salyut 5
Soyuz 25	Oct. 9, 1977	V. Kovalenok V. Ryumin	48:46	failed to dock with Salyut 6
Soyuz 26	Dec. 10, 1977	G. Grechko Y. Romanenko	2,314:00	docked with Salyut 6; crew returned in Soyuz 27 spacecraft after setting crew duration record
Soyuz 27	Jan. 10, 1978	V. Dzhanibekov O. Makarov	142:59	docked with Salyut 6; first double docking; Soyuz 27 crew returned in Soyuz 26 spacecraft
Soyuz 28	March 2, 1978	A.A. Gubarev V. Remek	190:17	docked with Salyut 6; Czechoslovakia became third country to send a citizen (Remek) into space

Source: Through 1976, National Aeronautics and Space Administration, *Aeronautics and Space Report of the President, 1976 Activities.*

Viking missions

tory of the U.S. space program began with the launching of Vikings 1 and 2 (1975). Each spacecraft consisted of two sections, one that would remain in orbit around Mars, while the other completed a soft landing on the surface. Both Viking 1 and 2 landers made successful descents to the surface in July and September 1976, where they conducted the first long-term *in situ* analysis of the surface material, atmosphere, and possible life-forms of another planet.

The first spacecraft to fly by Jupiter were Pioneers 10 and 11, launched by the U.S. in 1972 and 1973. Both returned physical data and hundreds of revealing photographs of the planet. Pioneer 10 remains on a trajectory and travelling at sufficient velocity to cause it to become the first spacecraft to leave the solar system. On board is a plaque for communication with extraterrestrial intelligence should any be encountered. Pioneer 11 is on a trajectory to pass close to the rings of Saturn.

Voyagers 1 and 2 were launched by the U.S. in August and September 1977 on a fly-by mission to Jupiter and Saturn.

MANNED PROGRAMS

By the late 1970s, only the United States and the Soviet Union had launched manned satellites (Table 3).

Vostok and Mercury. Both the United States and the Soviet Union early made manned flight a major goal in their space programs. The Soviet intercontinental ballistic missile was much larger than the United States counterpart, having been designed to carry a heavier warhead. The first Soviet manned spacecraft, Vostok 1, weighed over 10,000 pounds (4,500 kilograms), while the U.S. Mercury weighed only about a third as much.

The Soviet Union launched Yury Alekseyevich Gagarin into a single Earth orbit on April 12, 1961. In August 1962 the Soviet Union launched two manned space-

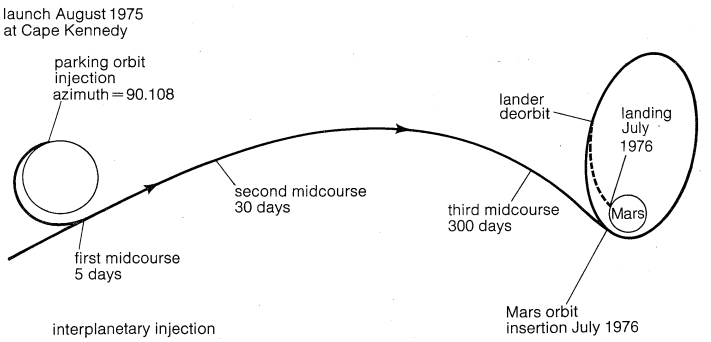


Figure 4: A typical Viking mission profile.
By courtesy of the National Aeronautics and Space Administration

craft in nearly identical orbits one day apart. In June 1963 a similar set of launches was made, one containing the first woman cosmonaut, Valentina Tereshkova.

At that time the U.S. had made only four orbital manned flights, and Soviet cosmonauts had logged nearly eight times as many man-hours in space as had U.S. astronauts. By May 1978, however, the U.S. had accumulated 22,503 hours of flight time compared with 19,823 hours by the Soviet Union.

Gemini and Voskhod. While the Vostok and Mercury flights had demonstrated that man could function while weightless and in space, neither spacecraft had operational manoeuvrability. The U.S. Gemini program was designed to develop objectives of rendezvous with a target vehicle and of docking, a capacity vital to the Apollo Moon-landing program. The dynamics of rendezvous with a target vehicle launched a day earlier called for extreme precision in tracking, timing of launch, and placement of spacecraft in almost identical orbit. A delay of one second in launch would allow the target vehicle to fly five miles beyond the rendezvous point. The Soviet Union meanwhile advanced to a three-man flight with Voskhod 1, October 12–13, 1964. On March 18, 1965, a two-man flight was made in which the cosmonaut Aleksey Leonov passed through an airlock to become the first man to float free in space. On Gemini 4, in June 1965, the United States astronaut Edward H. White II duplicated the feat. Called extravehicular activity (EVA), this manoeuvrability demonstrated the increasing ability of man to function in space. It also proved that man could adapt to weightlessness for as long as two weeks and readapt to the gravity of Earth.

First EVA's

By courtesy of the National Aeronautics and Space Administration

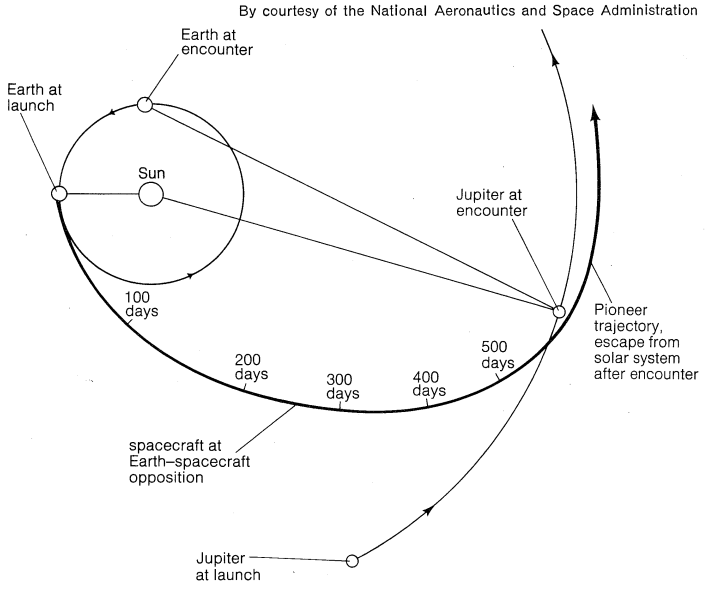


Figure 5: The Jupiter trajectory flown by Pioneer 10, launched March 2, 1972.

Soyuz. In April 1967 the Soviet Union launched Soyuz 1, called the largest, most complex spacecraft yet flown. V.M. Komarov, its cosmonaut, was killed in an apparent failure of the parachute recovery system. In October 1968 Soyuz 3 made rendezvous several times with the unmanned Soyuz 2, sent aloft a day earlier. Early in 1969 Soyuz 4 and Soyuz 5, launched on successive days, made rendezvous and docked; two cosmonauts left Soyuz 5 and transferred to Soyuz 4, undocked, re-entered, and landed. In October 1969, Soyuz 6, 7, and 8 were launched on successive days; welding experiments were performed, and the spacecraft landed with their combined crew of seven after five days in orbit.

In June 1970 Soyuz 9 established an endurance record for manned space flight, nearly 18 days. The cosmonauts reportedly required nearly two days to regain coordination and normal sleeping habits upon return to Earth. Soyuz 10, in April 1971, rendezvoused with a previously orbited "space station," Salyut. The flight lasted only two days, however, and it has been presumed that operational difficulties occurred. In June 1971, Soyuz 11 carried three cosmonauts to rendezvous with Salyut. Twenty-seven hours after launch, the cosmonauts transferred to the 47-foot- (14-metre-) long Salyut space laboratory. This successful flight ended in tragedy when a minor malfunction of a door seal during re-entry caused loss of the air in the spacecraft and the death of all three cosmonauts.

The flight of Soyuz 9

The remaining Soyuz missions flown by 1978 carried two-man crews. Soyuz 12 provided an operational test of new systems developed since the Soyuz 11 tragedy. Soyuz 13 performed a variety of scientific experiments. Soyuz 14, 15, 17, 18, and 20–27 were all related to one of the Salyut orbiting laboratories launched after April 1971.

The Salyut vehicle is equipped to conduct extended missions of scientific or manned military reconnaissance in Earth orbit. Interior space is approximately 100 cubic metres. Each of the first five Salyut stations eventually was abandoned and destroyed on re-entry. Soyuz 26 docked with Salyut 6 (a vehicle equipped with a docking port at each end) on December 11, 1977. On January 11, 1978, Soyuz 27 docked with the space station, bringing up supplies to the crew, and returned to Earth five days later. Soyuz 28 docked with the space station on March 3, 1978, and on March 16 the crew of Soyuz 26 returned to Earth in Soyuz 27, having set a duration record of 96 days in space. Soyuz 28 carried a Czechoslovakian cosmonaut, the first non-Soviet, non-U.S. person in space.

Apollo program. In May 1961 Pres. John F. Kennedy committed the United States to landing a manned spacecraft on the Moon "before this decade is out." The name given to the program was Apollo.

Several techniques for the lunar mission appeared to be feasible and were studied in detail. These included a direct method, whereby a single giant launch vehicle would launch a spacecraft toward the Moon. Near the Moon the spacecraft would be slowed down by its own propulsion system, called the lunar landing stage. Another propulsion system, the lunar takeoff stage, would launch the spacecraft from the Moon and send it toward Earth.

A second proposal, called the Earth-orbit rendezvous method, differed from the direct method in that two somewhat smaller launch vehicles would separately transport two parts of the spacecraft (*i.e.*, the lunar landing stage, and the lunar takeoff stage and spacecraft proper) into Earth Orbit. There they would be joined and then propelled as a single craft toward the Moon.

A third method, called the lunar-orbit rendezvous technique, was selected. This method would employ a single Saturn V launch vehicle to send the three-man Apollo spacecraft, weighing 100,000 pounds (45,000 kilograms), on a trajectory toward the Moon. Approximately 2½ days later, as the craft approached the Moon, its own propulsion system would place it in an orbit around the Moon. In lunar orbit one element of the spacecraft, called a Lunar Module (LM), with two men aboard, would separate from the mother ship and be flown toward the Moon, while the third man circled the Moon in the other

The lunar-orbit rendezvous technique

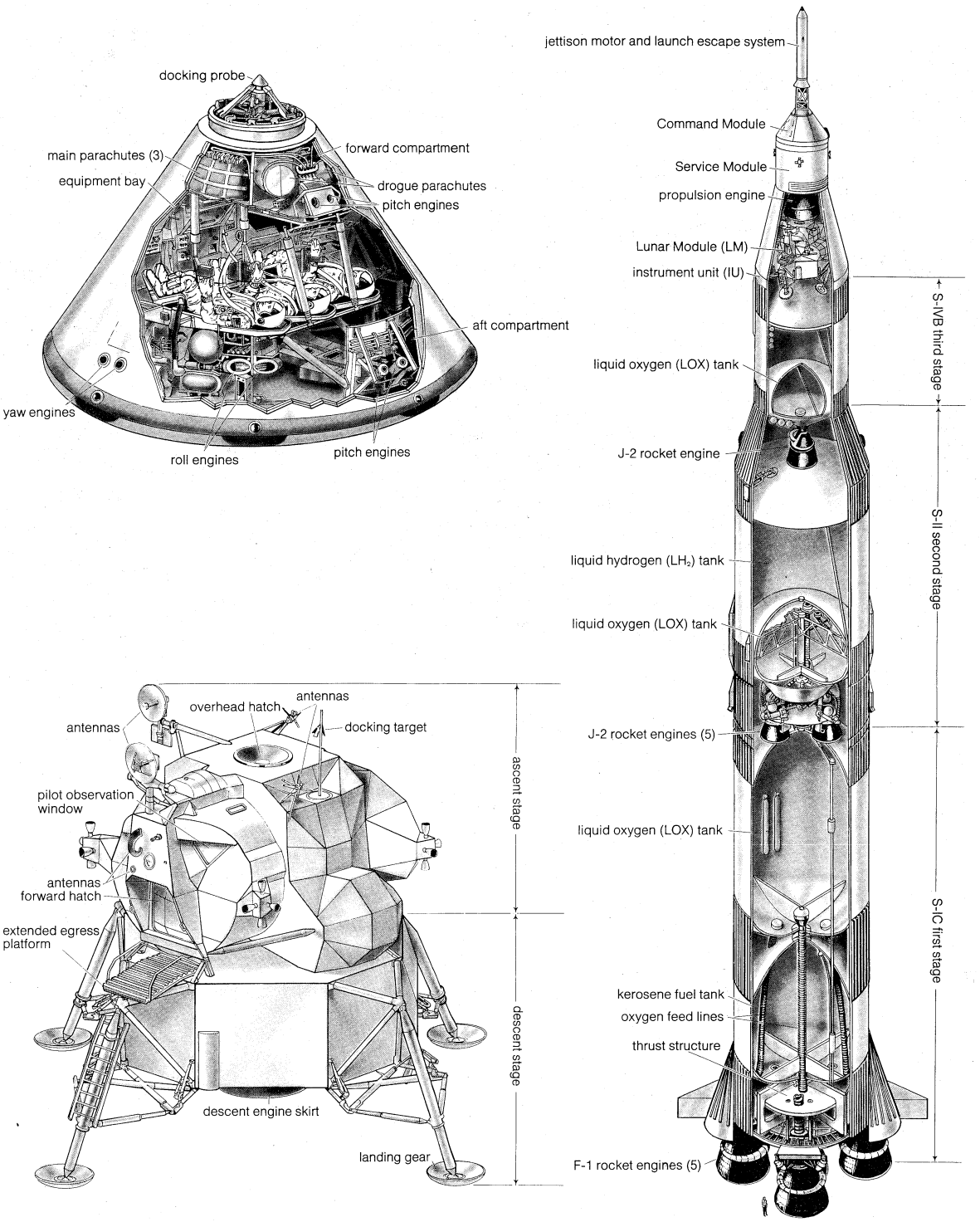


Figure 6: *The Apollo program.*
(Right) Saturn V launch vehicle and Apollo spacecraft, with details of (top left) the Command Module and (bottom left) the Lunar Module.

section of the spacecraft, the Command Module (CM). After executing a soft landing on the lunar surface, the two lunar explorers would make scientific observations and collect geological samples. After about one day, they would take off from the surface and fly toward a rendezvous with the Command Module in lunar orbit. Spacecraft propulsion would then eject the spacecraft from lunar orbit and send it back toward the Earth. Precise guidance and control on the return trip would allow the spacecraft to enter within the limits of the narrow re-entry corridor described above under *Re-entry and recovery*. Within the atmosphere the Apollo Command Module would have a limited degree of manoeuvrability.

Apollo modules

The Apollo spacecraft (Figure 6) consisted of three modules: command, service, and lunar. The Command Module was the spacecraft's control centre and housed the crew of three astronauts. It was conical in shape, 12 feet high and 13 feet in diameter, and weighed 10,000 pounds. The launch escape system, a tower-like structure above the Command Module, provided power to lift the Command Module up and away from the Saturn launch vehicle should an emergency occur after fuelling and before injection into orbit. Solid-propellant rockets provided this standby emergency power.

Below the Command Module was the Service Module, 12.8 feet in diameter, 22 feet long, and containing the

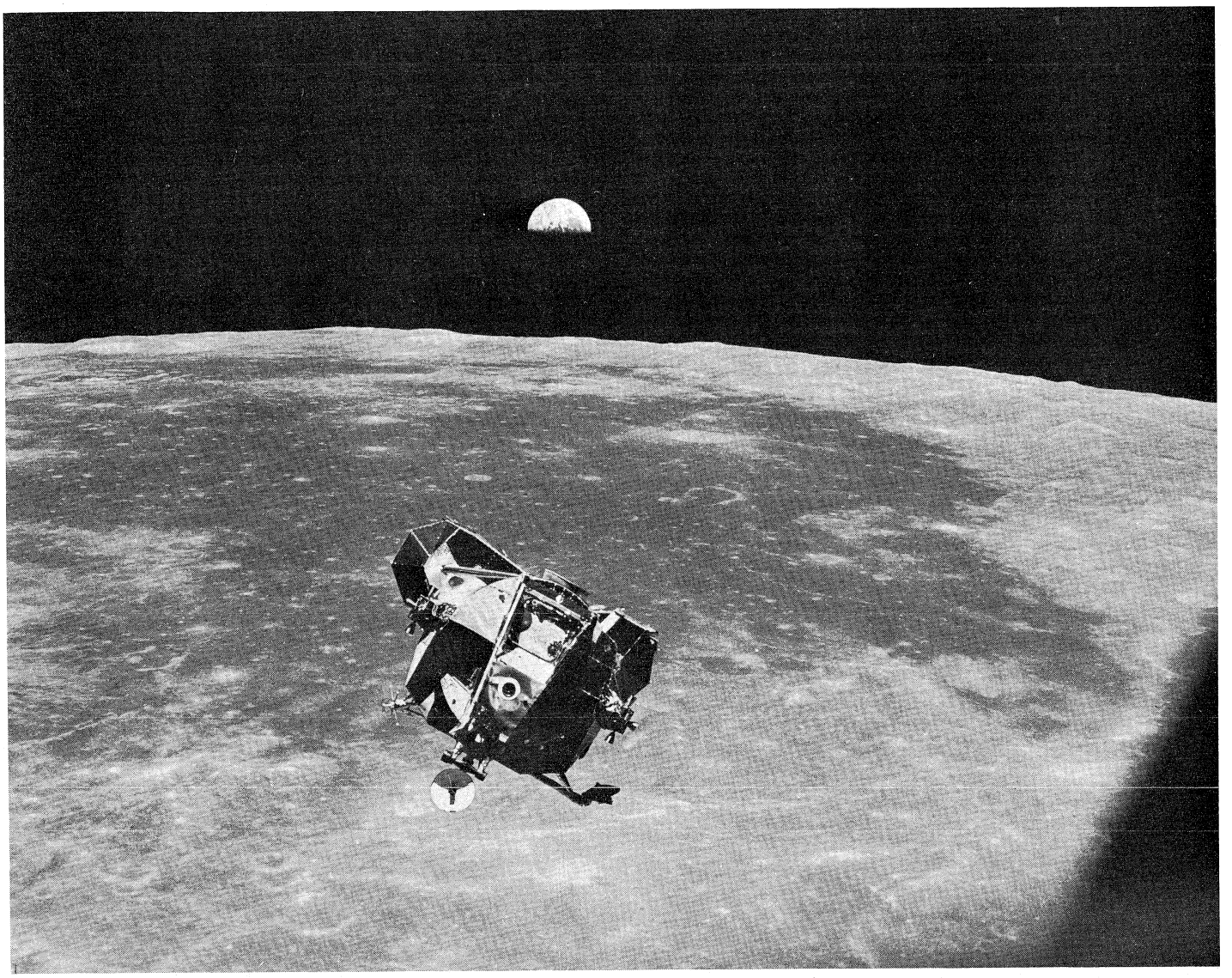


Figure 7: *Return of the first Lunar Module to the Apollo spacecraft.*
The Apollo 11 Lunar Module with astronauts Neil A. Armstrong and Edwin E. Aldrin, Jr., aboard, photographed from the Command Module by Michael Collins as it ascended from the lunar surface to rendezvous in orbit on July 21, 1969. The large, dark-coloured area in the background is Smyth's Sea (Mare Smythii), centred at longitude 85° E and latitude 2° S on the lunar nearside. The Earth is seen rising in the west above the lunar horizon.
By courtesy of the National Aeronautics and Space Administration

propulsion system for midcourse corrections, retrofire to achieve lunar orbit, and thrust to return from lunar orbit into Earth trajectory. The pressure-fed rocket engine was capable of developing 22,000 pounds of thrust, was ablating cooled, and used nitrogen tetroxide and hydrazine-type fuel.

Between the Service Module and the Saturn V launch vehicle was the Lunar Module (LM) adapter. This unit was 29 feet long and 21.7 feet in diameter where it joined the S-IVB stage, tapering to 13 feet where it joined the Service Module. The LM housed in this section was designed to carry two astronauts from lunar orbit to the surface of the Moon and return them to the Command Module. The LM was approximately 21 feet high and 11 feet in diameter, weighing 30,000 pounds (Earth weight). There were two engines, one with controllable thrust for descent to the surface of the Moon and the other with a fixed 35,000-pound thrust for ascent to rendezvous with the Command Module in lunar orbit. Both engines were pressure fed, ablating cooled, and used the same propellants as the Service Module engine.

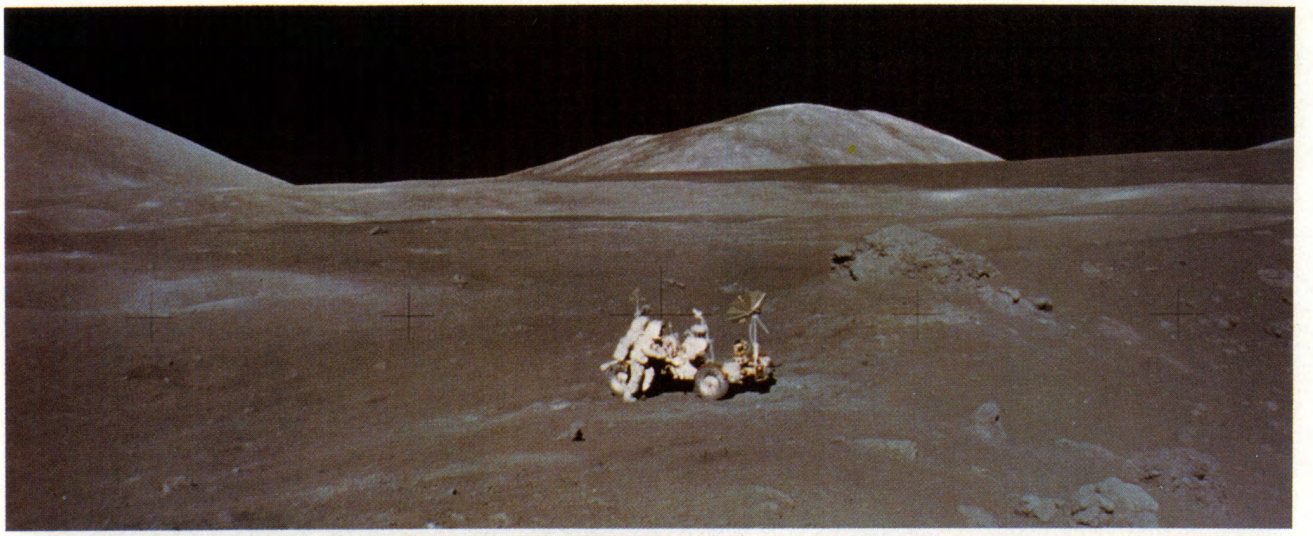
The first manned Apollo flight, in Earth orbit, scheduled for February 1967, was delayed when on January 27, 1967, during a countdown rehearsal, a fire broke out inside the spacecraft cabin and spread rapidly in the concentrated oxygen environment at higher than sea level pressure. Three astronauts lost their lives.

In November 1967 a Saturn V placed an unmanned Apollo in orbit. Two more unmanned Apollo flights were made and on October 11, 1968, Apollo 7, with three astronauts, tested the Command Module in a 163-orbit flight. Apollo 8, launched on December 21, 1968, after Earth orbit was injected into a translunar trajectory, inserted into lunar orbit for 10 orbits, and then returned safely to Earth. Live television broadcasts were made during the flight and while orbiting the Moon.

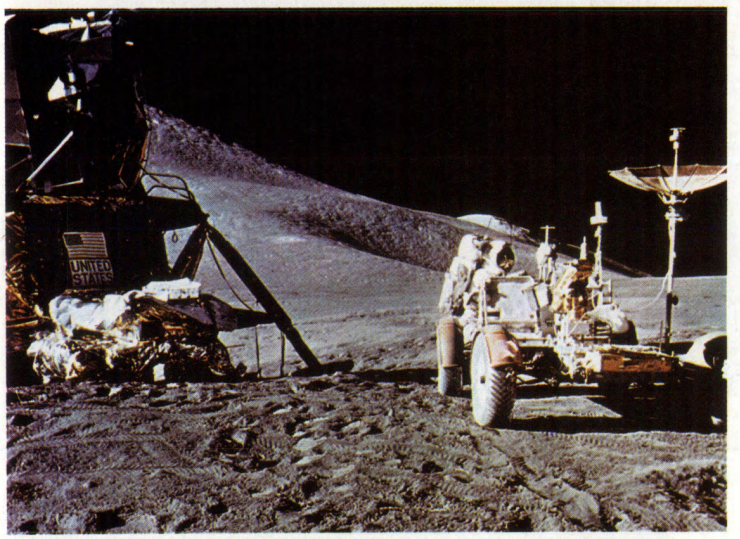
Apollo 9, launched on March 3, 1969, on a 10-day trip, checked out the Lunar Module operation in Earth orbit. Apollo 10 lifted off on May 18. Paralleling Apollo 8's flight into lunar orbit, two astronauts transferred to the LM and lowered their orbit to 46,900 feet (14,300 metres) above the Moon.

Apollo 11, launched July 16, 1969, manned by Neil A. Armstrong, Edwin E. Aldrin, Jr., and Michael Collins, repeated Apollo 10's flight to lunar orbit. Armstrong and Aldrin transferred to the Lunar Module, and descent and landing on the Moon were made July 20 at 8:17 PM Greenwich Mean Time, at 23.26° E and 0.41° N. Armstrong stepped onto the Moon's surface at 2:56 AM GMT. His first words were, "That's one small step for a man,

". . . one giant leap for mankind"

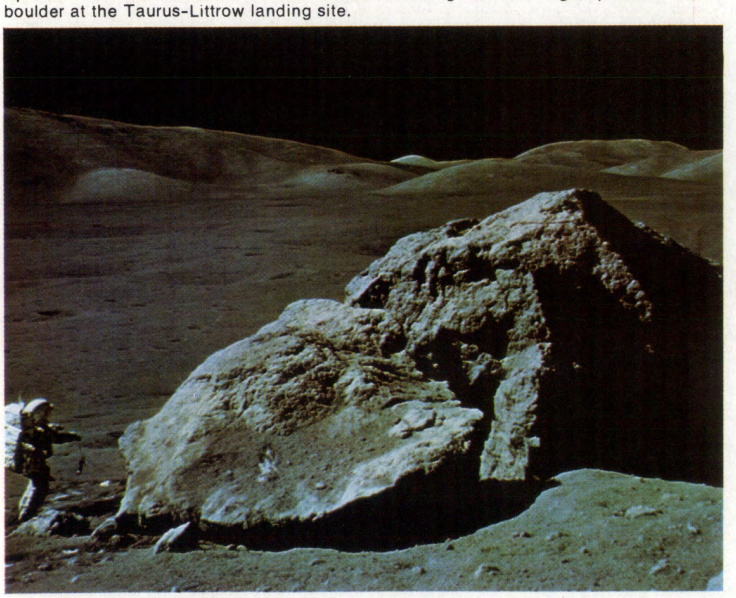


Apollo 17 scientist-astronaut Harrison Schmitt working at the Lunar Roving Vehicle at the Taurus-Littrow landing site. This is the area where Schmitt first spotted orange soil. Apollo 17 was launched December 7, 1972.



Apollo 15 astronaut James Irwin standing in back of the Lunar Roving Vehicle; the Lunar Module (LM) is at left with the modular equipment storage assembly (MESA) in front of it. Apollo 15 was launched July 26, 1971.

Apollo 17 scientist-astronaut Harrison Schmitt standing next to a huge, split lunar boulder at the Taurus-Littrow landing site.



Man on the Moon

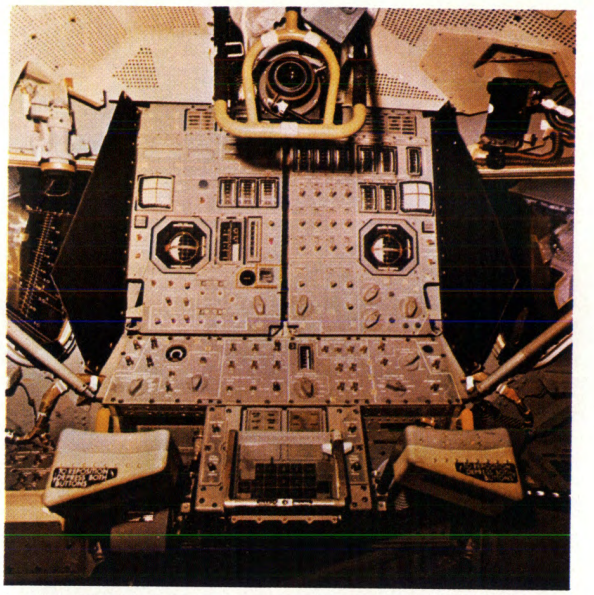
Astronaut Edwin Aldrin, Jr., walking near the Lunar Module during the Apollo 11 extravehicular activity (EVA). Apollo 11 was launched July 16, 1969.



Plate 1: By courtesy of National Aeronautics and Space Administration



Apollo 15 space vehicle as it lifted off from Cape Kennedy, Florida. A camera mounted at the mobile launcher's 360-foot level recorded this photograph.

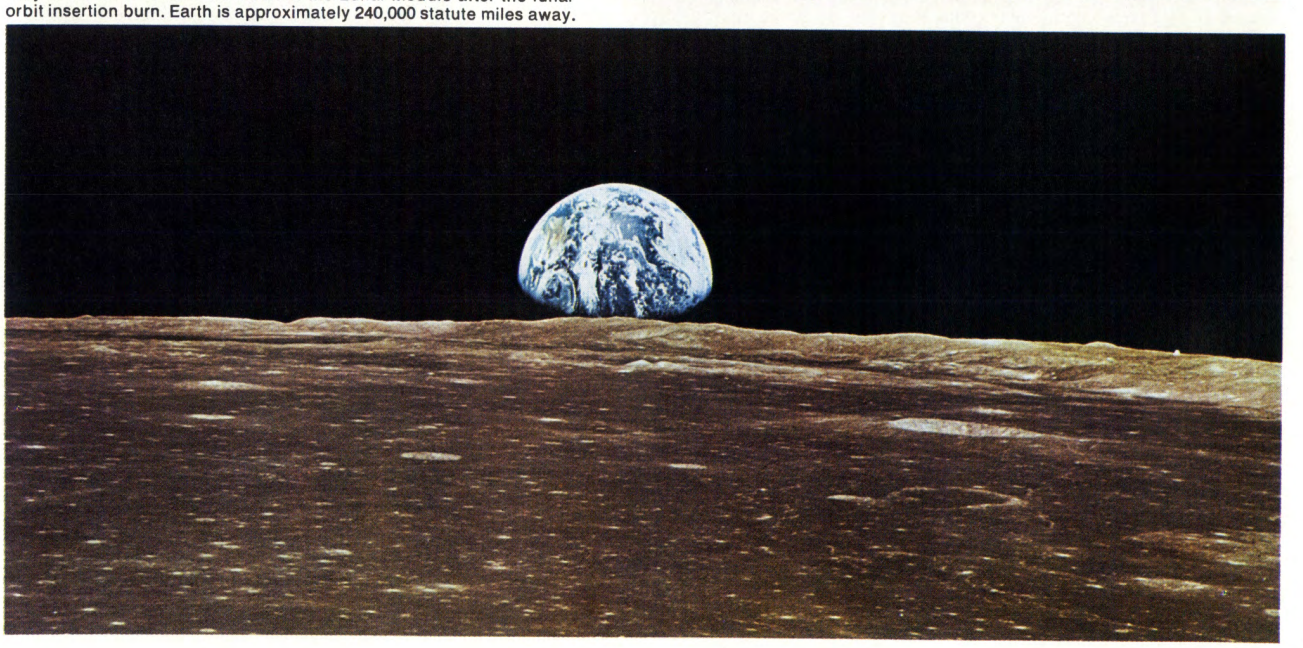


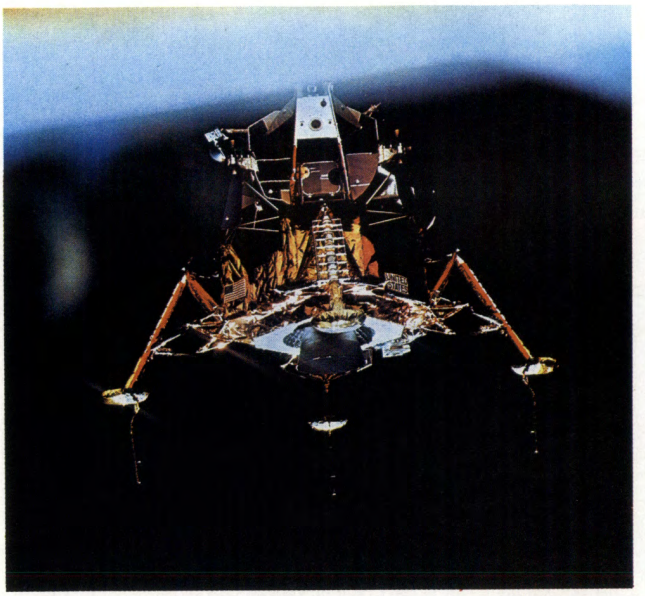
Interior view of the Apollo 11 Lunar Module showing the main flight control panels.

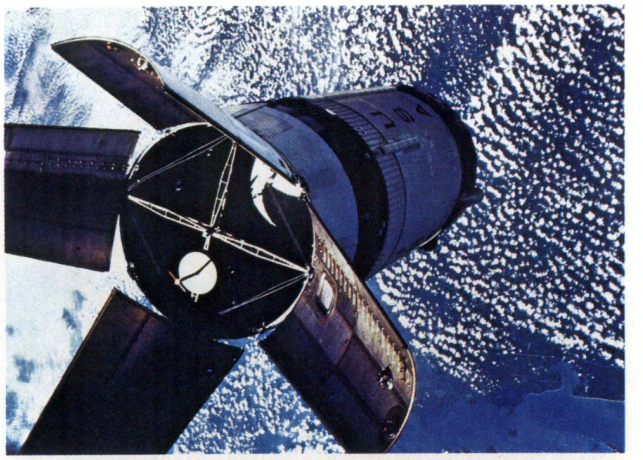


(Right) Firing room 2 at Launch Control Center, Kennedy Space Center, Florida, where personnel monitored the Apollo 14 countdown activities on January 31, 1971.

This view of the rising Earth greeted the Apollo 11 astronauts as they came from behind the Moon in the Lunar Module after the lunar orbit insertion burn. Earth is approximately 240,000 statute miles away.





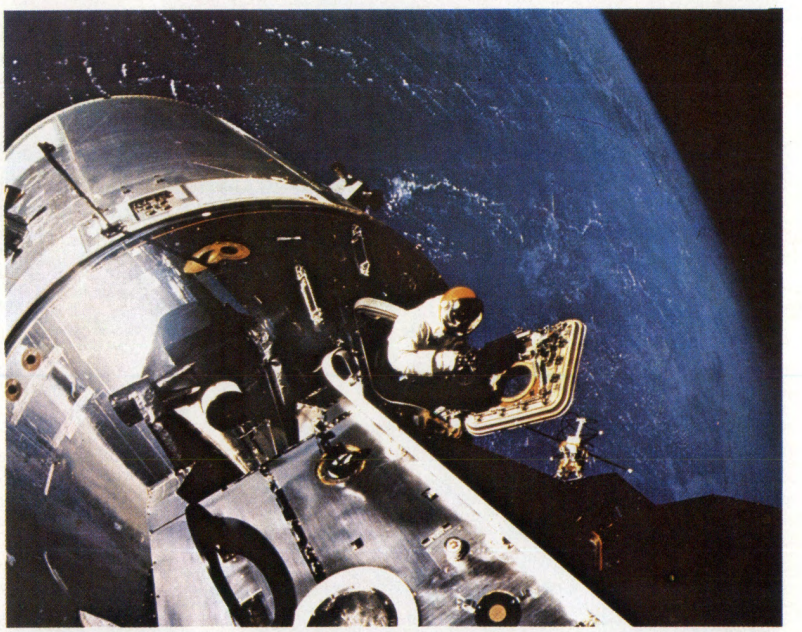


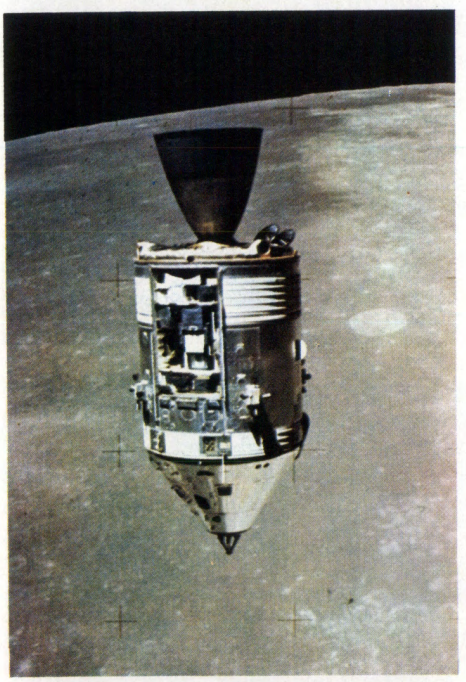
The expended Saturn IVB stage as photographed from the Apollo 7 spacecraft during transposition and docking manoeuvres near the end of the second revolution of the Earth at an approximate altitude of 124 nautical miles. Apollo 7 was launched October 11, 1968.

(Left) Apollo 11 Lunar Module with its three landing gear foot pads deployed. This photograph was taken from the Command Module (CM) as the two spacecraft moved apart.

Apollo program

(Right) With the Earth in the background, astronaut David Scott opens the hatch of the Apollo 9 Command and Service modules to view the just completed docking with the Lunar Module. Apollo 9 was launched March 3, 1969.



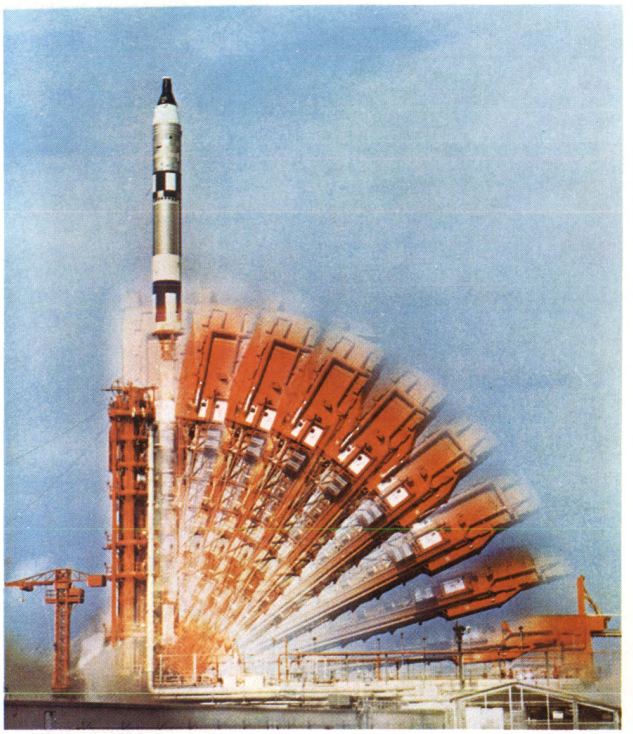


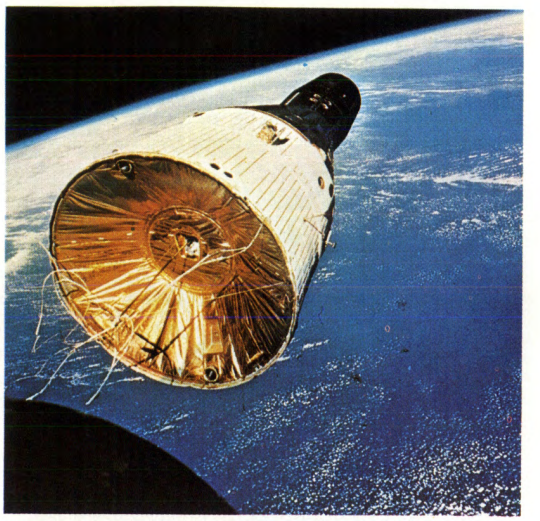
Apollo 15 Command and Service modules in lunar orbit with the lunar surface in the background, as photographed from the Lunar Module. The Scientific Instrument Module (SIM) bay can be seen on the front of the Service Module (SM).

(Right) The prime recovery helicopter hovering over the Apollo 16 spacecraft after splashdown on April 27, 1972, 11 days after launch. Navy swimmers in the life raft secure the flotation collar on the capsule.



Plates 2 and 3: By courtesy of National Aeronautics and Space Administration





The Gemini 7 spacecraft as photographed from Gemini 6 during rendezvous and station keeping manoeuvres. Gemini 7 was launched December 4, 1965, and Gemini 6 was launched 11 days later on December 15.

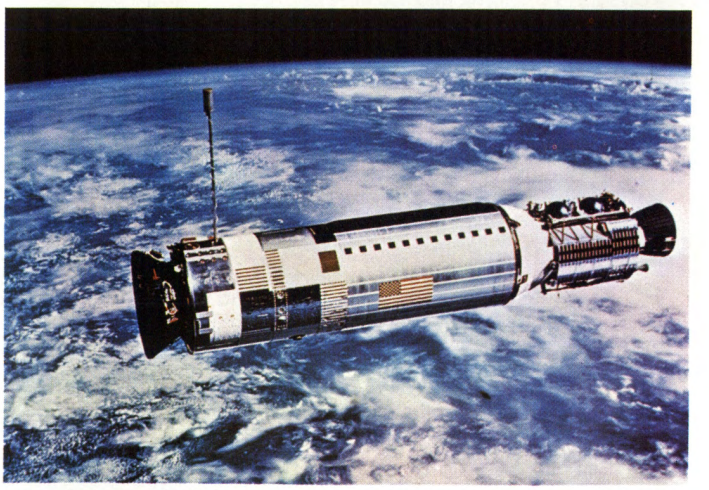
(Left) Gemini 10 blast-off from Cape Kennedy, Florida, on July 18, 1966. The photograph is the result of the combination of ten exposures.

Gemini program



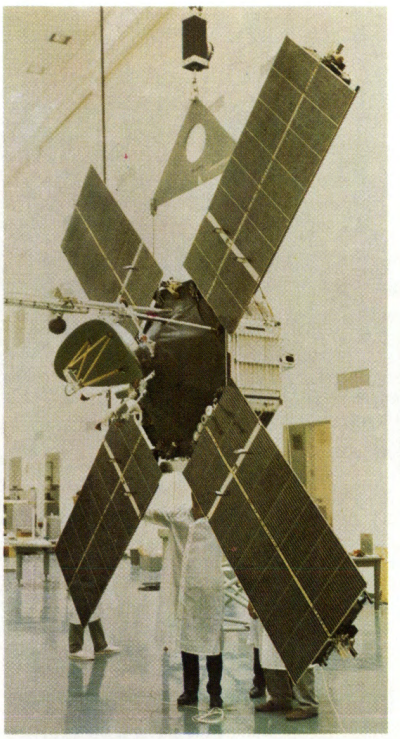
(Right) Mission Control Center, Manned Spacecraft Center, Houston, during the flight of Gemini 3, which was launched March 23, 1965.





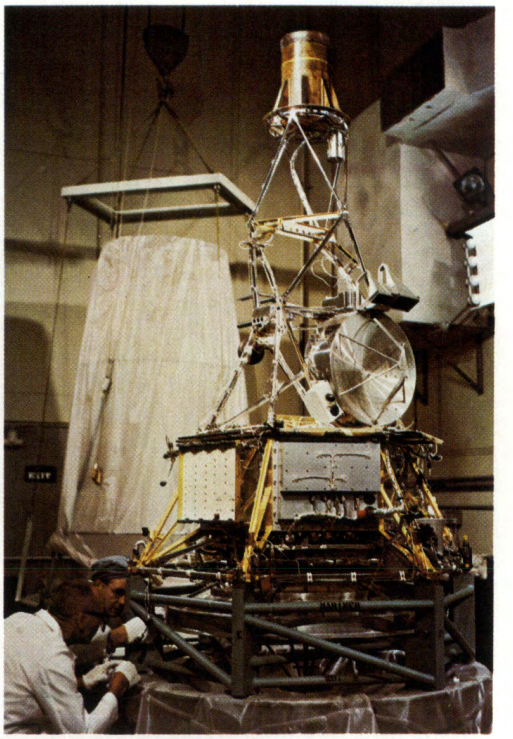
Agena, the target vehicle for the Gemini 12 rendezvous and docking, launched two hours before the Gemini spacecraft on November 11, 1966.

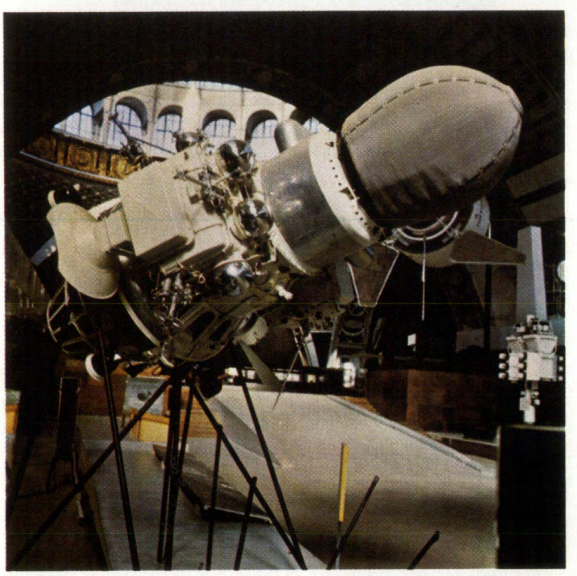
(Left) Gemini 4 astronaut Edward White during his 21-minute space-walk, or extravehicular activity. White is secured to the Gemini spacecraft by a 25-foot umbilical and tether line. He is holding a self-manoeuvring unit, which helped him to move about. Gemini 4 was launched June 3, 1965.



(Top left) The U.S. Mariner 4, which was launched November 28, 1964, and passed within 6,118 miles of Mars on July 14, 1965. It returned photographs of the surface of Mars and transmitted information on the Martian atmosphere.

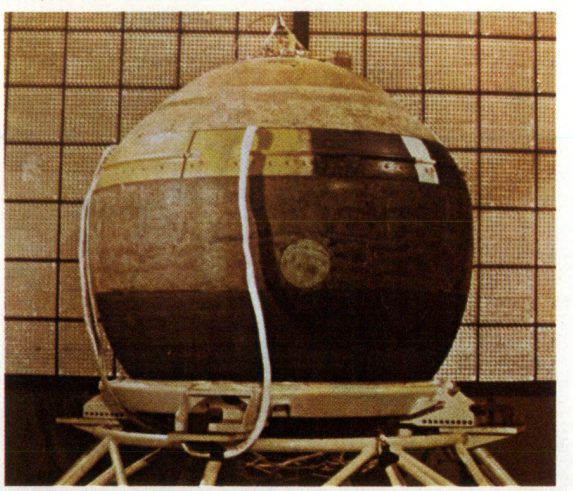
(Top right) The U.S. Mariner 5, which was launched June 14, 1967, and passed within 2,500 miles of Venus on October 19, 1967. It transmitted data from atmospheric and surface experiments.





The Soviet Luna 9, the first spacecraft to soft-land on the Moon. It was launched January 31, 1966, and returned photographs of the lunar surface for three days.

The Soviet Venera 4, the first spacecraft to soft-land on Venus. It was launched June 12, 1967, and transmitted information about atmospheric temperature and pressure and atmospheric composition.



Lunar and planetary spacecraft

(Right) Lunokhod 1, a Soviet remote-controlled lunar roving vehicle carried to the Moon by Luna 17, which was launched November 10, 1970. Lunokhod 1 travelled 6.5 miles on the lunar surface and relayed television pictures and scientific data.

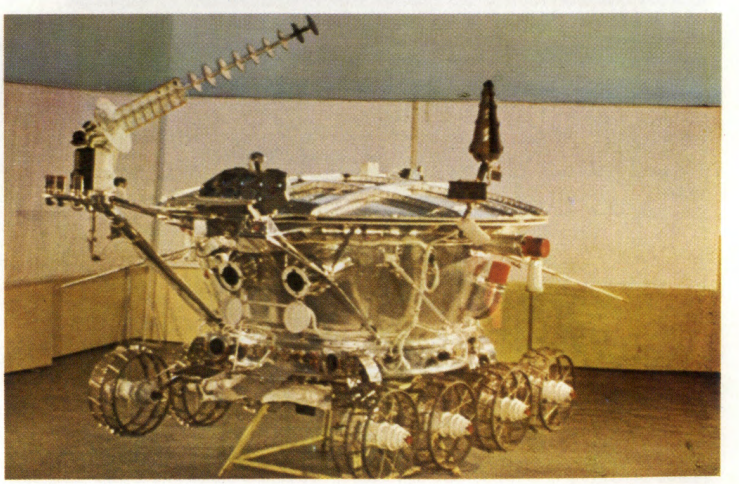
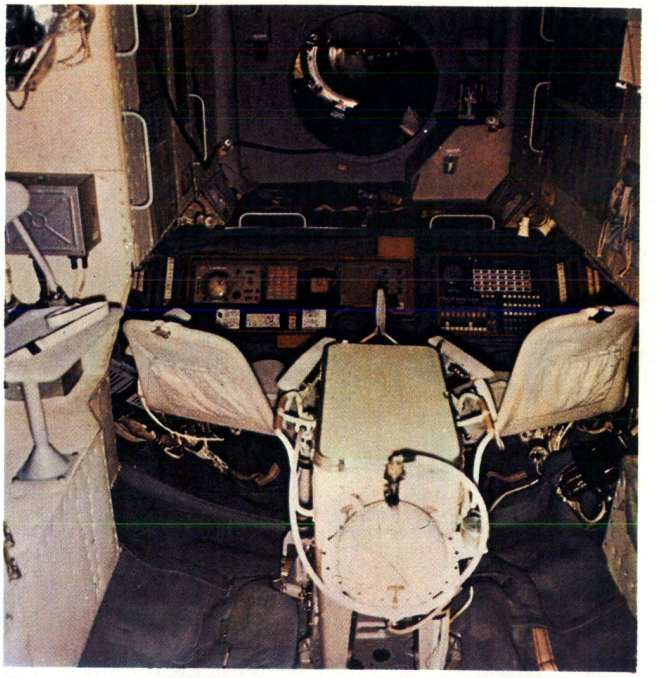


Plate 4: By courtesy of National Aeronautics and Space Administration

Plate 5: By courtesy of (top left, top right) National Aeronautics and Space Administration; photographs, (centre left) Novosti Press Agency, (centre right, bottom) Tass—Sovfoto



Vostok launch vehicle showing the two-stage rocket booster liquid propellant engines.

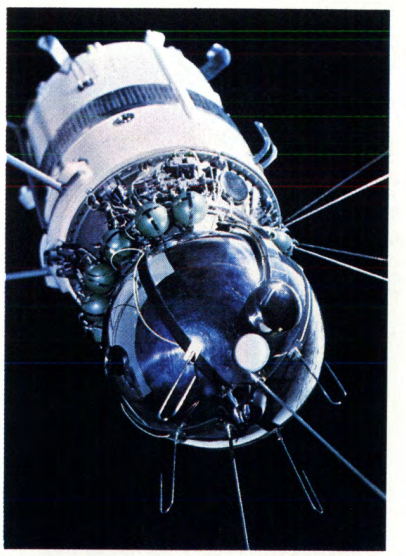


The main section and working compartment of the Salyut 1 space station. In the centre is the command and control panel and the seats for the crew. Behind the panel is the entrance to the tunnel connecting Salyut 1 with the Soyuz spacecraft. Salyut 1 was launched April 19, 1971.

Soviet manned and applications satellites

Molniya 1L, a television communications satellite launched April 11, 1969.



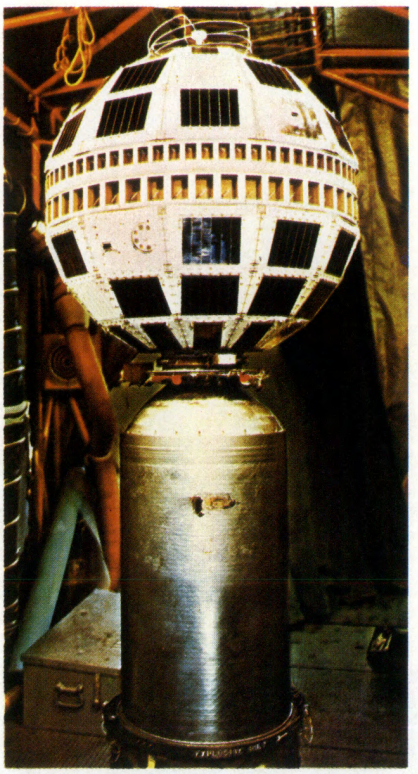


Vostok 6 spacecraft in which the first woman cosmonaut. Valentina Tereshkova, orbited the Earth for three days. Vostok 6 was launched June 16, 1963.

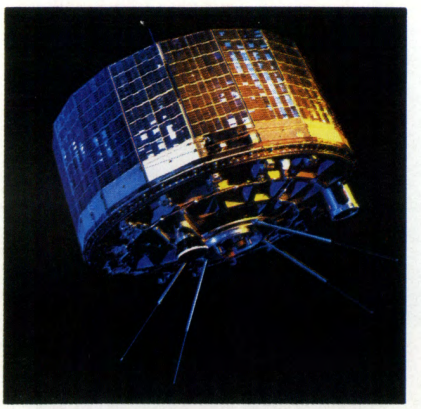


Soyuz 10 before being positioned for launch. Soyuz 10 was launched April 23, 1971, and docked with the Salyut 1 space station, which was launched four days earlier.

Plate 6: (All but centre left) Novosti Press Agency, (centre left) Sovfoto
Plate 7: By courtesy of National Aeronautics and Space Administration

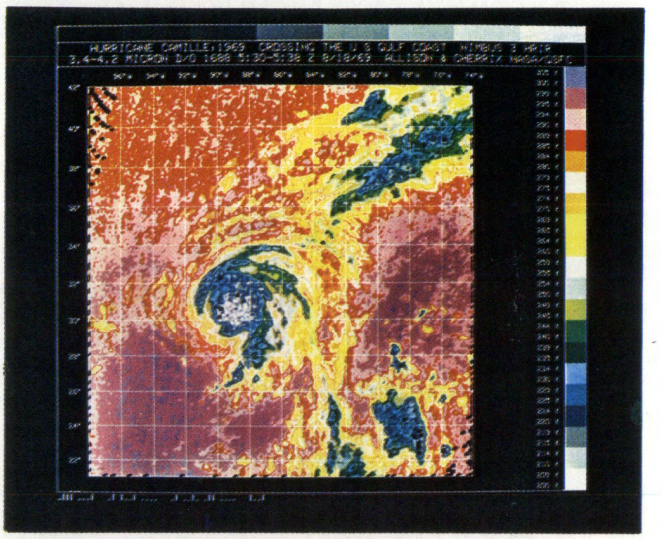


Telstar 1, launched July 10, 1962, relayed the first transatlantic television signals.

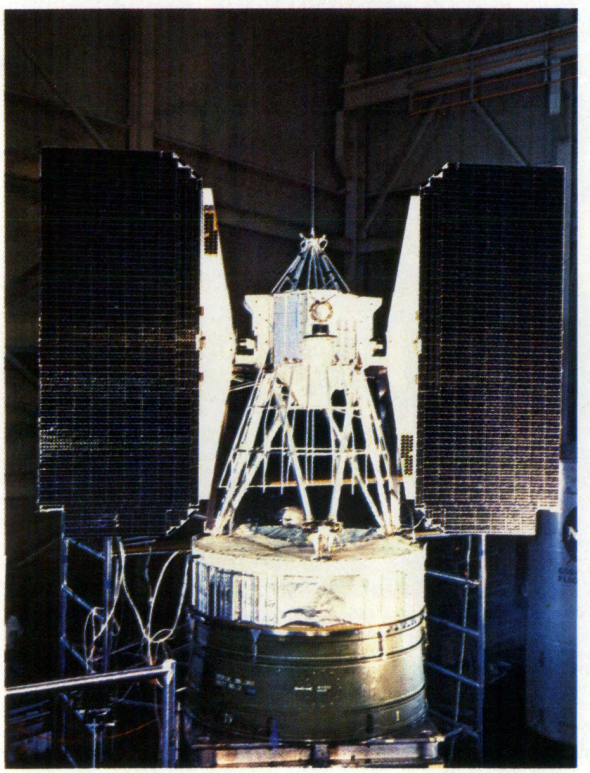


TIROS 7 (Television and Infra-Red Observation Satellite 7), launched June 19, 1963, transmitted cloud-cover photographs enabling meteorologists to track, forecast, and analyze storms.

U.S. applications satellites



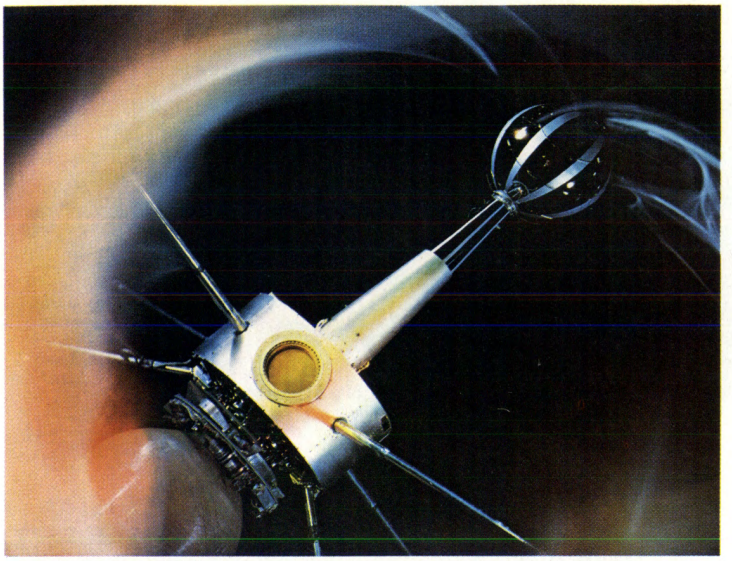
(Right) Photograph of Hurricane Camille that is the result of computer processing of data supplied by Nimbus 3 meteorological satellite. By observing hurricanes in colour, meteorologists can determine cloud heights using cloud temperatures. For example, ground or sea surface temperatures, or low clouds, appear purple; middle clouds appear yellow; high clouds appear blue; while the highest clouds, in the eye of the hurricane, appear gray. Nimbus 3 was launched April 14, 1969.



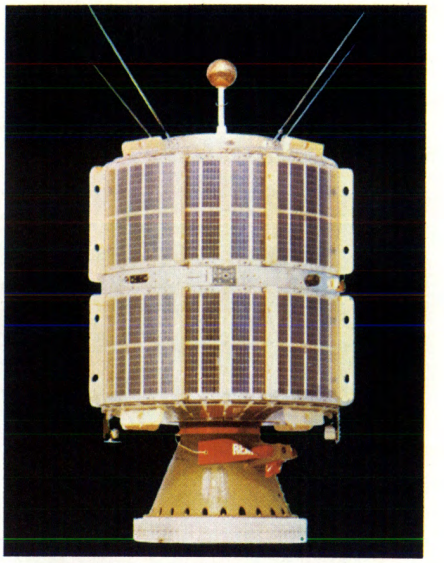


Composite photograph of the Olympia-Seattle-Tacoma, Washington, area transmitted from the Earth Resources Technology Satellite 1 (ERTS 1). Green, red, and infrared are recorded separately by the satellite and then combined to make the photograph. Healthy crops, trees, and other green plants appear bright red; barren land, cities, and industrial areas appear green or gray. Clear water is completely black and snow or ice shows as bright white. ERTS 1 was launched July 23, 1972.

(Left) Nimbus 1, launched August 28, 1964, transmitted cloud-cover photographs.



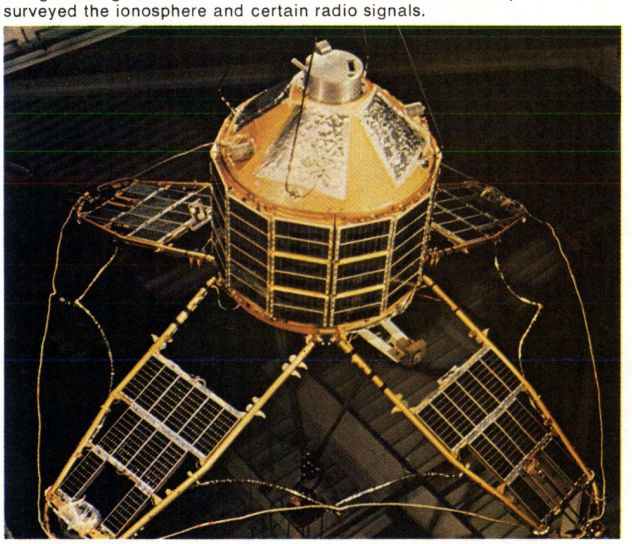
The U.S. Explorer 10 shown undergoing testing in a NASA (National Aeronautics and Space Administration) laboratory. Explorer 10 was launched March 25, 1961, and relayed magnetic field data.



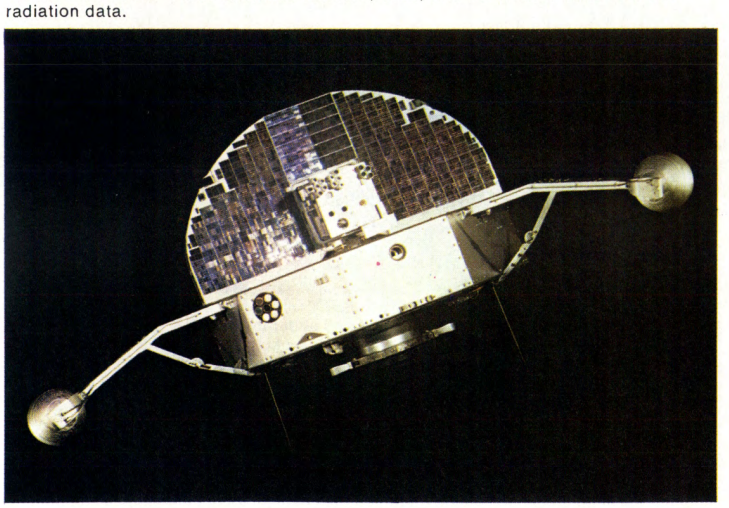
The European-designed and -built satellite Aurorae (ESRO 1A) that investigated the Aurora Borealis and other related phenomena of the polar ionosphere. Aurorae was launched October 3, 1968 by NASA.

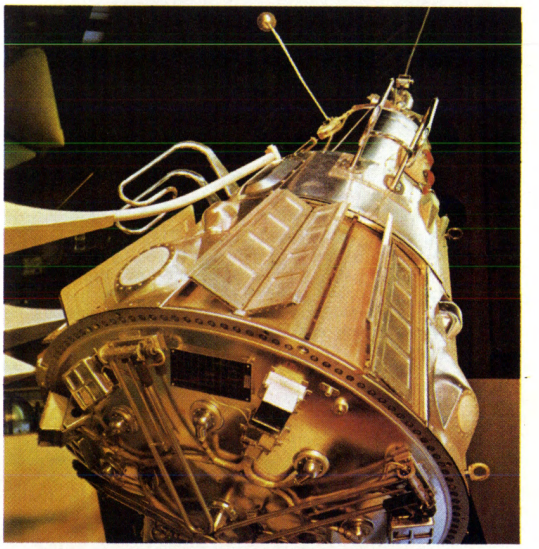
Scientific satellites and space probes

The British Ariel 4 (U.K. 4) shown suspended with its solar panels deployed in flight configuration. It was launched December 11, 1971, by NASA and surveyed the ionosphere and certain radio signals.

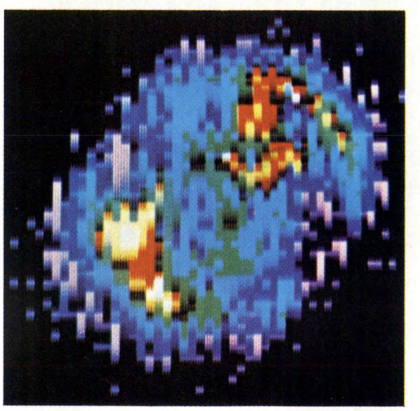


The U.S. Orbiting Solar Observatory 1 (OSO-1) shown in a simulated space environment. OSO-1 was launched March 7, 1962, and transmitted solar flare and radiation data.





The Soviet Sputnik 3, the first instrumented multipurpose space laboratory. It was launched May 15, 1958, and transmitted information on atmospheric composition and pressure, positive ion concentration, primary cosmic ray data, etc.



Photograph of a solar storm transmitted by the U.S. Orbiting Solar Observatory 7 (OSO-7). The white regions show the most intense storm activity, with yellow and red clusters denoting lesser activity. OSO-7 was launched September 29, 1971.

Plate 8: By courtesy of (all but centre right) National Aeronautics and Space Administration; photograph, (centre right) Tass—Sovfoto

one giant leap for mankind." He reported sinking approximately one-eighth of an inch into the fine, powdery surface material, which adhered in a thin layer to his boots. Aldrin joined Armstrong, and together they spent about two hours taking photographs, collecting about 47.7 pounds (21.7 kilograms) of lunar soil (including two core samples), planting a United States flag, and deploying a solar-wind experiment, a seismic experiment package, and a laser beam reflecting device. They were seen on Earth television by means of a camera erected some distance from the Lunar Module.

The astronauts made their lunar ascent stage takeoff (Figure 7) on July 21, concluding a total lunar stay time of 21 hours 36 minutes. Transferring to the Command Module, they relanded on Earth in the Pacific on July 24. Because of possible contamination by organisms with unknown effect upon terrestrial life, decontamination and quarantine were employed.

Apollo 11 decontamination

Apollo 12, launched November 14, followed a flight plan similar to that of Apollo 11. The Lunar Module landed on the Moon on November 19 in the Ocean of Storms. Two Moon walks were made; a nuclear isotope-powered surface experiments package was emplaced, 75.6 pounds (34 kilograms) of lunar samples were collected, and an inspection was made of the Surveyor 3 spacecraft that had landed 2½ years before. Parts of the Surveyor were brought back to Earth for study.

Apollo 14, launched January 31, 1971, continued exploration of the Moon. Landing within 60 feet of the target area, the astronauts made two surface explorations totalling more than nine hours. A surface experiments package was emplaced, and 94.4 pounds (43 kilograms) of lunar rocks were brought back.

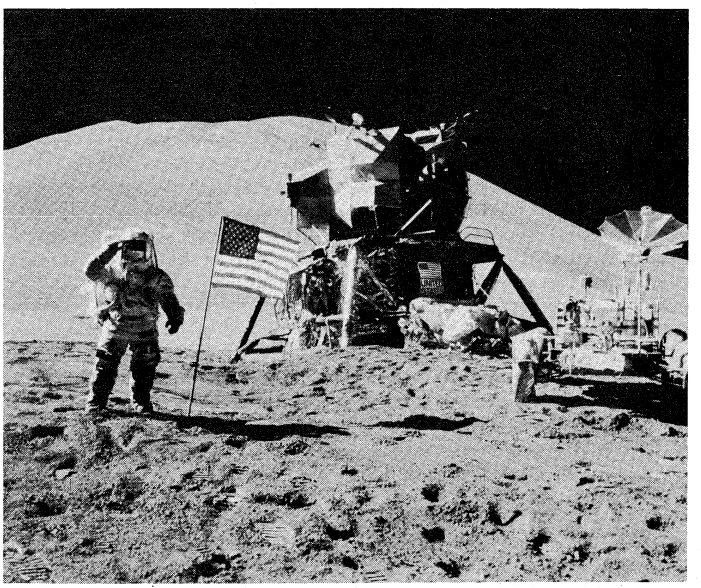


Figure 8: The Apollo 15 space mission's Lunar Module "Falcon," and the Lunar Roving Vehicle (right). Astronaut James B. Irwin is saluting the flag.
By courtesy of the National Aeronautics and Space Administration

Apollo 15 was launched July 26, 1971, and returned August 7, 1971. Landing on the Moon near Hadley Rille and the Apennine Mountains, it carried a four-wheeled, battery-powered Lunar Roving Vehicle that permitted extensive exploration (Figure 8). Weight of lunar material returned was 170 pounds (77 kilograms).

On April 16, 1972, Apollo 16 was launched. Landing in the Descartes region, it brought back to Earth some 215 pounds (98 kilograms) of lunar rocks and soil.

Apollo 17, the last U.S. mission to the Moon, launched December 7, 1972, broke previous records of lunar exploration time, distance travelled, and quantity of lunar samples returned (225 pounds [116 kilograms] from the Taurus–Littrow region).

Skylab and ASTP. Several additional manned missions to the Moon had been planned initially. With their cancellation, Saturn IB (Earth orbital) and Saturn V (lunar) launch vehicles, as well as Apollo Command and Service modules, became excess. This equipment was used for Skylab and the Apollo–Soyuz Test Project (ASTP).

Skylab, the first U.S. space station, was launched on May 14, 1973. Damage sustained during launching required an extensive and successful repair program carried out by the first two crews to visit Skylab. The station consists of four basic sections. The Orbital Workshop, a converted third stage (Saturn IVB) of Saturn V, includes a hydrogen tank divided into two floors to provide the basic living and work area for the crew. Space-suited astronauts may exit and re-enter the Skylab for work in space through the Airlock Module (AM). A Multiple Docking Adapter (MDA) is provided for docking the Apollo spacecraft to the Skylab, while an Apollo Telescope Mount (ATM) is a noninhabitable area containing experiments.

Between May and November 1973, three crews of three astronauts each lived and worked aboard Skylab for 171 days 13 hours. These missions contributed to understanding of the space environment and of the medical and biological implications of long-term space habitation.

The Apollo–Soyuz Test Project is the only cooperative U.S.–Soviet manned space mission. Launched on July 15, 1975, from their respective national launch complexes, the three-man Apollo and the two-man Soyuz 19 spacecraft rendezvoused and docked in orbit two days later by means of a U.S.-built docking module. The two spacecraft remained linked for two days while the astronauts and cosmonauts conducted joint Earth survey, astronomical, medical, and technical experiments. At the conclusion of the two-day joint mission, the Soyuz spacecraft returned to Earth while the Apollo crew remained in orbit for an additional five days. The ASTP demonstrated the possibilities of international cooperation in space as well as the feasibility of using the docking module as a rescue device.

Apollo–Soyuz Test Project

APPLICATIONS SATELLITES

The search for scientific knowledge, exploration of the unknown, and establishment of man's capability in space have been, and remain, primary goals of national space programs. Increasingly, however, Earth-oriented satellites (called applications satellites) of direct economic benefit to man are being placed in operation.

There are three general classifications of such satellites: communications, Earth survey, and navigation.

Communications satellites. In the decade following the passive reflector balloon-satellite Echo 1, communications satellites (comsats) became a significant part of global communications. The Intelsat 4 comsat, for example, launched in January 1971, was capable of handling 3,000–9,000 telephone circuits or 12 colour television channels, or a combination of both.

After 1975 improved Intelsat IVA satellites began to replace the Intelsat IV spacecraft in the international network. In addition, domestic comsats, such as Comstar, Satcom, and Westar in the U.S. and Anik in Canada, now meet the specific needs of particular nations. Voice, television, facsimile or high-speed data (or both) are relayed. Palapa 1 is a comsat bought by the government of Indonesia and launched in 1976 to provide 5,000 two-way telephone circuits aimed at linking the almost 1,000 inhabited islands of the Indonesian archipelago.

Other specialized communications satellites in service since 1976 include three Maritime Satellite systems (Marisats), designed to provide greatly improved ship-to-shore communications. Previously, 90 percent of ship-to-shore communications was by hand-operated radiotelegraphy, and ionospheric disturbances in poor weather often limited communications. Marisat provides high-fidelity voice, data, Telex, and facsimile services. In addition, emergency rescue operations would be speeded by immediate connection to rescue authorities.

Marisats

Two developmental Symphonie communications satellites, launched cooperatively by France and West Germany in 1974–75, have been placed in geosynchronous orbit to provide telephone, television, and data service. (See also SATELLITE COMMUNICATION.)

The United States Department of Defense has a "secure" satellite communications system, using both synchronous and lower altitude orbits. The United Kingdom has a military Skynet comsat system launched by the U.S.

Soviet Molniya comsats

The Soviet Molniya comsats operate in a 65° eccentric orbit with a 300-mile perigee over the Southern Hemisphere and 24,500-mile apogee over the Northern Hemisphere. This technique provides a longer time of eight to nine hours over the portion of the Earth visible to Soviet ground terminals. Thus, three of these satellites, properly spaced, make possible 24-hour coverage.

Applications Technology satellites (ATS) are a series of geosynchronous orbit satellites launched by the United States, their purpose being to test new instrumentation and advanced communications techniques such as higher transmission frequencies and improved transmission antennas. ATS 6, placed in orbit in 1974, explored the possibilities of direct broadcast communications from space through the use of inexpensive ground receivers. This geostationary satellite was used to conduct experiments involving the provision of improved communications services for isolated communities. Direct broadcast classroom and community educational programming was provided initially for students in Appalachian and Rocky Mountain regions. The potential value of the satellite for improving health services was demonstrated by providing an ATS 6 communication link between rural Alaskan clinics and medical advisers in major hospitals in the Northwest. The following year the satellite was moved around the Equator to a point above East Africa, from where a year-long experiment was performed in cooperation with the Indian government. Indian programs on such subjects as agriculture, health, hygiene, and family planning were broadcast to more than 2,400 remote villages and relayed to another 2,500 villages throughout India. In 1976 the ATS 6 was again positioned above the Equator to broadcast educational programs to the U.S. NASA's Communications Technology Satellite (CTS), a single experimental vehicle placed in orbit in 1976, has been used in a joint U.S.–Canadian project to explore new communications technologies utilizing higher transmission power at higher frequencies.

Earth survey satellites. This broad category of satellites has progressed from initial relatively simple photography to include many forms of observation of the Earth.

Weather. Meteorological, or weather, satellites operate globally. From a polar orbit, photographs of the Earth's surface are taken by television scanning and transmitted to Earth receiver stations. Since the Earth rotates constantly, total global coverage is possible. In addition, meteorological satellites in geostationary orbit have the advantage of continuous watch over large regions of the Earth. Seventy percent of the Earth is covered by water, and there are few observation stations in desolate and jungle areas. Global observation of weather and its movement by satellites has become important to all nations. Tracking of hurricane formation and movement has permitted advance warning, saved numberless lives, and minimized property damage.

The first meteorological satellite was TIROS—Television and Infra-Red Observation Satellite. TIROS satellites transmitted cloud-cover pictures enabling meteorologists to track, forecast, and analyze storms.

Meanwhile, NASA was developing meteorological satellites of greater sophistication in the Nimbus series. Nimbus 3 and 4, launched in 1969 and 1970, were designed to measure atmospheric water-vapour content and vertical temperature sounding of the atmosphere. These measurements enabled meteorologists to draw an accurate picture of vertical temperature and humidity profiles, making possible more accurate forecasts.

Transmission of cloud pictures

By 1972 more than 500 Automatic Picture Transmission stations were active in some 50 nations, each capable of receiving facsimile-printed transmissions of an area of about 1,700 square miles. Cloud pictures may be processed for retransmission by geostationary satellites to local ground stations. Transmission of infrared nighttime images, which indicate ground, water, and cloud-top temperatures, was begun in November 1970.

In late 1970 the first operational "second generation" weather satellite was launched. Twice the size and weight of TIROS, it had an important new instrument, scanning radiometers to obtain direct-readout and stored images of the Earth's night side. It provides global coverage at 12-hour rather than 24-hour intervals.

The world's first Synchronous Meteorological Satellite, SMS 1, was launched by the U.S. in May 1974. Following the launching of SMS 2 in February 1975, a Geostationary Operational Environmental Satellite (GOES) was placed in orbit in October to provide 24-hour coverage of U.S. weather. In addition, the SMS and GOES craft collect and relay information transmitted by unmanned, ground- or ocean-based sensor stations in isolated locations. They also contain instruments to study solar activity, X-radiation, and the Earth's magnetic field. The SMS test satellites and the operational GOES take part in the Global Atmospheric Research Program (GARP), an international project aimed at achieving improved understanding of the mechanisms that govern the world's weather.

Other elements in this worldwide system include the Japanese Geostationary Meteorological Satellite (GMS) and the European Space Agency's Meteosat, launched in 1977. Soviet weather satellites, called Meteor, have distributed weather photographs internationally since 1969. The completed system is expected to provide continuous global coverage of the weather.

Geodesy. Earth satellites offer a method of improving the world's maps. If the altitude of a satellite is known, the angles of simultaneous widely separated observations may be used to calculate a base line between the two points of observation. Satellite observations have made possible the "tying together" of landmasses of the world not hitherto possible.

Geodetic satellites

NASA launched its first Geodetic Earth Orbiting Satellite (GEOS 1) in 1965; it was followed by GEOS 2, carrying a variety of equipment. In 1970, 43 stations in 23 countries made observations. The aim was to provide a unified three-dimensional framework connecting all landmasses of the world within an accuracy of 33 feet (10 metres). In 1971 an International Satellite Geodesy Experiment program using GEOS 2 included most major nations.

By the use of laser beam techniques, extremely precise measurements may make it possible to record Earth movements associated with earthquakes. This may make possible future forecasting of these natural phenomena. Lageos (Laser Geodynamic Satellite), launched in 1976, was a new approach to the study of geodetics. A completely passive satellite, Lageos is essentially a brass core encased in an aluminum sphere covered with 426 retroflectors that allow ground-based laser ranging stations to take measurements for the study of continental drift, tides, and fault motions.

Lageos was placed in a very stable orbit 3,600 miles (5,793 kilometres) high, where it is predicted it will remain for some 8,000,000 years.

Earth resources satellites. Many other benefits may be derived from Earth survey satellites. By viewing the Earth's surface in different portions of the radiant spectrum, revealing images are obtained.

The U.S. launched the first Earth Resources Technology Satellite (ERTS) in 1972. In polar orbit at an altitude of about 560 miles (910 kilometres), it transmitted to the Earth multispectral images that provided data to hundreds of scientific investigators in many disciplines: agriculture, forestry, mineral and land resources, land use, and water and marine resources.

ERTS 1 was renamed Landsat 1, and in January 1975 Landsat 2 was orbited successfully. The two spacecraft are 180° apart in orbit, thereby providing a view of the same local area with the same sun angle every nine days. With only one satellite, 18 days would be needed for the same view.

More than 100,000 black and white pictures and photographs in strange, artificial colours produced from the multispectral scanning system were taken by Landsat I. Based on those pictures, more than 300 separate investigations were conducted, providing better understanding

of the Earth's surface features. Some of the most valuable uses of Landsat 1 have been estimating crop acreage; monitoring urban development and planning future land use; locating air and water pollution; mapping strip-mine and forest-fire scars; locating geologic formations that may indicate the presence of minerals and petroleum; updating maps and navigation charts; monitoring the advance of glaciers; and studying flood hazards and managing water resources.

Military reconnaissance satellites. As of 1977 the U.S. and the Soviet Union were the only nations with both the technological ability and the wealth to utilize satellites for photographic and electronic monitoring. Such programs have been conducted by both nations since about 1959; in fact, reconnaissance is the oldest use of the applications satellites. Although there has never been official disclosure of such programs by either country, the use of reconnaissance satellites continues by tacit agreement. Such satellites are credited with being the main technological base in nuclear-arms control under the terms of the Strategic Arms Limitations Talks (SALT).

The U.S. in the 1960s began use of two other types of military detection satellites, launched in pairs in the Vela program. One satellite detects the launching of large missiles by the infrared radiation emitted, the other detects above-ground nuclear fission and fusion explosions from the gamma rays emitted.

Navigation. This specialized class of Earth-oriented satellites, first launched by the United States Navy in 1960, originally was classified secret. The prime purpose of the Transit series of satellites was to enable nuclear submarines accurately to fix their position under all surface weather conditions. It made star sighting unnecessary. In essence a beacon in the sky, the system was revealed in 1967. Operationally the system is based upon observation of the Doppler shift of the satellite's transmitted signal. A major military advantage is the fact that a ship does not have to emit an electromagnetic signal. Three Transit satellites were launched in 1967 and another in 1968. Use of this navigation system by merchant ships is limited because of the requirement for expensive equipment.

Transit satellites

The Soviet Union has also announced an operational navigation satellite system. Certain of the Cosmos satellites have flown in ways that suggest that they are navigation satellites similar to the Transit system. In accordance with Soviet policy, no specific information has been released.

IV. Contributions to the physical sciences

Dealt with in this section are selected examples of new understanding resulting from the exploration of space—by rockets, satellites, and space probes.

Discoveries about the Earth. Before the advent of artificial satellites it was accepted that the shape of the Earth was, except for some high mountains and deep valleys, an ellipsoid; *i.e.*, a sphere, slightly flattened. Accurate observation of variations of the orbital path of Vanguard 1 proved that, on the contrary, the Earth is slightly pear-shaped, the distance from the Earth's centre to the North Pole being greater than the distance to the South Pole. Further observations of anomalies in the orbits of satellites have shown that the Equator is elliptical and that the surface gravity of the Earth has several "hills" and "valleys" varying as much as 100–300 feet from the average.

Originally it was thought that cloud patterns and weather systems were relatively small, perhaps a few hundred miles across. Photographs from space have shown that the Earth's cloud cover is completely organized globally. Coherent cloud cover systems have been found to extend for thousands of miles and to be related to other systems of similar dimensions. It was discovered that weather systems could be directly identified by their cloud structure, and thus it was possible immediately to locate important atmospheric phenomena such as fronts, storms, and hurricanes, and to chart their courses daily with accuracy.

Atmospheric density and variation with altitude above the Earth had been studied for many years by balloons and to some extent by sounding rockets. Observations of atmospheric drag on satellites and resultant lowering of orbits, however, revealed that variations in atmospheric density are related directly to solar flares and the impact of electrically charged particles about 26 hours after marked solar activity. Further, solar flares have been found to increase the temperature of the atmosphere by interaction of short-wavelength radiation atoms and molecules in the atmosphere. Recombination of the ionized particles is accompanied by the liberation of heat.

Although it had been postulated that some charged particles might be trapped in the Earth's magnetic field, Explorers 1 and 4 revealed a major phenomenon in geophysics: the existence of two permanent doughnut-shaped belts of high-energy radiation caused by trapped charged particles, 2,000–10,000 miles (3,000–16,000 kilometres) above the Earth.

Discovery of high-energy radiation belts

The earlier view of the ionosphere was that from 40–400 miles (60–600 kilometres) distinct layers of ionized atoms and molecules, known as the D, E, and F layers, existed and were necessary for radio communications. Evidence from space exploration is that ionization in the atmosphere is actually continuous but that peaks of concentration of electrons occur at altitudes where the existence of layers had formerly been postulated. Still higher, ionization is practically complete and extends to great altitudes.

Previously it was believed that the magnetic field of the Earth was similar in shape and symmetry to a short bar magnet, with the magnetic field extending outward into space. It has been learned from satellites that the Earth's magnetic field on the side toward the Sun is sharply bounded because of the solar wind. On the dark side of the Earth, in the direction away from the Sun, the field streams out in a long, thin tail beyond the orbit of the Moon.

Solar phenomena. Since ultraviolet rays and X-rays are absorbed in the atmosphere, measurements of these radiations from the Sun were made by sounding rockets in some of their first flights. Solar spectra have been continuously refined by spectrographic cameras carried aboard many satellites. Variation in measurements correlated with observed solar activity is providing new knowledge of the processes taking place within the Sun. Although the radiation of the Sun in the visible portion of the spectrum is constant, the ultraviolet and X-radiation are variable. From the information gathered by space exploration it also has been learned that the Sun radiates in the manner of an incandescent gas, that ultraviolet and X-radiation are generated not in the region of the solar disk but higher in the solar atmosphere, in the chromosphere and the corona.

Previously the solar wind, a constant outpouring of high-speed particles from the Sun, had been only postulated. Measurements from spacecraft and from experiments made on the Moon by Apollo astronauts have measured the number and rate of flow of these particles. The velocity of the solar wind is highly variable and ranges from 250 to 310 miles (400 to 500 kilometres) per second.

Interplanetary phenomena. The earlier view was that interplanetary space was empty of matter except for that thrown out from the Sun and for the clouds of particles that became visible as meteoroids on striking the Earth's atmosphere at night. From space exploration it has been learned that the solar wind flows continuously and at such velocities that the particles escape the solar system.

Earlier it was believed that the existence of electric fields in space was unlikely. The evidence based upon measurements in space is that such electric fields do exist.

The earlier view of the number of meteoroids in space was based mainly on conjecture. Actual measurements by Explorer 16, Pegasus, and other satellites have given actual data and determined the number of meteoroids (and their hazard) to be considerably less than expected.

Meteoroid measurements

Exploration of the Moon and planets. Detailed photographs of the Moon have made possible a nearly complete map, including the hidden far side. Soft-landing vehicles have televised views of the surface by remotely

controlled cameras, determined the physical and load-bearing characteristics of the soil, and performed chemical analysis by isotope back-scatter technique. Seismometers, laser reflectors, solar wind and ion detectors, charged particle measuring devices, magnetometers, and other data-gathering instruments have been set up. Six Apollo missions brought back 841 pounds (382 kilograms) of selected lunar rocks and soil for study.

The physical nature of the Moon has been revealed by space exploration in a way impossible before. The list of discoveries is long and detailed. Perhaps the most significant is that the age of the surface material studied has been determined to be 4,600,000,000 years. For much of that time the Moon has not drastically changed, but lunar history appears to have been complex. Because it has changed so little, a record of the Sun can be traced on the lunar surface. This type of data has many implications for Earth history. Most rocks found on Earth are less than 50,000,000 years old because of destruction, erosion, covering, and other processes.

Some of the other findings about the Moon are: no evidence of life or water; lunar rocks are similar to Earth rocks but have different chemical compositions; seismometers record numerous "moonquakes" about the time the Moon is closest to the Earth each month; there is a lunar magnetic field, but it is much weaker than the Earth's; the solar wind is not disturbed as it approaches the Moon, signifying that there is no layer of charged particles similar to the Earth's ionosphere; laser reflection makes it possible to measure Earth–Moon distance to within about nine inches.

Through Soviet and United States spacecraft travelling to the vicinity of Venus, much greater understanding has been obtained of this planet. The Soviet technique of parachuting specially protected capsules through the dense Venusian atmosphere, apparently to the surface, has confirmed cloud-covered surface temperatures approaching 1,000° F (540° C). The thick atmosphere is primarily carbon dioxide, probably more than 95 percent; water vapour, about 0.4 percent.

Data on Venusian atmosphere

Prior to 1964 it generally had been accepted that the atmospheric pressure on Mars is about 85 millibars. (Terrestrial atmospheric pressure is about 1,000 millibars.) Mariner 4 data indicated that Martian atmosphere was about one-tenth this pressure, an important factor in the design of Martian landers.

U.S. Mariner spacecraft have flown within 860 miles (1,380 kilometres) of Mars, photographing the polar caps and revealing some widespread cratering, not unlike that of the Moon. There were, however, flat featureless plains, possibly covered by dust or sand. The polar caps are crusted with frozen carbon dioxide possibly to a thickness of two feet (0.6 metre). The Martian atmosphere is composed almost entirely of carbon dioxide with only traces of water. Unlike that of the Earth, Mars's atmosphere does not shield it from ultraviolet radiation, and most of it reaches the surface. Since intense ultraviolet radiation destroys most molecules associated with life—as mankind knows it—Mars may not support living things.

Scientists will be studying the data returned from Vikings 1 and 2, which sent landing vehicles to the Martian surface, for many years to come. Already, however, these spacecraft have enormously expanded our understanding of conditions on Mars. The atmosphere of the planet is primarily composed of carbon dioxide, but traces of nitrogen, argon, krypton, xenon, and oxygen have also been identified. Humidity varies significantly, the summer hemisphere being more humid than the winter. Viking Orbiter data indicate that the polar regions are composed of layered sheets of water ice. The biological experiments have so far failed to uncover any evidence of life on the surface, nor has the presence of organic chemicals been detected. Some test results have indicated the possibility of unusual soil chemistry.

Astronomical phenomena. The earlier view was that the temperatures of stars could be calculated from the visible radiations that penetrate to the surface of the Earth. From space research it appears that inclusion of the ultraviolet radiation causes significant changes in the calculated temperatures of individual stars. Some stars are found to be cooler than previously believed, and a class of supergiant stars are hypothesized to be losing mass into the interstellar medium at a greater rate than was previously thought.

The earlier view was that much of the mass of the Milky Way galaxy, containing the solar system, was molecular hydrogen. The evidence obtained from spacecraft is that the quantity of molecular hydrogen is much less than was thought and undetectable with available instruments.

It was thought originally that stars radiated most of their energy in the visible regions of their spectra. The evidence from spacecraft is that in the case of some stars the energy output in the visible spectrum is exceeded by that in the ultraviolet, radio frequencies, or infrared, or all three. Such tremendous outpourings of energy have caused theories of stellar evolution to be reconsidered. At the least, they suggest that there may be sources of energy that far exceed man's current knowledge of nuclear energy technology.

Radiation outside the visible spectrum

Since many of the stellar distances are dependent upon determining the intensity of stellar sources, these discoveries of much greater total energy outputs may cause astronomers to double the estimated physical size of the known universe. New and different types of stellar sources, such as gamma and X-rays, are being discovered at a rapid rate. Astrophysics poses one of the most challenging fields of professional endeavour as a result of the new and puzzling data being obtained.

V. Proposed future missions

A logical extension of space exploration and space exploitation (by applications satellites) may be expected. Many programs are possible, but funds and decisions to conduct new missions are controlled by national and international political, sociological, and economic factors. Major disasters such as climatological changes or major wars could severely limit progress in space activities.

Planetary missions. In 1978, Pioneer 11 was heading for an encounter with the rings of Saturn, scheduled to arrive in September 1979; it was expected to obtain the first detailed information and photographs of this planet. Voyagers 1 and 2 will commence studies of Jupiter in early 1979 and of Saturn in 1980.

Pioneer Venus 1, launched on May 20, 1978, will spend one year circling Venus, conducting surface mapping studies with radar, and obtaining data on the atmosphere of the planet. Pioneer Venus 2, scheduled to be launched in August 1978, will dispatch a series of three small probes and one large one that will descend through the atmosphere, relaying data before striking the surface.

The Voyager mission to Jupiter and Saturn may be followed by a Jupiter Orbiter Probe spacecraft that will circle the planet and dispatch small probes through the layers of cloud that shroud it.

The success of the Viking mission may lead to more ambitious programs of exploration for Mars, perhaps in 1984. These will almost certainly include a Mars rover capable of moving over the surface and transmitting data to Earth for a long time. Under consideration is the possibility of return of Martian surface samples by 1990, in a manner similar to that used by the Soviet Luna 16.

The exploration of the outer planets of the solar system may begin with a Voyager flyby of Uranus and will undoubtedly be continued by later spacecraft. Other possible missions may include visits to comets and asteroids and to the large moons of Jupiter and Saturn.

For a mission nearer to Earth, a Lunar Polar Orbiter (LPO) is being considered for launching in the 1980s. The five lunar science surface packages left by Apollo astronauts were shut down in 1977. The LPO is of interest for global observations of the Moon and as a prototype for spacecraft to survey other planets.

Lunar Polar Orbiter

An ambitious study has been made of a possible spacecraft rendezvous with Halley's Comet during its next passage close to the Earth in 1985–86, a technically

difficult mission. Launching might take place in 1982, and ion propulsion might be used for the unique trajectory. Since the comet moves in a highly inclined retrograde orbit, the spacecraft would have to move to a point along the comet's orbit, slow to a stop, and then proceed along the orbit, regaining speed to match that of the comet.

Manned space programs. The American space program will enter a new era with the first operational flights of the reusable Space Shuttle in 1980. The shuttle will be launched vertically by two large solid-propellant strap-on boosters and three liquid-propellant main engines. The liquid oxygen/liquid hydrogen propellants are fed to the main engines from a large external tank that is jettisoned when empty.

Reusable Space Shuttle

The Orbiter is the central unit of the Space Shuttle system. A winged vehicle, it is launched vertically but re-enters the atmosphere and glides to a landing like an airplane. Five successful glide landings were made in 1977 by the protoype Orbiter, "Enterprise."

The Orbiter can be reflown for up to 100 missions. The boosters are also recovered and refurbished to be flown again. The significant savings in cost, compared to standard launch vehicles that are expended during the course of a single flight, will enable NASA planners to schedule more frequent missions. By the mid-1980s, shuttle launchings may become weekly occurrences.

The central portion of the Orbiter contains the cargo bay in which satellites or other materials can be carried into space. A Spacelab being constructed by a nine-nation European consortium for the European Space Agency will be among the early payloads launched aboard the shuttle. The Spacelab makes use of an arrangement of a module, where a life-sustaining environment is maintained, and two instrument pallets. A variety of experiments is planned, ranging from Earth observation to materials processing, communications, navigation, and space sciences. Like the shuttle, the Spacelab will be reusable. Duration of missions will average seven days, although extensions of up to 30 days will be possible.

An Interim Upper Stage (IUS) is also being developed for use with the Space Shuttle. This large solid-propellant rocket motor will be used to boost payloads, such as synchronous communications satellites, into higher orbits than those that can be reached by the shuttle. Thought is also being given to the use of the IUS for the Jupiter Orbiter Probe being proposed for 1982.

The shuttle will make it possible to undertake large-scale construction projects in space. The large external propellant tank can be carried into orbit to serve as the nucleus for a space station. The Space Shuttle will provide a reasonably economical means of voyaging into space regularly.

Applications satellites. Because this class of satellites is of direct economic benefit to mankind, all varieties may be expected to increase in number, size, and sophistication of design. Of particular interest is the concept of direct broadcast satellites which could permit direct reception of television broadcasts from space. Initially the broadcast power of communications satellites was limited, requiring large dish antennas for reception. Large comsats, however, pioneered by the U.S. ATS 6, have demonstrated that direct satellite transmission to home television receivers is possible. Many political and sociological factors are obviously involved in such a development.

Direct broadcast satellites

Beyond 1985. During the 1970s there was active speculation on possible construction of large space structures for a variety of purposes, such as generating electric power from solar radiation and beaming the power to receiving stations on Earth. Large solar reflectors have been proposed that could illuminate populous areas of the Earth at night, reducing requirements for street and road lighting. Powerful communication relay stations in orbit may make possible personal "wristwatch" telephones. Because of weightlessness in orbit, and the presence of high vacuum and wide temperature variation, pure and applied research could be carried out as well as unusual varieties of manufacturing processes.

Human beings may be expected, by their nature, to continue outward exploration of the solar system. Returning to the Moon, a permanent lunar base and even a permanent colony could be established. Manned expeditions to the moons of Mars as well as to the Martian surface, a two-year journey, are possible. There has even been speculation on the establishment of permanent colonies, or habitats, in space—perhaps at the libration points (L-4 and L-5) of the Moon. These possibilities are based upon extrapolation of known technologies. New inventions and discoveries in applied science, such as controlled fusion power, may be anticipated.

Predictions cannot be made with confidence. Some of the space missions suggested above, however, or even flight to the stars, although not considered possible or feasible with today's technology, may occur in the future. History suggests that it is unwise to state categorically that they cannot happen.

BIBLIOGRAPHY

Historical and general background: WILLY LEY, *Rockets, Missiles, and Men in Space* (1968); W. VON BRAUN and F.I. ORDWAY III, *History of Rocketry and Space Travel,* 3rd rev. ed. (1975); ARTHUR C. CLARKE (ed.), *The Coming of the Space Age: Famous Accounts of Man's Probing of the Universe* (1967); DAVID IRVING, *The Mare's Nest* (1964); REGINALD TURNILL, *The Observer's Book of Manned Space Flight* (1977); A.M. WORDEN, *Hello Earth: Greetings from Endeavour* (1974); W.S. BAINBRIDGE, *The Spaceflight Revolution: A Sociological Study* (1976); K. GATLAND, *Robot Explorers* (1972), (with PHILIP BONO) *Frontiers of Space,* rev. ed. (1976), *Manned Spacecraft,* 2nd rev. ed. (1976); T.M. WILDING-WHITE, *Jane's Pocket Book of Space Exploration* (1976).

Popular and technical: ARTHUR C. CLARKE, *The Promise of Space* (1968), *Man and Space* (1964), *Voices from the Sky* (1965); W. VON BRAUN, *Space Frontier,* rev. ed. (1971); E.M. CORTRIGHT (ed.), *Exploring Space with a Camera* (1968), *Apollo Expeditions to the Moon* (1975); PATRICK MOORE, *Space* (1968); MITCHELL R. SHARPE, *Living in Space* (1969); O.W. NICKS, *This Island Earth* (1970); C.MCL. GREEN and M. LOMASK, *Vanguard: A History* (1971); WILLY LEY, *Events in Space* (1969); L.S. SWENSON, JR., JAMES M. GRIMWOOD, and CHARLES C. ALEXANDER, *This New Ocean: A History of Project Mercury* (1966); H.S.F. COOPER, JR., *A House in Space* (1976); P.J. KLASS, *Secret Sentries in Space* (1971); B.M. FRENCH, *The Moon Book* (1977); L.F. BELEW (ed.), *Skylab: Our First Space Station* (1977); R.O. FIMMEL, W. SWINDELL, E. BURGESS, *Pioneer Odyssey,* rev. ed. (1977).

Apollo moon landings: JOHN N. WILFORD, *We Reach the Moon* (1969); RICHARD S. LEWIS, *Appointment on the Moon,* rev. ed. (1969), *The Voyages of Apollo* (1974); N. ARMSTRONG, *et al., First on the Moon* (1970).

Biographical: MILTON LEHMAN, *This High Man* (1963), biography of Robert H. Goddard; ERIK BERGAUST, *Reaching for the Stars* (1960), biography of Wernher von Braun; SHIRLEY THOMAS, *Men of Space,* 8 vol. (1960–68); MICHAEL COLLINS, *Carrying the Fire* (1974); W. CUNNINGHAM and M. HERSKOWITZ, *The All American Boys* (1977).

Technical: S. GLASSTONE, *Sourcebook on the Space Sciences* (1965), *Book of Mars* (1968); PAUL E. PURSER, MAXIME A. FAGET, and NORMAN F. SMITH (eds.), *Manned Spacecraft* (1964); W.R. CORLISS, *Space Probes and Planetary Exploration* (1965), *Scientific Satellites* (1967); H.O. RUPPE, *Introduction to Astronautics,* 2 vol. (1966–67); R.C. HALL, *Lunar Impact* (1978).

Speculative: WALTER SULLIVAN, *We Are Not Alone: The Search for Intelligent Life on Other Worlds* (1964); I.S. SHKLOVSKII and CARL SAGAN, *Intelligent Life in the Universe* (1966; orig. pub. in Russian, 1965); J.G. STRONG, *Flight to the Stars: An Inquiry into the Feasibility of Interstellar Flight* (1965); G.K. O'NEILL, *The High Frontier: Human Colonies in Space* (1977); W.J. SAUBER, *The Fourth Kingdom* (1975); B. BOVA, *The Seeds of Tomorrow* (1977); C. SAGAN, *The Cosmic Connection* (1973); G.H. STINE, *The Third Industrial Revolution* (1975).

(F.C.D.III)

Space Law

Space law refers to those rules accepted by the nations of the world as legally binding on them and their nationals in conducting activities in outer space and on celestial bodies. While "outer space" has not as yet been formally defined by the nations, it is generally taken to mean the area beyond most of the earth's atmosphere at which an unpowered ar-

tificial satellite can remain in orbit, roughly 70–100 miles above the earth's surface and beyond.

Since the emergence of the modern nation-state in the 15th and 16th centuries, a substantial body of rules has developed by which the states regulate their dealings with each other. International law, reflecting the present largely unintegrated system of world order, is itself decentralized. The same states that make the law (by following custom until it is considered a legal rule, or increasingly, by making treaties) are "bound" by it, judge deviations from it, and on occasion enforce it. Yet in an increasingly technologically unified world, man's modest ability to penetrate beyond the limits of his atmosphere simply projects the possibility of conflict and misunderstanding into a new milieu. In the face of these new potentials it has been generally assumed that the most likely means to regulate relations would continue to be the appropriate, established rules of international law. As on the surface of the earth, international law sets the guides and limits that permit nations to operate, hopefully with a maximum possibility of accommodation and a minimum of conflict.

NEW DIMENSIONS AND OLD PRINCIPLES

Especially after the beginning of space flight with the Soviet satellite Sputnik 1 in 1957, diplomats and many writers began to consider which rules of international law might be applicable to man's activities in outer space and whether new rules were necessary. There was a general consensus that the bulk of already existing international law concerning the acceptable interactions of states, including the Charter of the United Nations with its limitations on the use of force, was indeed applicable to this new dimension. One doctrine of that law, governing the acquisition of title to new territory, required special attention.

As international law developed with respect to the earth's surface, it became established that certain areas, the high seas being the most notable example, were free for the peaceful use of all nations and were not subject to national appropriation (*res extra commercium*). Land areas that were empty or occupied by "barbaric" people were, however, considered as belonging to no one (*res nullius*) and hence available to whichever "civilized" nation claimed possession and could enforce this claim. Acquisition of sovereignty, the near absolute right of a state to control persons, things, and acts in a given area and to exclude nationals of all other states, required more than discovery and symbolic acts; reasonable control and effective occupation and use, at least to the degree possible, have been held necessary for a state to perfect such a claim.

From the outset, spokesmen for the space powers, the Soviet Union and the United States, for all other nations, and almost all scholars, uniformly agreed that outer space and the celestial bodies should not be considered *res nullius* in the traditional sense and that they were not to be subject to national claims of sovereignty. The concept that space and the celestial bodies are open to the peaceful use of all nations has now been formally made part of international law.

The first orbital space flights were achieved during the period of international scientific cooperation known as the International Geophysical Year (IGY 1957–58). This may help to explain why no state protested the flight over its territory of these satellites, even though in theory every nation has absolute sovereignty in the "airspace" over its land mass and territorial waters and permits no aerial flights there without its consent. This acquiescence has been taken by some as the basis at least of a rule of customary international law to the effect that such flights are outside national airspace, even when "above" a country, and that outer space is open for the peaceful use of all nations, although others argue that this acquiescence is not sufficient proof that states recognize an upper limit to their sovereignty.

THE ROLE OF THE UNITED NATIONS

The question of outer space use was first brought to the United Nations by President Eisenhower as part of the disarmament discussions of early 1957. Those discussions were not fruitful. Despite the fact that the intercontinental ballistic missile (ICBM) spends at least part of its flight time at altitudes that all would concede to be "outer space," control of the ICBM has not again been discussed as part of the outer space regime but has been left to general disarmament talks.

After the success of the Soviet Union's Sputnik 1 on Oct. 4, 1957, and of the United States' Explorer 1 on Jan. 31, 1958, both nations brought proposals for action concerning outer space to the UN in the fall of 1958. Discussions at the UN showed agreement on several legal issues: outer space was to be used for "peaceful" purposes only, general principles of law recognized by civilized nations were applicable to space; space was not subject to the extension of national sovereignty (as demonstrated by the lack of national objections to the satellites already flown). An ad hoc committee was established on Dec. 13, 1958, to consider legal and technical problems and possible international arrangements. This committee functioned through the spring of 1959 and, despite the Soviet Union's refusal to participate, due to its then widespread insistence on the inclusion of "neutral" and East European governments, its report in June 1959 set much of the tone of all that followed in the UN. It suggested, among other things, that it was premature to attempt to promulgate a comprehensive legal code or to create an internationally controlled regime for space activities, that the UN should play a modest role, and that it was unnecessary to define precisely the limits of airspace or outer space.

The Outer Space Committee

On Dec. 12, 1959, the ad hoc committee was transformed into a 24-member, permanent Committee on the Peaceful Uses of Outer Space ("Outer Space Committee") with five "neutral" states and seven Eastern European states, thus giving the Soviet Union the representation it demanded. Soviet-U.S. disputes over whether voting should be unanimous or by majority prevented Committee meetings in 1960 and 1961, however. Then, in the fall of 1961, the United States and the Western nations led the UN General Assembly in adopting, unanimously, a resolution (Res. 1,721, XVI, Dec. 20, 1961) that, for the first time, formally suggested principles to govern the exploration and use of outer space. While the UN's role continued to be deemphasized, the Secretariat was to provide a place for the registry of launchings "into orbit or beyond." Moreover, the resolution recognized: ". . . the common interest of mankind in furthering the peaceful uses of outer space and the urgent need to strengthen international cooperation in this important field," and commended to states the principles that: "(a) International law, including the Charter of the United Nations, applies to outer space and celestial bodies; (b) Outer space and celestial bodies are free for exploration and use by all States in conformity with international law and are not subject to national appropriation."

The Outer Space Committee, meeting as a whole and through legal and technical subcommittees, achieved some agreement in 1962 and early 1963, although Cold War differences between the U.S. and the U.S.S.R., especially over rights of observation from space, continued. By mid-1963, a great-power thaw set in, the principal demonstration of which was the Nuclear Test-Ban Treaty of August 1963, which, among other things, prohibited tests in the atmosphere and in outer space as well. On Oct. 17, 1963, with U.S.-U.S.S.R. agreement, the General Assembly adopted a resolution that recorded the understanding, achieved during the earlier Geneva disarmament negotiations, that no nuclear or other weapons of mass destruction would be stationed in outer space. Then, on Dec. 13, 1963, the General Assembly adopted the nine-point Declaration of Legal Principles Governing the Activities on States in the Exploration and Use of Outer Space. This declaration recognized the "common interest of all mankind" in space activities and that the exploration and use of outer space and the celestial bodies should be for the benefit of all men. Outer space and celestial bodies were "not subject to national appropriation by claim of sovereignty, by means of use or occupation, or by any other means," and were free for use by all states on a

basis of equality. All activities were to be carried on in accordance with international law, including the UN Charter; and states bore international responsibility for their acts in space, were liable for damages caused, and were to consult with other states before undertaking potentially hazardous activities. States retained ownership of their vehicles and control over their personnel; the astronauts were to be treated as "envoys of mankind" and were to receive all possible aid in the event of an emergency.

The declaration was hailed by some as the "Magna Carta" for outer space activities, and while General Assembly resolutions as such are not legally binding on members, several nations, including the United States, urged that this declaration, unanimously adopted, evidenced the present will and practice of the international community and was "international law as it is accepted by the members of the United Nations." Others stated that it was at least the basis for international rules. The Soviet Union and France were more cautious, urging that a treaty was essential to create binding rules.

THE OUTER SPACE TREATY

Whatever the view of the 1963 declaration as present law, the issue was resolved for most nations by the creation, under UN auspices, in December 1966 of an Outer Space Treaty (Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, Including the Moon and Other Celestial Bodies). The treaty was signed on Jan. 27, 1967, by representatives of 63 nations and is now in force for most of the world's states. In 17 articles it reaffirms all the UN resolutions noted above. Thus, Article I declares that space shall be free for exploration and use by all states and that states shall facilitate and encourage international cooperation in such investigations; Article II states that outer space, the moon, and other celestial bodies are "not subject to national appropriation by claim of sovereignty, by means of use or occupation, or by any other means"; and Article III provides that a nation's activities in exploration and use of outer space shall be in accordance with international law, in the interests of maintaining international peace and security, and in the interests of promoting international cooperation and understanding. Article IV prohibits stationing in outer space any object carrying a nuclear weapon or any other weapon of mass destruction and installing such weapons on celestial bodies, or in outer space in any other manner, and prohibits establishing military bases and fortifications, testing weapons, or conducting military manoeuvres on celestial bodies, though military personnel may be used for scientific research or for any other peaceful purposes. Article V again calls astronauts "envoys of mankind" and provides that, in event of accident, distress, or emergency landing on the territory of another state or on the high seas, they will be given all possible assistance, including safe return. On celestial bodies, astronauts of different states are to render all possible assistance to one another.

No sovereignty

Article VI places international responsibility on a state for its activities in space, whether by governmental or nongovernmental entities. Nongovernmental entities must have authorization and "continuing supervision" by the state concerned. Article VII provides that the launching state and the state from whose territory or facility an object is launched are internationally liable for any resulting damages to the persons of another state. Article VIII states that jurisdiction and control is retained by the state on whose registry an object is carried and over the personnel thereof in outerspace or on a celestial body. Such objects or component parts are to be returned to the state of registry if found in another state. Article IX provides for cooperation and mutual assistance among states, who are to have due regard for the corresponding interests of other states and are to avoid harmful contamination of celestial bodies and of the environment of the earth through space activities. If potentially harmful interference with activities of another treaty state is foreseen, appropriate international consultations are to be undertaken upon the request of any of the states concerned before the planned activity is undertaken.

Article X calls for opportunities for states to observe the flight activities of other states as determined by agreements between them. Article XI provides for informing the UN secretary general as well as the public and the scientific community, to the greatest extent feasible and practicable, of the nature, conduct, locations, and results of a state's activities in outer space. Article XII advances this concept by providing that all space vehicles and installations on the moon or other celestial bodies will be open for inspection by other parties to the treaty, who are limited, however, by the need for advance consultation to assure that adequate safety precautions are taken and to avoid interference with normal operation of the space facility.

THE CONTINUING DEVELOPMENT OF SPACE LAW

As a result of the continuing work of the Outer Space Committee, an Agreement on the Rescue and Return of Astronauts and the Return of Objects Launched into Outer Space was signed on April 22, 1968, by 44 nations and is now in force for the space powers and most other states. It makes more precise the requirements of safe and unqualified return of astronauts, provides that the launching authority may be obliged to share in the cost of recovering a vehicle but not an astronaut, and asserts again the permanent ownership and control of the launching state. It was not until 1970 that a draft treaty on liability was substantially completed under UN auspices as a supplement to earlier UN resolutions and the 1967 Outer Space Treaty.

Communications. The area of most intense concern over the peaceful uses of outer space is that of communications. Satellites already have demonstrated their utility in adding to existing techniques for communicating between widely separated earth points and promise revolutionary changes in decreased costs and widespread international use. The United States, in the Communications Satellite Act of 1962, established a hybrid entity, the Communications Satellite Corporation (Comsat), with stock held in part by the public and in part by U.S. communications companies and with several directors appointed by the president to provide service through the space medium. Comsat has, in turn, become the U.S. "representative" and the manager of the International Telecommunications Satellite Consortium (Intelsat), functioning under special and interim agreements and linking countries generating some 90 percent of all present international telecommunications traffic through communications satellites. The Soviet bloc has a similar arrangement, based on the Soviet Union's Molniya ("lightning") system.

Problems as well as encouraging prospects exist. In addition to the possible creation of competitive systems, there are demands that, internationally, the United States yield greater control in Intelsat to other nations and, within the United States, that Comsat give up its monopoly on domestic use of communications satellites. Since some locations in space are better than others for placing satellites, disagreements over "usurpation" may occur. For decades the International Telecommunications Union (ITU) has allocated radio frequencies, which are a limited resource, among the nations; here too the demand for frequencies for space use and problems of conflicting uses are growing increasingly acute. In addition to these forms of competition, there is a potent source of future conflict in direct broadcast satellites, capable of communicating directly into a home receiver, which may be used as vehicles for the transmission of information and ideas in propaganda wars.

Other arrangements. In addition to the general agreements discussed above, a large number of other international agreements exist. The United States, for example, has entered into some hundred bilateral treaties, executive agreements, or memoranda of understanding with other nations and separate jurisdictions concerning such disparate matters as tracking and data acquisition, launchings for other states, communications testing, meteorology, training for scientists, and the provision of a satellite-television educational system for India. Joint U.S.-U.S.S.R.

arrangements exist in the fields of meteorology and communications, and the Soviet Union also has made a number of bilateral agreements. In addition, the Western European countries have cooperated in the European Space Research Organization (ESRO) and in the effort to develop a launch capability in the European Space Vehicle Launcher Development Organization (ELDO).

International organizations

In addition, a number of international organizations with very widespread membership among the nations have participated in one phase or another of space developments and hence in the creation of the background for law. The ITU has already been mentioned. The International Civil Aviation Organization (ICAO) now sets standards for international flight in airspace and may exercise a role where air and space flight interact. Others that have considered space matters include the World Meteorological Organization (WMO), the United Nations Educational, Scientific, and Cultural Organization (UNESCO), the World Health Organization (WHO), and the International Atomic Energy Agency (IAEA). Moreover, the space nations have utilized a special group of national science experts, meeting as the Committee on Space Research (Cospar), for some coordination of national space programs, for the exchange of information, and for discussion of potentially harmful space experiments and the problems of contamination.

Problems and future developments. Though space law has seen a remarkable growth in a relatively short time, through customary law, treaty, and resolution, problems remain unresolved. These include the need for a more precisely defined limit to airspace (and hence to national sovereignty) and for resolution of the problem of control and ownership of whatever resources there may be on celestial bodies, quite apart from claims of sovereignty. Disagreement continues as to whether observation from space is always permissible, which has been the U.S. view, or whether it is "spying" and should therefore be barred, as the Soviets maintained (though this is no longer vigorously pressed since both space powers are observing from space); and no solution has yet been found to the dispute as to whether the restriction of space to "peaceful" uses bars *all* military activity (soviet view) or only aggressive military activity (U.S. view). There is also a need for the creation of a treaty dealing more precisely with liability questions; for an agreement on the future role of the UN itself; on the appropriate entities for the operation of international communications nets; and on appropriate devices for multinational cooperation in space ventures. Of these, the successful negotiation of a liability convention seems the most imminent. Despite this list, consensus to date on the scope of space law is extensive and, in view of the great costs of space exploration and the likely absence of serious international conflict over space activities in the short run, the legal structure, with some additions, is probably adequate for the present. Whether it would be able to assure the peaceful exploitation of a major breakthrough in space resource technology remains unclear.

BIBLIOGRAPHY. S.H. LAY and H.J. TAUBENFELD, *The Law Relating to Activities of Man in Space* (1970), latest comprehensive coverage of all international aspects through 1969; M.S. MCDOUGAL, H.D. LASSWELL, and I.A. VLASIC, *Law and Public Order in Space* (1963), a broad treatment with emphasis in placing outer space arrangements within the pattern of the authors' treatment of all international law; P.C. JESSUP and H.J. TAUBENFELD, *Controls for Outer Space and the Antarctic Analogy* (1959), a pioneering treatise focussing on the role of international arrangements for the control of "new" and "international" areas; C.Q. CHRISTOL, *The International Law of Outer Space* (1965), a concise treatment of the authorities in the early 1960s; G. GAL, *Világürjog* (1964; Eng. trans., *Space Law*, 1969), an interesting analysis emphasizing the Eastern European and Soviet points of view; N.M. MATTE, *Aerospace Law* (1969), a general, theoretical treatment; C.W. JENKS, *Space Law* (1965), an earlier effort to chart an overall pattern; A.G. HALEY, *Space Law and Government* (1963), an interesting view by a pioneer in the field.

(H.J.T.)

Space Perception

Human beings have been interested in the perception of objects in space at least since the question of how a picture or image arises in the human eye was first asked. It was popularly thought in ancient Greece that objects are seen because they emit what was imagined to be a continuous series of extremely thin "membranes" in their own image, which fall upon the eye and there merge into the picture that is perceived. Only after long experimental research was a more tenable conception reached, in which space was described in terms of three dimensions or planes: height (vertical plane), width (horizontal plane), and depth (sagittal plane). These planes all intersect at right angles, and their single axis of intersection is defined as being located within perceived three-dimensional space—that is, in the "eye" of the perceiving individual. Man does not ordinarily perceive a binocular space (a separate visual world from each eye), but sees a so-called Cyclopean space, as if the images from each eye fuse to produce a visual field akin to that of Cyclops, a one-eyed giant in Greek mythology. The horizontal, vertical, and sagittal planes divide space into various sectors: something is perceived as "above" or "below" (the horizontal plane), as "in front of" or "behind" (the vertical plane), or as "to the right" or "to the left" (of the sagittal plane).

The study of space perception developed rapidly in the second half of the 19th century and still more rapidly during the 20th.

Many modern psychologists who deal with perceptual function hold that the study of space perception is rapidly becoming a distinct branch of psychology in its own right. This special field of investigation within psychology concentrates on the factors contributing to the perceptual articulation (organization) of objects in space (*e.g.*, on cues to depth perception, movement, form, colour, and their interactions), or dwells on particularly interesting special problems such as that of amodal perception (*e.g.*, the question of how one perceives that there are six sides on a cube, even though only three of them can be directly visible at once). Another special problem is the effect of the so-called mirror-ego on perception. It is questioned, for example, whether, when one's skin is touched while he is looking at himself in front of a mirror, he experiences the touch as if it were in the mirror image or perceives it to be in contact with his own body as it actually appears in front of the mirror.

Perception and orientation

Space perception research also has tended to become a distinct specialty because of indisputable evidence that behaviour in this area is involved in orienting the individual to his environment. Specifically, man's orientation in space typically seems to reflect his strivings (*e.g.*, to seek food or to avoid injury). He would not be capable of effective orientation behaviour, however, unless the environmental information reaching him through the various sense organs were articulated (perceptually organized) in one way or another, permitting objects in space to be perceived at least approximately in a way that corresponds to their physical "reality" (*i.e.*, to their "actual" condition as stimuli). Such perception is called veridical perception, despite philosophical disputes about perceptual truth or verity (see EPISTEMOLOGY). Without some degree of veridicality of what is perceived to be physical space, one should be unable to efficiently seek food, flee from his enemies, or take part in social interaction. This means that, in service to one's veridical perception (direct perception of stimuli as they exist), one must also experience changing stimuli as if they were stable (invariant); that is, one experiences so-called perceptual invariances. Even though the sensory image of an approaching tiger grows larger, for example, one tends to perceive that the animal's size remains unchanged. One must be capable of perceiving objects in his environment as having relatively constant characteristics (as to size, colour, and so on) despite considerable variations in stimulus conditions.

Primary gravitational effects. Not all perception of space, however, is veridical; *i.e.*, it may fail to correspond to reality. A large number of perceptual effects have been demonstrated in which space as perceived deviates from reality in some systematic way. These are cases of nonveridical perception; for example, it has been demonstrated experimentally that the three basic spatial planes (hor-

izontal, vertical, and sagittal) dominate the ability of the individual to localize visual objects in nearby space. Carefully controlled experiments show that objects tend to be perceived as lying closer to these basic dimensions or planes than they really are. Part of the explanation of these perceptual discrepancies in visual experience may lie in the force of gravity, which undeniably affects the individual from the moment of conception onward, pulling a man's viscera and muscles "down" toward the centre of the earth. The various perceptual phenomena indicative of nonveridical perception nevertheless do not generate chaos in one's perception of space, as might theoretically be expected; indeed, most often they serve to simplify and to articulate the perceived characteristics of surrounding space. If all of the mass of sensory information available at a given moment were perceived veridically, the flood of data might so confuse the perceiver that he would be unable to orient himself appropriately. In other words, some degree of selectivity in perception in general appears to be demanded for the survival of the individual. Ideally, information about the environment is perceived only as it is relevant to the goals, needs, or physiological state of the individual at a given moment.

Selective perception

Visual factors in space perception. On casual consideration, it might be concluded that the perception of space is based exclusively on vision. After closer study, however, this so-called visual space is found to be supplemented perceptually by what are clearly identifiable as cues from auditory (sense of hearing), kinesthetic (sense of bodily movement), olfactory (sense of smell), and gustatory (sense of taste) experience. In addition to these, a number of other spatial cues, such as vestibular stimuli (sense of balance) and other modes for sensing body orientation, must be taken into account. These various (*e.g.*, visual, olfactory) "spaces" are not found to be perceptually independent of one another; powerful experimental evidence shows them to interact in producing unified perceptual experiences.

It is nevertheless undeniable that the individual ordinarily receives the largest part of his information about the environment through the sense of sight; balance or equilibrium (vestibular sense) seems to rank next in importance. It has been shown that in a state of total darkness the orientation of an individual in space is mainly dependent on sensory data deriving from vestibular stimuli. The dominance of visual stimuli in the perception of space becomes particularly understandable when it is noted that vision is a so-called distance sense; it is capable of gathering information from extremely distant points in the environment, reaching out to the stars themselves. Hearing is also considered a distance sense, as is smell, though the space they encompass is considerably more restricted than that of vision. All the other senses, such as touch and taste, are usually considered to be proximal (proximate or direct) senses—*i.e.*, conveying information about the space in direct contact with the perceiving individual's body. Strictly speaking, the latter senses may, on occasion, function over distances. The smell of rotten eggs (produced by hydrogen sulfide gas) several yards away may be associated with a taste experience if the gas dissolves in the perceiver's saliva.

The eye has often been compared to a camera; this is, of course, a very rough comparison, but it is possible to think of the retina (the back surface of the inside of the eye) as if it were the film in a camera and of the lens within the eye as analogous to the single lens of the camera (see EYE AND VISION, HUMAN). Just as in a portrait photographer's camera, the picture (image) that is projected from the environment onto the retina is upside down. While it might be expected that the perceiver therefore would experience space as turned upside down, normally this does not happen. Some sort of perceptual mechanisms within the individual enable him to see his environment right side up; the exact nature of these mechanisms is under study, yet remains poorly understood. But the process of perception seems to involve at least two inversions: one (optical) inversion of the image on the retina; another (perceptual) associated with nerve impulses in the visual tissues of the brain.

Experiments have been carried out with people who were wearing special optical devices (*e.g.*, as spectacles) that reverse the right–left or up–down dimensions of images as normally received at the surface of the retina. At first the person becomes disoriented, but when he wears such a device all the time, he gradually learns to cope with space correctly while the special glasses are on. Indeed, he reorients himself to his environment to the point that he visually begins to perceive objects as if they were right side up again. This is evidence that the individual can become visually adapted in his perception of space to a new set of stimulus cues that deviate considerably from those previously learned. The process changes direction when the optical distorting device is removed; at first the basic visual dimensions appear reversed to the subject who has adapted to the spectacles. Within a short time, however, a new adaptation occurs; the subject quickly reorients himself to the earlier, well-learned, normal visual cues and perceives his environment as normal once more.

PERCEPTION OF DEPTH AND DISTANCE

The perception of depth and distance depends on information transmitted through various sense organs. Sensory cues indicate the distance at which objects in the environment are located from the perceiving individual and from each other. Such sense modalities as seeing and hearing are known to transmit depth and distance cues and are in substantial degree independent of one another. Each modality by itself can produce consistent perception of the distances of objects. Ordinarily, however, the individual relies on the collaboration of all his senses, (so-called intermodal perception).

Gross tactual–kinesthetic cues. In his perception of the distances of objects located in nearby space, one depends on his tactile (tactual or touch) sense. In tactile experience, however, kinesthesic experience (sensations of muscle movements and of movements of the sense-organ surfaces) normally is so closely associated that investigators lump the two as tactual–kinesthetic cues.

Tactual–kinesthetic sensations enable one to differentiate his own body from the surrounding environment. The body of the individual seems to function as a perceptual frame of reference—that is, as a standard against which the distances of objects are gauged. One's perception of his own body may vary from time to time, and so its role as a perceptual standard is not always consistent. It has been found that the way in which the environment is perceived also affects one's perception of his own body. Thus, for example, it has been shown experimentally that one experiences his own arm as growing longer when he is using it to point at some object off in the environment.

Man's own body as a perceptual standard

Cues from the eye muscles. When one looks at an object at a distance, the effort arouses activity in two eye-muscle systems called the ciliary muscles and the rectus muscles. The ciliary effect is called accommodation (focussing the lens for near or far vision), and the rectus effect is called convergence (moving the entire eyeball). Each of these muscle systems contracts as a perceived object approaches. The effect of accommodation in this case is to make the jellylike lens more convex (fatter), while the rectus muscles rotate the eyes to converge on the object as it comes nearer. One's experience of these muscle contractions provides him with cues to the distance of objects.

Beyond the cues of accommodation and convergence, the size of the retinal image also serves to indicate how far one is from an object. The larger the image on the retina, the closer one judges the object to be. Some workers theorize that these sensory cues activate inherited (nativistic) tendencies immediately to perceive such sensory attributes as size without the need for learning. Modern efforts to study these cues have been especially directed toward physiological changes in the body that may be related to depth perception; whether one's perception of depth is totally inborn (nativistic), and thus independent of learning, however, remains most controversial.

Accommodations and convergence provide reliable cues when the perceived object is at a distance of less than about 30 feet (9 metres) and when it is perceived binocularly (with both eyes at once).

Visual cues. Perhaps the most important perceptual cues of distance and depth depend on so-called binocular disparity; that is, since the eyes are imbedded at different points in the skull, they receive slightly different (disparate) images of any given object. The two retinal images of the same object seem to be combined perceptually in the brain into one three-dimensional experience. The degree of disparity between the two retinal images, so-called binocular parallax, depends on the difference between the angles at which an object is fixed by the right eye and by the left eye. Thus, in looking at the indicator needle on a pressure gauge, for example, one will make slightly different readings because of parallax if he uses first the left eye alone and then the right eye. The greater the parallax difference between the two retinal images, the closer the object is perceived to be.

Binocular disparity

The phenomenon of binocular disparity functions primarily in near space because with objects at considerable distances from the viewer the angular difference between the two retinal images diminishes. Visual disparity nevertheless can be used in estimating greater distances by simply using optical devices that magnify the parallax distance separately for each eye. Such devices include artillery range-finding devices and old-fashioned, three-dimensional picture viewers called stereoscopes.

In what is called visual movement parallax, distance cues are obtained from retinal changes that depend on the interposition of objects in space. Thus, when the individual moves his head either from side to side, or forward and backward, the retinal image of a nearby tree moves more, while that of a distant tree moves less. Unlike binocular disparity, which functions only in binocular vision, movement parallax is especially important for judging distance when only one eye is used (monocular vision).

Another group of visual images called perspective projections provide perceptual cues that are independent of whether there is monocularity or binocularity of vision. Although estimates of distance based on such perspective effects as the apparent distant fusing of railroad tracks in a single point are incompletely understood, they are often considered to depend heavily on learning. Such phenomena illustrate the tendency of the individual to integrate perceptions into consistent and invariant wholes. Experiences of perspective may be generated by putting appropriate lines in an oil painting (linear perspective), by gradations in the tint of the paint (colour perspective), and by viewing the surface of the earth from an airplane (aerial perspective).

Still another group of visual cues of depth and distance consists of apparent differences in object brightness. In experimental studies it is found that the brighter an object appears, the closer it seems to be. Thus, a white card against a dark background seems to recede or to move forward as the level of illumination on the card is experimentally varied. Similar effects can be induced by changing the colour (hue) of an object—*e.g.*, from bright red to dark red.

Auditory cues. Auditory cues for depth perception include sound intensity (loudness), auditory pitch, and the time lapse between visual perception and auditory perception (for example, one hears a distant cannon after seeing the flash and smoke of the explosion).

Hearing in depth

Changes in pitch function as depth cues because when a moving object (*e.g.*, an automobile) is emitting sound waves (*e.g.*, from its horn), the pitch of the sound seems to rise when the object is approaching the perceiver; to fall when it is moving away. (For additional discussion of these points, see PERCEPTION OF MOVEMENT; ILLUSIONS AND HALLUCINATIONS.)

INTERRELATIONS AMONG THE SENSES

The development of human ability to perceive space normally depends on the interaction of the senses (modalities) of sight and touch. Toward the end of its first year, a child eagerly begins to touch and to explore objects with his hands. Compared with his visual apparatus, which begins to function more efficiently at a later time, the child's sense of touch at this stage of development is delicately and effectively sensitive. By touching objects that are placed at various heights and distances from him, the child learns excellent tactual ways of perceptually evaluating what must be a highly ambiguous world as transmitted by his relatively immature visual apparatus.

The part played by other modalities (*e.g.*, hearing) does not appear to be as fundamental in perceptual learning among young children. By contrast, people are most seriously hampered in learning a detailed, well-articulated perceptual appreciation of space without vision or touch; most people (including those who are blind) may find it difficult to imagine how impoverished an understanding of space one would have if there were only auditory cues. It is well known that people with full sensory endowment learn to locate the sources of sounds by consulting both their visual and tactual experiences. There are subtler forms of sensory interaction as well; for example, one is more accurate in turning a pointer toward a distant source of sound (*e.g.*, a buzzer) if the pointer is simultaneously illuminated whenever there is buzzing. Such anchoring of one sense to another as a basis for spatial perception is vividly illustrated by experiments with laboratory animals.

Experiments with animals

Young monkeys have been taught to discriminate, on the basis of tactual perception alone, differently shaped objects (balls, cubes, cones, and cylinders, all about the size of a matchbox). That is, each animal was given two of these objects at a time and was allowed to feel them with its hands without seeing them. The animal's choice of one specific shape (say, a cube) and its rejection of another (say, a cone) were rewarded with bits of food. When this selective tactual response had been learned, the animal was given the same tasks of selection, this time to be performed visually (by looking at pictures of the same objects). Under these conditions the animal often was able to discriminate the visually presented figures correctly. This phenomenon (called cross-modal learning) may be explained as a transfer effect based on earlier visual–tactual learning (see TRANSFER OF TRAINING).

Success in orientation, in moving about effectively and without accident in everyday pursuits, is most probable when environmental information at any given time is available through as many senses as possible. When the range of sensory stimuli that forms the usual basis for the experience of perceptual space is experimentally reduced, one's success in orientation is impaired. When visual cues are sharply reduced for sighted people, they complain of disorientation, akin to groping around in a diffuse, undifferentiated fog. Indeed, injuries are all too common even when one seeks in the dark to negotiate such familiar territory as his own bedroom.

Apparently, by learning about systematic relationships that exist among a number of simultaneously available stimuli, people can perceive distances more or less correctly. The retinal image of an object 100 yards (91 metres) away is so much smaller than the image of the same object at a distance of 10 yards (9 metres) that (if they relied solely on the sizes of retinal images) people might perceive moving objects to shrink or grow rather than to recede or approach.

Experiments have shown that the distance (in depth) between selected objects in photographs is most accurately estimated when the objects have been filmed in a richly organized environment—*e.g.*, many people standing at different distances from the camera. On the other hand, it is most difficult indeed to reliably perceive even the relative depth of two vertical rods when they are presented against a background in which other cues have been reduced or eliminated.

Perceptual constancies. In general, the considerable range of sensory information offered by isolated local stimuli is typically controlled, sifted, and corrected by the perceiver. The specific nature of these "corrections to conform to reality" generally depends on the unique combination of simultaneously presented stimuli at any

given moment. In this way, one's spatial perception tends to ensure that he experiences the continually changing circumstances of the environment in which he moves and lives with some degree of stability or constancy.

This "realistic" perception, based on one's awareness of the apparently real, physical world, serves as the basis for one's practical adjustment and survival and appears clearly in a group of perceptual experiences of so-called object constancy. In the experience of size constancy, for example, within a radius of a few hundred yards from the observer, the size of objects is perceived to remain roughly constant no matter where they are. Constancy of form means that an object is perceived to retain its fixed characteristic shape regardless of variation in the angle at which it is observed; for example, a pencil seen end-on only shows a small circular profile but is still perceived as a pencil. Colour constancy clearly illustrates the way in which the brightness and hue experienced over the surface of an object is determined by direct comparison with other objects; a lump of coal still is perceived as black whether the sun is shining brightly or whether there is a dull, leaden sky.

Path recognition: navigation in space. Different animal species are equipped in various ways for the recognition of their path of movement. Earthbound species of animals often use olfactory (smell) signals in recognizing paths of varying distance; this is encountered both among social insects (*e.g.*, ants) and among many mammals (the dog sniffing from tree to tree). Certain insect larvae can retrace their path of movement by using as guides extremely fine webs or filaments spun during their advance. Many species seem to navigate by the sun, the angle between the rays of the sun and the organism's own path being held constant. The ability of migratory birds to orient themselves by stars in twilight or at night is most impressive (see MIGRATION, ANIMAL). In addition to cues offered by the stars and other heavenly bodies, such stimuli as the effects of gravity, temperature changes, and direct visual observation of landmarks such as rivers all seem to contribute to path recognition among animals.

Use of smell to judge distance

With the burgeoning of aviation and rocket-flight technology, there has been a growing research effort to increase understanding of the sensory basis for human navigation in space. Reliable perception of the vertical and horizontal dimensions and preservation of perceptual constancy for these dimensions during flight are based on the parallel activity of vision and the vestibular sense (balance, or equilibrium). Even in flying small, privately owned aircraft it has been shown that the pilot becomes disoriented unless he preserves visual control of the horizontal dimension; he cannot safely rely on the information conveyed by his sense of balance to inform him that a wing tip, for example, is dipping dangerously low. While the sense of balance ordinarily depends on the effect of gravity, the movement of an airplane produces additional (so-called centrifugal and centripetal) forces, particularly as the plane tilts or changes direction, which easily mislead one's vestibular (balance) receptors. For this reason, in high-altitude flying the horizontal line of the surface of the earth is simulated for the pilot by an optical display in front of him that works on the same principles as does a television screen.

Even greater demands on the human senses of vision and balance are made in space flights, since one is effectively weightless in outer space. At one laboratory maintained by the U.S. Navy, an enormous, very slowly rotating cylindrical chamber has been constructed. Variations in the perceptual sensitivity of people remaining for variable periods (even days at a time) in this simulated "outer space" environment are studied in an effort to anticipate the short-term and long-term effects of interplanetary flight.

Social and interpersonal aspects of space perception. Many animal species that use nests, lairs, or dens and care for their young act to defend a specifically delimited territory against interlopers. This process is clearly observable in birds and among seals during breeding season. Apparently territorial behaviour depends on a rather precise perception of space, since the animal ceases its defensive manoeuvres when an interloper passes out of the territory by moving across the "border." The social distances maintained by primates (*e.g.*, people and apes) are theorized to be the result of territorial groupings. Modern architecture is held to be dominated by an inborn human tendency to divide groups into small, separate family territories (just as birds do), the result being such structures as apartment houses. Geographical distance may also be maintained to separate individuals who belong to different social classes or races, as is observed in the ethnic neighbourhoods or ghettos of large cities.

Territorial distance

Theories of space perception. An early theory put forth by the Anglican bishop George Berkeley at the beginning of the 18th century was that the third dimension (depth) cannot be directly perceived in a visual way since the retinal image of any object is two-dimensional, as in a painting. He held that the ability to have visual experiences of depth is not inborn (nativistic) but can only result from logical deduction based on observational (empirical) learning through the use of other senses (*e.g.*, touch).

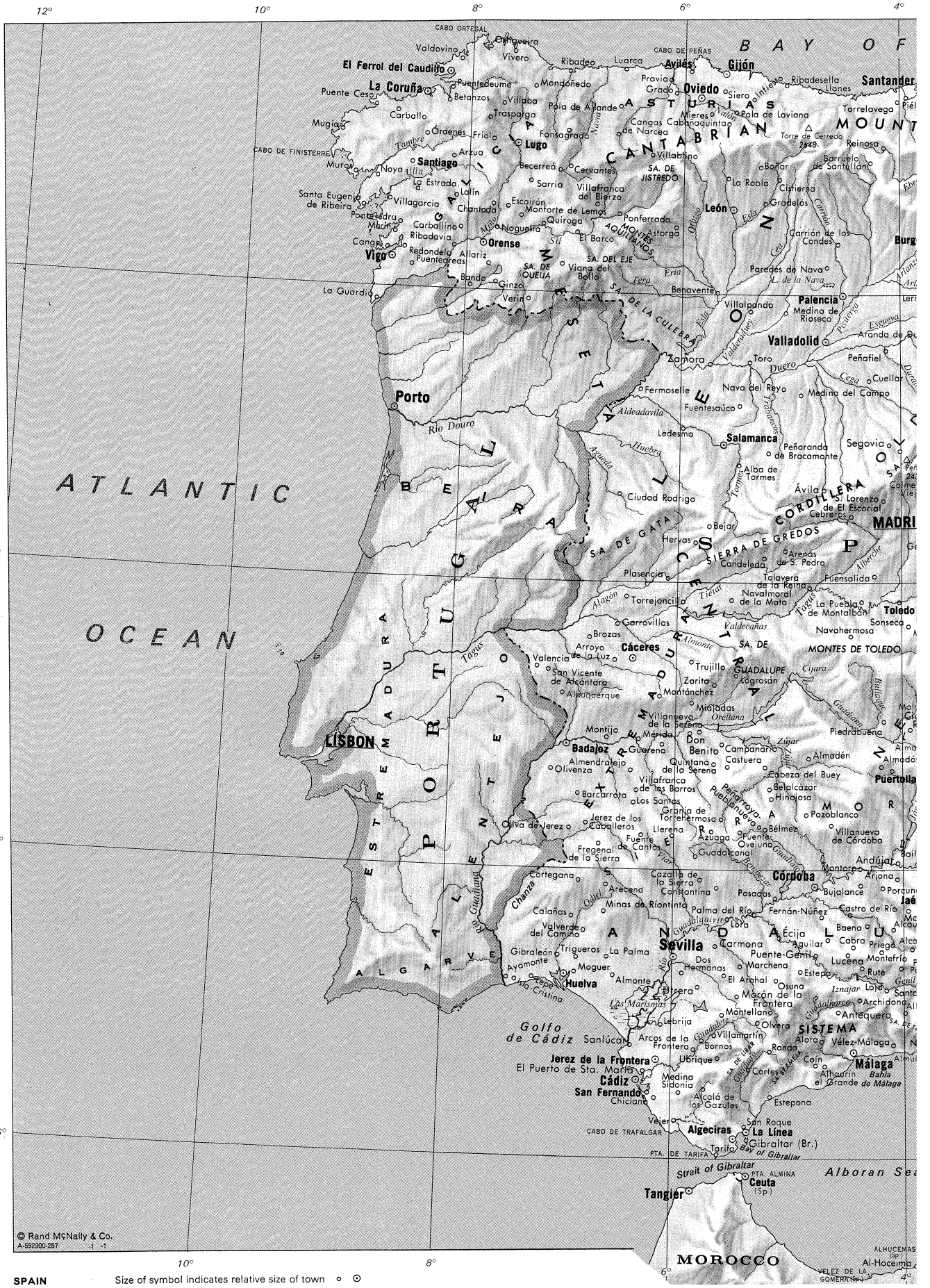
Although modern research fails to verify the importance Berkeley placed on reason in perception, contemporary theories still include both nativistic and empiricistic considerations. According to empiricistic theorists, spatial perception develops primarily as the result of learning; indeed, during the early stages of his development, the individual gradually does seem to learn something about the significance of observable (empirical) spatial cues. It is just as difficult, however, to deny the nativistic concept that space perception is based on the innate (hereditary) structure of sense receptors and the nervous system; for example, the ability to detect differences in the size of two-dimensional figures seems to be inborn (that is, the differences seem to be immediately recognizable without training or learning). For a considerably more detailed discussion of early and modern theories of space perception and of perceptual phenomena in general, see PERCEPTION.

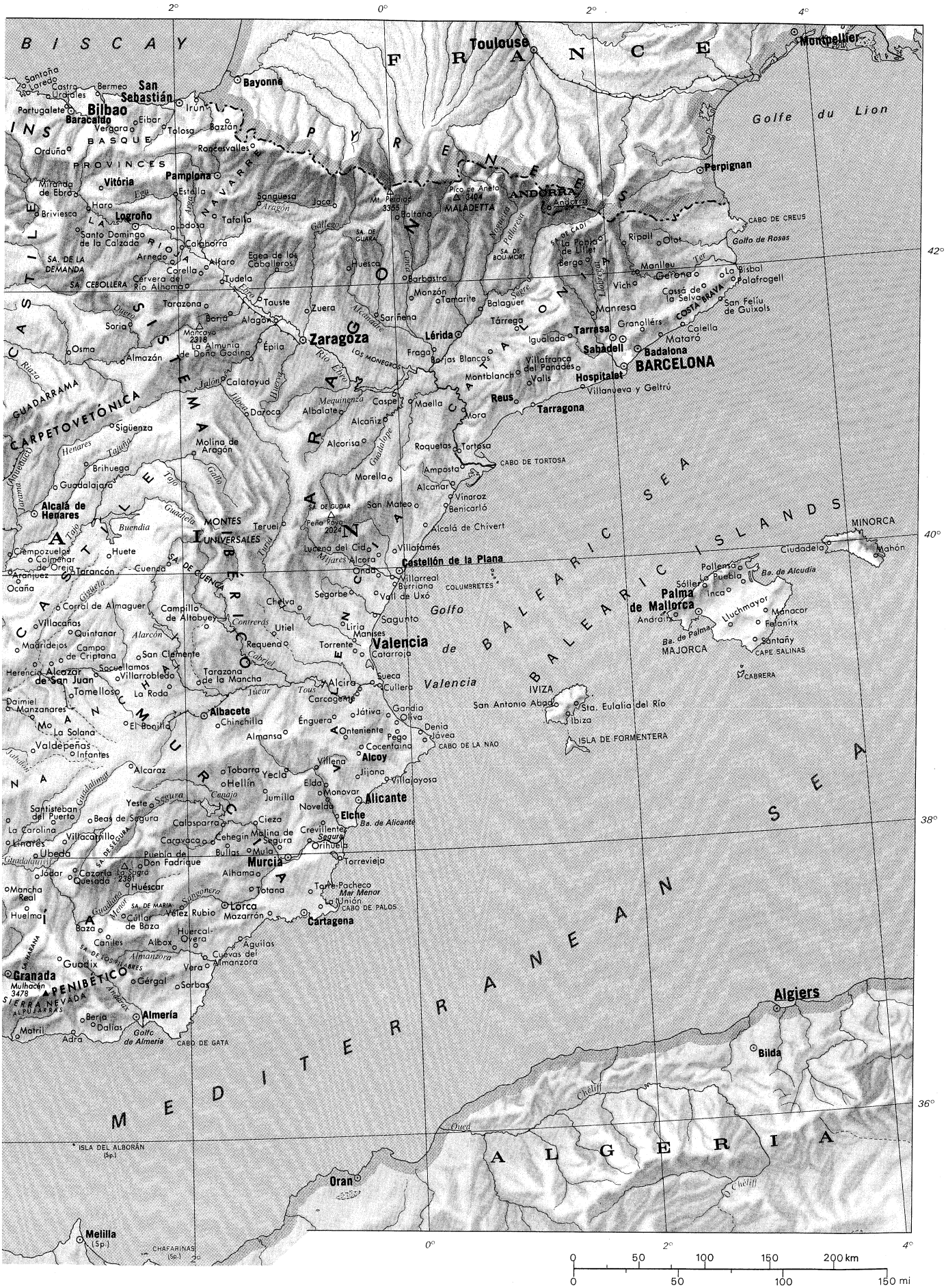
BIBLIOGRAPHY. K. VON FIEANDT, *The World of Perception*, 2nd ed. (1968), a textbook on classical and modern psychophysics and perceptual theory; E.G. BORING, *Sensation and Perception in the History of Experimental Psychology* (1942), a textbook on the growth of perceptual research and theory; J.J. GIBSON, *The Senses Considered As Perceptual Systems* (1966), a detailed discussion of the role of different cues in space perception and in the perceptual process in general; C.H. GRAHAM (ed.), *Vision and Visual Perception* (1965), a comprehensive book of experimentally oriented readings; R.N. HABER (ed.), *Contemporary Theory and Research in Visual Perception* (1968); S. KOCH (ed.), *Psychology: A Study of a Science*, vol. 1 (1959), readings of special interest in space perception; C.E. OSGOOD, *Method and Theory in Experimental Psychology* (1953), an introduction to the methodology of research in space perception; S.S. STEVENS (ed.), *Handbook of Experimental Psychology* (1951), detailed information on fundamental concepts and experiments in space perception; R.S. WOODWORTH and H. SCHLOSBERG, *Experimental Psychology*, 3rd rev. ed. (1954).

(K.V.J.v.F./E.J.J./P.Y.K.)

Spain

Lying at the southwestern end of the European continent, Spain occupies about 85 percent of the Iberian Peninsula. With Portugal, which forms its western boundary, it is shaped like the hide of a bull, the neck pointing toward Africa in the south and a large base formed by the Pyrenees mountains and the Bay of Biscay in the north. It is separated from France by the small principality of Andorra and by the Pyrenees; and in the far south lies the small peninsula of Gibraltar in the Bahía de Algeciras, which has been under British sovereignty since 1713. Off Spain's eastern coast lie the Balearic Islands, and the Canary Islands, in the Atlantic Ocean, off the coast of Africa, also belong to Spain. Other Spanish territories include the Spanish Sahara in northwest Africa and the cities of Ceuta and Melilla in northern Morocco. The total area of Spain is 297,670 square miles (770,963 square kilometres), of which the Spanish Sahara comprises 102,703 square miles, while the remaining 194,967 square miles constitute what is officially defined as the national


12°
10°
8°
6°
4°
42°
40°
38°
36°
ATLANTIC OCEAN
B A Y O F
CABO ORTEGAL
CABO DE PEÑAS
CABO DE FINISTERRE
CABO DE TRAFALGAR
El Ferrol del Caudillo
La Coruña
Gijón
Avilés
Oviedo
Santander
Santiago
Lugo
Orense
Vigo
Porto
León
Palencia
Valladolid
Zamora
Salamanca
Segovia
Ávila
MADRID
Toledo
Cáceres
Badajoz
LISBON
Sevilla
Córdoba
Huelva
Jerez de la Frontera
Cádiz
Málaga
Algeciras
La Línea
Gibraltar (Br.)
Ceuta (Sp.)
Tangier
MOROCCO
Strait of Gibraltar
Golfo de Cádiz
Alboran Sea
Bay of Gibraltar
CANTABRIAN MOUNTAINS
CORDILLERA CENTRAL
SISTEMA CENTRAL
SIERRA DE GREDOS
SA. DE GATA
SA. DEL EJE
SA. DE QUEIJA
SA. DE LA CULEBRA
MONTES DE TOLEDO
SA. DE JISTREDO
MONTES AQUILANOS
ESTREMADURA
ALGARVE
ANDALUCIA
Duero
Tajo
Guadiana
Guadalquivir
Rio Douro
Rio Tinto
Rio Odiel
Rio Tejo
Rio Guadiana
Rio Guadalquivir
Rio Guadalete
Rio Guadalhorce
Rio Genil
Rio Segura
Rio Tinto
Rio Odiel
Rio Chanza
Rio Piedras
Rio Tepe
Rio Almonte
Rio Guadalete
Rio Guadalhorce
Rio Genil
Rio Segura
Rio Guadalquivir
Rio Guadiana
Rio Tajo
Rio Douro


BISCAY
FRANCE
Toulouse
Montpellier
Bayonne
San Sebastián
Bilbao
Baracaldo
Golfe du Lion
PYRENEES
Perpignan
ANDORRA
Pamplona
Vitória
Logroño
Zaragoza
Lérida
Tarragona
Tarrasa
Sabadell
Badalona
BARCELONA
Hospitalet
Reus
Castellón de la Plana
Valencia
Golfo de Valencia
Palma de Mallorca
MAJORCA
MINORCA
IVIZA
ISLA DE FORMENTERA
Alicante
Elche
Alcoy
Albacete
Murcia
Cartagena
Lorca
Almería
Granada
Algiers
Bilda
Oran
Melilla
ISLA DEL ALBORÁN (Sp.)
CHAFARINAS (Sp.)
CABO DE GATA
CABO DE TORTOSA
CABO DE LA NAO
CABO DE PALOS
MEDITERRANEAN SEA
BALEARIC ISLANDS
ALGERIA
SISTEMA CENTRAL
SISTEMA IBÉRICO
CORDILLERA CANTÁBRICA
CARPETOVETÓNICA
SISTEMA PENIBÉTICO
SIERRA NEVADA
2°
0°
2°
4°
42°
40°
38°
36°
0 50 100 150 200 km
0 50 100 150 mi

MAP INDEX

Cities and towns

Adra 36·44n 3·01w
Aguilar 37·31n 4·39w
Águilas 37·24n 1·35w
Alagón 41·46n 1·07w
Albacete 38·59n 1·51w
Alba de Tormes 40·49n 5·31w
Albalate [del Arzobispo] 41·07n 0·31w
Albox 37·23n 2·08w
Albuquerque 39·13n 7·00w
Alcalá de Chivert 40·18n 0·14e
Alcalá de Henares 40·29n 3·22w
Alcalá de los Gazules 36·28n 5·44w
Alcalá la Real 37·28n 3·56w
Alcanar 40·33n 0·29e
Alcañiz 41·03n 0·08w
Alcaraz 38·40n 2·29w
Alcaudete 37·36n 4·05w
Alcázar de San Juan 39·24n 3·12w
Alcira 39·09n 0·26w
Alcora 40·04n 0·12w
Alcorisa 40·53n 0·23w
Alcoy 38·42n 0·28w
Alfaro 42·11n 1·45w
Algeciras 36·08n 5·30w
Alhama [de Granada] 37·00n 3·59w
Alhama [de Murcia] 37·51n 1·25w
Alhaurín el Grande 36·38n 4·41w
Alicante 38·21n 0·29w
Allariz 42·11n 7·48w
Almadén 38·46n 4·50w
Almagro 38·53n 3·43w
Almansa 38·52n 1·05w
Almazán 41·29n 2·32w
Almendralejo 38·41n 6·24w
Almería 36·50n 2·27w
Almodóvar [del Campo] 38·43n 4·10w
Almonte 37·15n 6·31w
Almuñécar 36·43n 3·41w
Alora 36·48n 4·42w
Amposta 40·43n 0·35e
Andraitx 39·35n 2·25e
Andújar 38·03n 4·04w
Antequera 37·01n 4·33w
Aracena 37·53n 6·33w
Aranda de Duero 41·41n 3·41w
Aranjuez 40·02n 3·36w
Archidona 37·05n 4·23w
Arcos de la Frontera 36·45n 5·48w
Arenas de San Pedro 40·12n 5·05w
Arjona 37·56n 4·03w
Arnedo 42·13n 2·06w
Arroyo de la Luz 39·29n 6·35w
Arzúa 42·56n 8·09w
Astorga 42·27n 6·03w
Ávila 40·39n 4·42w
Avilés 43·33n 5·55w
Ayamonte 37·13n 7·24w
Azuaga 38·16n 5·41w
Badajoz 38·53n 6·58w
Badalona 41·27n 2·15e
Baena 37·37n 4·19w
Bailén 38·06n 3·46w
Balaguer 41·47n 0·49e
Bande 42·02n 7·58w
Barbastro 42·02n 0·08e
Barcarrota 38·31n 6·51w
Barcelona 41·23n 2·11e
Barruelo de Santullán 42·54n 4·18w
Baza 37·29n 2·46w
Baztán 43·15n 1·33w
Beas de Segura 38·15n 2·53w
Becerreá 42·51n 7·10w
Béjar 40·23n 5·46w
Belalcázar 38·34n 5·10w
Bélmez 38·16n 5·12w
Benavente 42·00n 5·41w
Benicarló 40·25n 0·26e
Berga 42·06n 1·51e
Berja 36·51n 2·57w
Bermeo 43·26n 2·43w
Betanzos 43·17n 8·12w
Bilbao 43·15n 2·58w
Boltaña 42·27n 0·04e
Boñar 42·52n 5·19w
Borja 41·50n 1·32w
Borjas Blancas 41·31n 0·52e
Bornos 36·48n 5·44w
Brihuega 40·48n 2·52w
Briviesca 42·33n 3·19w
Brozas 39·37n 6·46w
Bujalance 37·54n 4·22w
Bullas 38·03n 1·40w
Burgos 42·21n 3·42w
Burriana 39·53n 0·05w
Cabañaquinta 43·10n 5·38w
Cabeza del Buey 38·43n 5·13w
Cabra 37·28n 4·27w
Cáceres 39·29n 6·22w
Cádiz 36·32n 6·18w
Calahorra 42·18n 1·58w
Calañas 37·39n 6·53w
Calasparra 38·14n 1·42w
Calatayud 41·21n 1·38w
Calella 41·37n 2·40e
Campanario 38·52n 5·37w
Campillo de Altobuey 39·36n 1·48w
Campo de Criptana 39·24n 3·07w
Candeleda 40·09n 5·14w
Cangas 42·16n 8·47w
Cangas de Narcea 43·11n 6·33w
Caniles 37·26n 2·43w
Caravaca 38·06n 1·51w
Carballino 42·26n 8·04w
Carballo 43·13n 8·41w
Carcagente 39·08n 0·27w
Carmona 37·28n 5·38w
Carrión de los Condes 42·20n 4·36w
Cartagena 37·36n 0·59w
Caspe 41·14n 0·02w
Cassa de la Selva 41·53n 2·53e
Castellón de la Plana 39·59n 0·02w
Castro de Río 37·41n 4·28w
Castro Urdiales 43·23n 3·13w
Castuera 38·43n 5·33w
Catarroja 39·24n 0·24w
Cazalla de la Sierra 37·56n 5·45w
Cazorla 37·55n 3·00w
Cebreros 40·27n 4·28w
Cehegín 38·06n 1·48w
Cervantes 42·52n 7·04w
Cervera del Río Alhama 42·01n 1·57w
Ceuta 35·52n 5·20w
Chantada 42·37n 7·46w
Chelva 39·45n 0·59w
Chiclana [de la Frontera] 36·25n 6·08w
Chinchilla 38·55n 1·55w
Cieza 38·14n 1·25w
Cistierna 42·48n 5·07w
Ciudadela 40·02n 3·50e
Ciudad Real 38·59n 3·56w
Ciudad Rodrigo 40·36n 6·32w
Cocentaina 38·45n 0·26w
Coín 36·40n 4·45w
Colmenar de Oreja 40·06n 3·23w
Colmenar Viejo 40·40n 3·46w
Constantina 37·52n 5·37w
Córdoba 37·53n 4·46w
Corella 42·07n 1·47w
Corral de Almaguer 39·46n 3·11w
Cortegana 37·55n 6·49w
Cortes [de la Frontera] 36·37n 5·20w
Crevillente 38·15n 0·48w
Cuéllar 41·25n 4·19w
Cuenca 40·04n 2·08w
Cuevas del Almanzora 37·18n 1·53w
Cúllar de Baza 37·35n 2·34w
Cullera 39·10n 0·15w
Daimiel 39·04n 3·37w
Dalías 36·49n 2·52w
Daroca 41·07n 1·25w
Denia 38·51n 0·07e
Don Benito 38·57n 5·52w
Dos Hermanas 37·17n 5·55w
Écija 37·32n 5·05w
Egea de los Caballeros 42·08n 1·08w
Éibar 43·11n 2·28w
El Arahal 37·16n 5·33w
El Barco [de Valdeorras] 42·25n 6·59w
El Bonillo 38·57n 2·32w
Elche 38·15n 0·42w
Elda 38·29n 0·47w
El Ferrol [de Caudillo] 43·29n 8·14w
El Puerto de Santa María 36·36n 6·13w
Enguera 38·59n 0·41w
Épila 41·36n 1·17w
Escairón 42·40n 7·40w
Estella 42·40n 2·02w
Estepa 37·18n 4·54w
Estepona 36·26n 5·08w
Fermoselle 41·19n 6·23w
Felanitx 39·28n 3·08e
Fernán-Núñez 37·40n 4·43w
Fonsagrada 43·08n 7·04w
Fraga 41·31n 0·21e
Fregenal de la Sierra 38·10n 6·39w
Friol 43·02n 7·47w
Fuensalida 40·03n 4·12w
Fuente de Cantos 38·15n 6·18w
Fuente Ovejuna 38·16n 5·25w
Fuentesaúco 41·14n 5·30w
Gandía 38·58n 0·11w
Garrovillas 39·43n 6·33w
Gérgal 37·07n 2·33w
Gerona 41·59n 2·49e
Getafe 40·18n 3·43w
Gibraleón 37·23n 6·58w
Gijón 43·32n 5·40w
Ginzo [de Limia] 42·03n 7·43w
Gradefes 42·37n 5·13w
Grado 43·23n 6·04w
Granada 37·13n 3·41w
Granja de Torrehermosa 38·19n 5·35w
Granollers 41·37n 2·18e
Guadalajara 40·38n 3·10w
Guadalcanal 38·06n 5·49w
Guadix 37·18n 3·08w
Guareña 38·51n 6·06w
Haro 42·35n 2·51w
Hellín 38·31n 1·41w
Herencia 39·21n 3·22w
Hervás 40·16n 5·51w
Hinojosa [del Duque] 38·30n 5·09w
Hospitalet 41·22n 2·08e
Huelma 37·39n 3·27w
Huelva 37·16n 6·57w
Huércal-Overa 37·23n 1·57w
Huesca 42·08n 0·25w
Huéscar 37·49n 2·32w
Huete 40·08n 2·41w
Ibiza 38·54n 1·26e
Igualada 41·35n 1·38e
Inca 39·43n 2·54e
Infantes 38·44n 3·01w
Infiesto 43·21n 5·22w
Irún 43·21n 1·47w
Isla Cristina 37·12n 7·19w
Jaca 42·34n 0·33w
Jaén 37·46n 3·47w
Játiva 38·59n 0·31w
Jávea 38·47n 0·10e
Jerez [de la Frontera] 36·41n 6·08w
Jerez de los Caballeros 38·19n 6·46w
Jijona 38·32n 0·30w
Jódar 37·50n 3·21w
Jumilla 38·29n 1·17w
La Almunia de Doña Godina 41·29n 1·22w
La Bisbal 41·57n 3·03e
La Carolina 38·15n 3·37w
La Coruña 43·22n 8·23w
La Estrada 42·41n 8·29w
La Guardia 41·54n 8·53w
Lalín 42·39n 8·07w
La Línea 36·10n 5·19w
La Palma [del Condado] 37·23n 6·33w
La Pobla de Lillet 42·15n 1·59e
La Puebla 39·46n 3·01e
La Puebla de Montalbán 39·52n 4·21w
Laredo 43·24n 3·25w
La Robla 42·48n 5·37w
La Roda 39·13n 2·09w
La Solana 38·56n 3·14w
La Unión 37·37n 0·52w
Lebrija 36·55n 6·04w
Ledesma 41·05n 6·00w
León 42·36n 5·34w
Lepe 37·15n 7·12w
Lérida 41·37n 0·37e
Maella 41·08n 0·09e
Lerma 42·02n 3·45w
Linares 38·05n 3·38w
Liria 39·38n 0·36w
Llanes 43·25n 4·45w
Llerena 38·14n 6·01w
Lluchmayor 39·29n 2·54e
Lodosa 42·25n 2·05w
Logroño 42·28n 2·27w
Logrosán 39·20n 5·29w
Loja 37·10n 4·09w
Lora [del Río] 37·39n 5·32w
Lorca 37·40n 1·42w
Los Santos [de Maimona] 38·27n 6·23w
Luarca 43·32n 6·32w
Lucena 37·24n 4·29w
Lucena del Cid 40·08n 0·17w
Lugo 43·00n 7·34w
Madrid 40·24n 3·41w
Madridejos 39·28n 3·32w
Mahón 39·53n 4·15e
Maladetta 42·40n 1·00e
Málaga 36·43n 4·25w
Malagón 39·10n 3·51w
Manacor 39·34n 3·12e
Mancha Real 37·47n 3·37w
Manises 39·29n 0·27w
Manlléu 42·00n 2·17e
Manresa 41·44n 1·50e
Manzanares 39·00n 3·22w
Marchena 37·20n 5·24w
Marín 42·23n 8·42w
Martos 37·43n 3·58w
Mataró 41·32n 2·27e
Mazarrón 37·36n 1·19w
Medina del Campo 41·18n 4·55w
Medina de Ríoseco 41·53n 5·02w
Medina Sidonia 36·27n 5·55w
Melilla 35·19n 2·57w
Mérida 38·55n 6·20w
Miajadas 39·09n 5·54w
Mieres 43·15n 5·46w
Minas de Ríontinto 37·42n 6·35w
Miranda de Ebro 42·41n 2·57w
Moguer 37·16n 6·50w
Molina de Aragón 40·51n 1·53w
Molina de Segura 38·03n 1·12w
Mondoñedo 42·26n 7·22w
Monforte de Lemos 42·31n 7·30w
Monóvar 38·25n 0·47w
Montánchez 39·13n 6·09w
Montblanch 41·22n 1·10e
Montefrío 37·19n 4·01w
Montellano 37·00n 5·34w
Montijo 38·55n 6·37w
Montoro 38·01n 4·23w
Monzón 41·55n 0·12e
Mora 39·41n 3·46w
Mora [de Ebro] 41·05n 0·38e
Morella 40·37n 0·06w
Morón de la Frontera 37·08n 5·27w
Motril 36·45n 3·31w
Mugia 43·06n 9·13w
Mula 38·03n 1·30w
Murcia 37·59n 1·07w
Muros 42·47n 9·02w
Nava del Rey 41·20n 5·05w
Navahermosa 39·38n 4·28w
Navalmoral de la Mata 39·54n 5·32w
Nerja 36·44n 3·52w
Nogueira 42·18n 7·31w
Novelda 38·23n 0·46w
Noya 42·47n 8·53w
Ocaña 39·56n 3·31w
Oliva 38·55n 0·07w
Oliva de Jerez 38·16n 6·55w
Olivenza 38·41n 7·06w
Olot 42·11n 2·29e
Olvera 36·56n 5·16w
Onda 39·58n 0·15w
Onteniente 38·49n 0·37w
Órdenes 43·04n 8·24w
Orduña 42·59n 3·00w
Orense 42·20n 7·51w
Orihuela 38·05n 0·57w
Ortigueira 43·41n 7·51w
Osma 41·35n 3·05w
Osuna 37·14n 5·07w
Oviedo 43·22n 5·50w
Palafrogell 41·55n 3·10e
Palencia 42·01n 4·32w
Palma del Río 37·42n 5·17w
Palma de Mallorca 39·34n 2·39e
Pamplona 42·49n 1·38w
Paredes de Nava 42·09n 4·41w
Pego 38·51n 0·07w
Peñafiel 41·36n 4·07w
Peñaranda de Bracamonte 40·54n 5·12w
Peñarroya-Pueblonuevo 38·18n 5·16w
Piedrabuena 39·02n 4·10w
Piélagos 43·22n 3·57w
Pinos-Puente 37·15n 3·45w
Plasencia 40·02n 6·05w
Pola de Allande 43·16n 6·37w
Pola de Laviana 43·15n 5·34w
Ponferrada 42·33n 6·35w
Pollensa 39·53n 3·01e
Pontevedra 42·26n 8·38w
Porcuna 37·52n 4·11w
Portugalete 43·19n 3·01w
Posadas 37·48n 5·06w
Pozoblanco 38·22n 4·51w
Pravia 43·29n 6·07w
Priego [de Córdoba] 37·26n 4·11w
Puebla de Don Fadrique 37·58n 2·26w
Puenteáreas 42·11n 8·30w
Puente Ceso 43·14n 8·54w
Puentedeume 43·24n 8·10w
Puente-Genil 37·23n 4·47w

MAP INDEX (continued)

Puertollano.....38·41n 4·07w
Quesada.....37·51n 3·04w
Quintana de la Serena.....38·44n 5·40w
Quintanar [de la Orden].....39·34n 3·03w
Quiroga.....42·29n 7·16w
Redondela.....42·17n 8·36w
Reinosa.....43·00n 4·08w
Requena.....39·29n 1·06w
Reus.....41·09n 1·07e
Ribadavia.....42·17n 8·08w
Ribadeo.....43·32n 7·02w
Ribadesella.....43·28n 5·04w
Ripoll.....41·29n 2·12e
Roncesvalles....43·01n 1·19w
Ronda.....36·44n 5·10w
Roquetas.....40·49n 0·30e
Rute.....37·19n 4·22w
Sabadell.....41·33n 2·06e
Sagunto.....39·41n 0·16w
Salamanca.....40·58n 5·39w
San Antonio Abad.....38·58n 1·18e
San Clemente....39·24n 2·26w
San Feliu de Guixols.....41·47n 3·02e
San Fernando....36·28n 6·12w
Sangüesa.....42·35n 1·17w
San Ildefonso o la Granja.....40·54n 4·00w
San Lorenzo de El Escorial.....40·35n 4·09w
Sanlúcar [de Barrameda].....36·47n 6·21w
San Mateo.....40·28n 0·11e
San Roque.....36·13n 5·24w
San Sebastián...43·19n 1·59w
Santa Eugenia de Ribeira.....42·33n 9·00w
Santa Eulalia del Río.....38·59n 1·31e
Santa Fé.....37·11n 3·43w
Santander.....43·28n 3·48w
Santañy.....39·22n 3·07e
Santiago [de Compostela]....42·53n 8·33w
Santisteban del Puerto.....38·15n 3·12w
Santo Domingo de la Calzada...42·26n 2·57w
Santoña.....43·27n 3·27w
San Vicente de Alcántara.....39·21n 7·08w
Saragossa, see Zaragoza
Sariñena.....41·48n 0·10w
Sarria.....42·47n 7·24w
Segorbe.....39·51n 0·29w
Segovia.....40·57n 4·07w
Sevilla (Seville)..37·23n 5·59w
Siero.....43·23n 5·40w
Sigüenza.....41·04n 2·38w
Socuéllamos.....39·17n 2·48w
Sóller.....39·46n 2·42e
Sonseca.....39·42n 3·57w
Sorbas.....37·07n 2·07w
Soria.....41·46n 2·28w
Sueca.....39·12n 0·19w
Tafalla.....42·31n 1·40w
Talavera de la Reina.....39·57n 4·50w
Tamarite [de Litera].....41·52n 0·26e
Tarancón.....40·01n 3·00w
Tarazona [de Aragón].....41·54n 1·44w
Tarazona de la Mancha.....39·15n 1·55w
Tarifa.....36·01n 5·36w
Tarragona.....41·07n 1·15e
Tarrasa.....41·34n 2·01e
Tarrega.....41·39n 1·09e
Tauste.....41·55n 1·15w
Teruel.....40·21n 1·06w
Tobarra.....38·35n 1·41w
Toledo.....39·52n 4·01w
Tolosa.....43·08n 2·04w
Tomelloso.....39·10n 3·01w
Toro.....41·31n 5·24w
Torrejoncillo.....39·54n 6·28w
Torrelavega.....43·21n 4·03w
Torrente.....39·26n 0·28w
Torre-Pacheco...37·44n 0·57w
Torrevieja.....37·59n 0·41w
Tortosa.....40·48n 0·31e
Totana.....37·46n 1·30w
Trasparga.....43·11n 7·51w
Trigueros.....37·23n 6·50w
Trujillo.....39·28n 5·53w
Tudela.....42·05n 1·36w
Úbeda.....38·01n 3·22w
Ubrique.....36·41n 5·27w
Utiel.....39·34n 1·12w
Utrera.....37·11n 5·47w
Valdepeñas.....38·46n 3·23w
Valdoviño.....43·36n 8·08w
Valencia.....39·28n 0·22w
Valencia [de Alcántara].....39·25n 7·14w
Valladolid.....41·39n 4·43w
Vall de Uxó.....39·49n 0·14w
Valls.....41·17n 1·15e
Valverde del Camino.....37·34n 6·45w
Vejer [de la Frontera].....36·15n 5·58w
Vélez-Málaga....36·47n 4·06w
Vélez Rubio.....37·39n 2·04w
Vera.....37·15n 1·52w
Vergara.....43·07n 2·25w
Verín.....41·56n 7·26w
Viana del Bollo...42·11n 7·06w
Vich.....41·56n 2·15e
Vigo.....42·14n 8·43w
Villablino.....42·56n 6·19w
Villacañas.....39·38n 3·20w
Villacarrillo.....38·07n 3·05w
Villafamés.....40·07n 0·03w
Villafranca del Bierzo.....42·36n 6·48w
Villafranca de los Barros.....38·34n 6·20w
Villafranca del Panadés.....41·21n 1·42e
Villagarací [de Arosa].....42·36n 8·45w
Villajoyosa.....38·30n 0·14w
Villalba.....43·18n 7·41w
Villalpando.....41·52n 5·24w
Villamartín.....36·52n 5·38w
Villanueva de Córdoba.....38·20n 4·37w
Villanueva de la Serana.....38·58n 5·48w
Villanueva y Geltrú.....41·14n 1·44e
Villarreal.....39·56n 0·06w
Villarrobledo.....39·16n 2·36w
Villena.....38·38n 0·51w
Vinaroz.....40·28n 0·29e
Vitoria.....42·51n 2·40w
Vivero.....43·38n 7·35w
Yecla.....38·37n 1·07w
Yeste.....38·22n 2·18w
Zamora.....41·30n 5·45w
Zaragoza (Saragossa).....41·38n 0·53w
Zorita.....39·17n 5·42w
Zuera.....41·52n 0·47w

Physical features and points of interest

Agueda, *river*....41·02n 6·56w
Alagon, *river*.....39·44n 6·53w
Alarcon, *reservoir*.....39·36n 2·10w
Alberche, *river*...39·58n 4·46w
Alborán, Isla del, *island*.....35·58n 3·02w
Alboran Sea.....36·00n 4·00w
Alcanadre, *river*..41·37n 0·12w
Alcudia, Bahía de, *bay*.....39·48n 3·13e
Aldeadavila, *reservoir*.....41·10n 6·40w
Algeciras, Bahía de, see Gibraltar, Bay of
Alhucemas, *island*.....35·13n 3·53w
Alicante, Bahía de, *bay*.....38·18n 0·27w
Almanzora, *river*.....37·14n 1·46w
Almería, see Andorax
Almería, Golfo de *gulf*.....36·46n 2·30w
Almonte, *river*....39·42n 6·28w
Alpujarras, *mountains*.....36·55n 3·15w
Andalucía (Andalusia), *historic region*...37·36n 4·30w
Andarax (Almería), *river*..36·48n 2·26w
Aneto, Pico de, *peak*.....42·38n 0·40e
Aquilianos, Montes, *mountains*.....42·27n 6·40w
Aragón, *historic region*.....41·00n 1·00w
Aragón, *river*.....42·13n 1·44w
Arga, *river*.....42·18n 1·47w
Arlanza, *river*....42·06n 4·09w
Arlanzón, *river*...42·03n 4·17w
Asturias, *historic region*.....43·20n 6·00w
Atlantic Ocean...40·00n 10·00w
Baetic Cordillera, see Penibético, Sistema
Balearic Islands (Islas Beleares).39·30n 3·00e
Balearic Sea.....40·00n 2·00e
Basque Provinces (Vascongadas), *historic region*...43·00n 2·45w
Bembezar, *river*..37·45n 5·13w
Bermeja, Sierra, *mountains*.....36·28n 5·14w
Biscay, Bay of (Baie de Biscaye).....44·00n 4·00w
Bou Mort, Sierra de, *mountains*...42·13n 1·07e
Brava, Costa, *coast*.....41·45n 3,04e
Buendia, *reservoir*.....40·25n 2·43w
Bullaque, *river*...38·59n 4·17w
Cabrera, *island*...39·09n 2·56e
Cabriel, *river*.....39·14n 1·03w
Cadí, Sierra de, *mountains*.....42·17n 1·42e
Cádiz, Golfo de, *gulf*.....36·50n 7·10w
Cantabrian Mountains (Cordillera Cantábrica).....43·00n 5·00w
Carpetovetónica, Cordillera (Cordillera Central), *mountains*.....40·30n 5·00w
Carrión, *river*....41·53n 4·32w
Castilla la Nueva, see New Castile
Castilla la Vieja, see Old Castile
Catalonia (Cataluna), *historic region*...42·00n 2·00e
Cea, *river*.....42·00n 5·36w
Cebollera, Sierra, *ridge*.....42·00n 2·40w
Cega, *river*.....41·33n 4·46w
Cenajo, *reservoir*.38·20n 1·50w
Central, Cordillera, see Carpetovetónica, Cordillera
Cerredo, Torre de, *mountain*....43·10n 4·58w
Chafarinas, *islands*.....35·11n 2·26w
Chanza, *river*....37·33n 7·31w
Cijara, *reservoir*..39·18n 4·52w
Cinca, *river*.....41·26n 0·21e
Columbretes, *islands*.....39·52n 0·40e
Contreras, *reservoir*.....39·32n 1·30w
Creus, Cabo de, *cape*.....42·19n 3·19e
Cuenca, Serrania de, *mountains*...39·55n 1·50w
Culebra, Sierra de la, *mountains*...41·54n 6·20w
Demanda, Sierra de la, *mountains*.....42·15n 3·05w
Duero, *river*.....41·02n 6·56w
Duratón, *river*....41·37n 4·07w
Ebro, *river*.....40·43n 0·54e
Ega, *river*.....42·19n 1·55w
Eje, Sierra del, *mountains*.....42·18n 6·54e
Eria, *river*.....42·03n 5·44w
Esgueva, *river*....41·40n 4·43w
Esla, *river*.....41·29n 6·03w
Extremadura, *historic region*...39·00n 6·00w
Filabres, Sierra de los, *mountains*.....37·13n 2·20w
Finisterre, Cabo de, *cape*.....42·53n 9·16w
Formentera, Isla de, *island*.....38·42n 1·28e
Galicia, *historic region*.....43·00n 8·00w
Gállego, *river*....41·39n 0·51w
Gallo, *river*.....40·48n 2·09w
Gata, Cabo de, *cape*.....36·43n 2·12w
Gata, Sierra de, *mountains*.....40·14n 6·45w
Genil, *river*.....37·42n 5·19w
Gibraltar, Bay of (Bahía de Algeciras).....36·09n 5·25w
Gibraltar, Strait of.....35·57n 5·36w
Gigüela, *river*....39·08n 3·44w
Gredos, Sierra de, *mountains*.....40·18n 5·05w
Guadalete, *river*..36·35n 6·13w
Guadalhorce, *river*.....36·41n 4·27w
Guadalimar, *river*.....37·59n 3·44w
Guadalope, *river*.41·15n 0·03w
Guadalquivir, *river*.....36·47n 6·22w
Guadalupe, Sierra de, *mountains*...39·26n 5·25w
Guadarrama, Sierra de, *mountains*.....40·55n 4·00w
Guadiana, *river*...37·14n 7·22w
Guadiana Menor, *river*.....37·56n 3·15w
Guadiaro, *river*...36·17n 5·17w
Guadiato, *river*...37·48n 5·04w
Guadiela, *river*...40·22n 2·49w
Guara, Sierra de, *mountains*.....42·17n 0·10w
Gudar, Sierra de, *mountains*.....40·27n 0·42w
Harana, Sierra, *mountains*.....37·20n 3·30w
Henares, *river*....40·24n 3·30w
Huebra, *river*....41·02n 6·48w
Huerva, *river*.....41·39n 0·52w
Ibérico, Sistema (Iberian Mountains).....40·30n 5·00w
Iviza, *island*.....39·00n 1·25e
Iznajar, *reservoir*.37·15n 4·30w
Jabalón, *river*....38·53n 4·05w
Jalón, *river*.....41·47n 1·04w
Jándula, *river*....38·03n 4·06w
Jarama, *river*.....40·02n 3·39w
Jiloca, *river*.....41·21n 1·39w
Jistredo, Sierra de, *mountains*...42·45n 6·25w
Júcar, *river*.....39·09n 0·14w
La Mancha, *plain*.....39·05n 3·00w
La Nao, Cabo de, *cape*.....38·44n 0·14e
La Rioja, *historic region*...42·20n 2·20w
La Sagra, *mountain*.....37·57n 2·34w
Las Marismas, *marsh*.....37·00n 6·15w
Leon (León), *historic region*...42·00n 6·00w
Libar, Sierra de..36·39n 5·22w
Llobregat, *river*...41·19n 2·09e
Los Monegros, *physical region*..41·29n 0·03w
Majorca (Mallorca), *island*.....39·30n 3·00e
Málaga, Bahía de, *bay*.....36·42n 4·24w
María, Sierra de, *mountains*.....37·39n 2·14w
Menor, Mar, *lagoon*.....37·43n 0·48w
Mediterranean Sea.....37·30n 0·30e
Mequinenza, *reservoir*.....41·15n 0·05w
Meseta Central (Meseta Iberica), *plateau*.....41·00n 4·00w
Mijares, *river*.....39·55n 0·01w
Miño, *river*.....42·10n 8·10w
Minorca, *island*...40·00n 4·00e
Moncayo, *mountains*.....41·46n 1·50w
Morena, Sierra, *mountains*.....38·00n 5·00w
Mulhacén, *peak*..37·03n 3·19w
Murcia, *historic region*.....38·30n 1·45w
Nalón, *river*.....43·32n 6·04w
Nava, Laguna de la, *lake*.....42·03n 4·39w
Navarre (Navarra), *historic region*...43·00n 1·30w
Navia, *river*.....43·33n 6·44w
Nevada, Sierra, *mountains*.....37·05n 3·10w
New Castile (Castilla la Neuva), *historic region*...40·00n 3·45w
Noguera Pallarese, *river*..42·15n 0·54e
Odiel, *river*.....37·10n 6·54w
Old Castile (Castilla la Vieja), *historic region*...41·30n 4·00w
Orbigo, *river*.....41·58n 5·40w
Orellana, *reservoir*.....39·00n 5·30w
Ortegal, Cabo, *cape*.....43·45n 7·53w
Palma, Bahía de, *bay*.....39·27n 2·35e
Palos, Cabo de, *cape*.....37·38n 0·41w
Peñalara, *mountain*.....40·51n 3·57w
Peñas, Cabo de, *cape*.....43·39n 5·51w

MAP INDEX (continued)

Penibético, Sistema (Baetic Cordillera), *mountains*.....37·00n 3·30w
Perdido, Mont, *mountain*.....42·40n 0·05e
Pisuerga, *river*...41·33n 4·52w
Pyrenees, (Pirineos), *mountains*.....42·40n 1·00e
Queija, Sierra de, *mountains*.....42·15n 7·18w
Riaza, *river*.....41·42n 3·55w
Rosas, Golfo de, *gulf*.....42·10n 3·15e
Roya, Peña, *peak*.40·24n 0·35w
Salinas, Cape....39·16n 3·03e
Sangonera, *river*.37·59n 1·04w
Segre, *river*.....41·40n 0·43e
Segura, *river*.....38·06n 0·54w
Segura, Sierra de, *mountains*...38·00n 2·43w
Sil, *river*.....42·27n 7·43w
Tagus (Tajo), *river*.....38·40n 9·24w
Tajo, *river*.....39·40n 7·30w
Tajuña, *river*.....40·07n 3·35w
Tambre, *river*....42·49n 8·53w
Tarifa, Punta de, *point*.....36·00n 5·37w
Tejeda, Sierra de, *mountains*.....36·54n 4·04w
Ter, *river*.....42·01n 3·12e
Tera, *river*.....41·54n 5·44w
Tietar, *river*.....39·50n 6·01w
Toledo, Montes de, *mountains*...39·33n 4·20w
Tormes, *river*....41·18n 6·29w
Tortosa, Cabo de, *cape*.....40·43n 0·55e
Tous, *reservoir*...39·10n 0·40w
Trabancos, *river*..41·27n 5·11e
Trafalgar, Cabo de, *cape*.....36·11n 6·02w
Turia, *river*.....39·27n 0·19w
Ulla, *river*.....42·39n 8·44w
Universales, Montes, *mountains*.....40·18n 1·33w
Valdecañas, *reservoir*.....39·45n 5·30w
Valderaduey, *river*.....41·31n 5·42w
Valencia, *historic region*.....39·30n 0·40w
Valencia, Golfo de, *gulf*.....39·30n 0·00
Vascongadas, see Basque Provinces
Vélez de la Gomera, *island*..35·11n 4·21w
Viar, *river*.....37·36n 5·50w
Zújar, *reservoir*...38·50n 5·20w
Zujar, *river*.....39·01n 5·47w

territory: Peninsular Spain; the Balearic and Canary Islands; the cities of Ceuta and Melilla; and certain small islands off northern Morocco. By the 1970 census, Spain was the home of just over 34,000,000 people, almost 76,500 of whom lived in the Spanish Sahara, with the remainder inhabiting the national territory as defined above.

One of the most ancient countries of Europe, Spain was also long one of the most powerful and, from the 16th to the 19th century, ruled a large empire in the Americas. Except for Brazil, Latin America today owes its language and essential culture to Spain. The people of Spain have been notable for their regional separatism; and history has borne witness to their peculiar pride and code of honour and to their austere and fervent devotion to the Roman Catholic faith, most remarkably in Old and New Castile on the bare central plateau, the Meseta. Spanish culture, with a heritage as African and Mediterranean in character as it is European, has given much to the Continent and to the world since Roman times.

The present constitution

Under the present constitution, Spain is a kingdom ruled by General Francisco Franco under emergency powers given to him during the Spanish Civil War (1936–39). His proclaimed heir is Prince Juan Carlos de Borbón y Borbón, who will assume the powers of head of state or be designated king in the event of Franco's death or disability. From an economic point of view, Spain has followed a liberal course since the mid-1950s; and this has favoured foreign investments and tourism, leading to impressive economic development. In the sphere of international relations, Spain is a member of the western bloc, linked to the North Atlantic Treaty Organization (NATO) through bilateral agreements with the United States. Spain is a member of the United Nations and its specialized agencies, as well as of most European technical organizations. It cooperates with the Council of Europe on cultural matters and has signed a preferential agreement with the European Economic Community, or Common Market. Spain maintains special relations with Latin American countries and Interamerican organizations and has established diplomatic relations with some countries of the socialist bloc including the Soviet Union but not the People's Republic of China.

This article concentrates principally on peninsular Spain. For related information, see MADRID; BARCELONA; PYRENEES RANGE; GUADALQUIVIR RIVER; TAGUS RIVER; SPAIN, HISTORY OF.

The article is divided into the following sections:

I. The land
  The natural environment
  The human imprint
II. The People
  Historical origins
  Demography
III. The national economy
  Sources of national income
  The management of the economy
  Economic problems and prospects
  Transportation
IV. Administration and social conditions
  Constitutional and political framework
  Social Services and social conditions
V. Cultural Life and institutions
  The cultural milieu
  The major arts
  Traditional survivals
  Popular culture
  Cultural institutions
  Communications media

I. The land

THE NATURAL ENVIRONMENT

Physical geography. With the exception of the southern triangle of Jerez–Cadiz–Gibraltar, the Iberian Peninsula forms an irregular quadrangle, bordered on the east and southeast by the Mediterranean and on the west, southwest, and most of the north by the Atlantic Ocean. Most of the peninsula is occupied by the Meseta, a vast plateau with approximately the same boundaries as the historical Kingdom of Castile. The average height of the plateau is 2,000 feet, although its northern part is slightly higher than the southern section. The Meseta is almost completely surrounded by mountain ranges: the Cordillera Cantábrica (Cantabrian Mountains) in the north; the Sistema Ibérico (Iberian System) from the Cantábrica to the southeast; the Sierra Morena in the south; and the lower mountains on the border with northern Portugal in the northwest. The southern plateau descends gently towards Portugal, whereas the transition from the northern plateau to Portugal is marked by rugged terrain and mountain ranges. The Meseta is divided by the Central Range, which extends across the plateau to the north of Madrid and runs all the way from the Sistema Ibérico to Portugal.

The Meseta

Around the central plateau, five natural regions can be identified. First, in the northeastern Spain, there is the valley of the Río Ebro, enclosed by the Sistema Ibérico in the southeast, the Pyrenees and the Cordillera Cantábrica in the north, and the mountains of Catalonia and other coastal ranges in the west. Second, there is the valley of the Guadalquivir River in the south, which is separated from the plateau by the Sierra Morena and from the Mediterranean by the Sistema Penibético (Baetic Cordillera) in the southeast but which has unimpeded passage to the Atlantic in the southwest. Third, there is the region of the coastal area of Valencia, in the east, descending from the southern Iberian ranges toward the Mediterranean, down the valleys of several small rivers. Catalonia, a mountainous region between the valley of the Ebro, the eastern Pyrenees, and the Mediterranean, forms the fourth natural region. Finally, there is northern Spain from the Pyrenees to the west, a mountainous country forming several small regions separated from each other by high mountains.

Geological features. Geologically, the peninsula forms a great massif (not to be confused with the Meseta, which it partly overlaps), bordered on the north, east, and south by mountain folds. This massif is a fragment of the former Hercynian continent, which formed across Europe at the close of the Carboniferous period (about 345,000,000 to 280,000,000 years ago). During the Mesozoic (225,000,000 to 65,000,000 years ago), the Hercynian system was shattered, and large portions sank below the sea. During the Tertiary Period (between 65,000,000 and 2,500,000 years ago), deposits of the Jurassic and Cretaceous seas, together with the earlier Tertiary beds, were crushed against the old massif to form the ranges bordering the Meseta. Later movements contorted and fractured the peninsula to give it its present general shape.

The Hercynian massif

The Balearic Islands are composed of mountain ranges of moderate altitude, not exceeding 5,000 feet. The Ca-

nary Islands, however, are of volcanic origin and have the highest peak among the Spanish territories, the Teide (12,190 feet [3,716 metres]), in the island of Tenerife.

Drainage and soils. The most important drainage systems are found in the plateau and the two large valleys of the Ebro and the Guadalquivir. The plateau has three main drainage systems, all flowing into the Atlantic: the Río Duero (Douro) in the northern plateau and the Tagus (Tajo) and Guadiana in the southern plateau. While the Duero and the Tagus follow an almost parallel course from the Iberian range to the west coast of Portugal, the Guadiana turns south and empties into the Gulf of Cadiz (Golfo de Cádiz) on the southern coast, at the frontier between Portugal and Spain. The Ebro follows an almost straight course from the northern mountains to the Mediterranean, collecting waters from the Pyrenees and the Iberic Range. The Guadalquivir flows from the east to the southwest, gathering waters from the Sierra Morena and the Sistema Penibético. In the southeastern peripheral regions, the two most important rivers are the Júcar and the Segura.

From a mineralogical point of view, three main regions can be distinguished in Spain. Crystalline rocks predominate in the central and peripheral basins, whereas the east and north have an abundance of limestones. The soils may be described as generally acid-humid in the northwest, calcareous in eastern and southern Spain, and siliceous in the Cantábrica, northern Iberian, Carpetovetónica, and Sierra Morena ranges, as well as in the Duero Valley and the mountains of Zamora and Salamanca. As in most of the western Mediterranean, erosion has played a very important role in the degradation of the soils, especially in the high areas of the plateau and the eastern and southern zones. In addition to natural erosion, man has contributed considerably through the destruction of forests. Since the 1920s, however, the government has carried out an impressive job of reforestation. Soils are generally better preserved in the north and in the Balearic and Canary islands.

Soil erosion

Climate. Spain is located in the temperate zone of the Northern Hemisphere, between the latitudes of 27° and 44° N. With the Balearic Islands and the cities of Ceuta and Melilla, Spain belongs climatically to the western Mediterranean, with some internal differences.

It is exposed both to the westerly North Atlantic winds, which dominate during most of the year, and to the warm, dry airstream that blows less frequently from the Sahara. The Pyrenees, however, play a very important role in the Spanish climate, for they constitute the northern boundary of the subtropical airstream during the summer months, thereby maintaining dry weather conditions over most of the country.

Three basic types of climate predominate within the peninsula: the coastal climate of the Atlantic and Mediterranean seaboards; the continental climate of the plateau and the valleys of the Ebro and the Guadalquivir; and the mountain climate of areas above 3,000 feet. Northern Spain, from Galicia to northern Catalonia, is characterized by moderate, even temperatures and high rainfall. Thus, La Coruña, in the northwest, ranges from 48° F (9° C) in winter to 64° F (18° C) in summer, while the annual rainfall is about 38 inches. The plateau has a continental climate, with lower rainfall and greater temperature variations. Thus, Albacete, in the southeast of the plateau, varies between 40° F (4° C) in the winter and 75.0° F (24° C) in the summer, while the annual rainfall is below 15 inches. The valleys of the Ebro and the Guadalquivir also have a continental climate, drier and colder in the Ebro and warmer and more humid in the Guadalquivir. Catalonia, Valencia, and the Balearic Islands enjoy more temperate weather, with higher rainfall in Catalonia.

Vegetation and animal life. *Vegetation.* There are four main vegetation areas: north, centre, south, and southeast. Northern Spain, enjoying humid weather, is covered with deciduous forests and meadows. It is basically a green country, good for forestry and pastures, as well as the cultivation of maize (corn), apples, and other products that thrive in humid areas. The northern plateau and the Ebro Valley, with low levels of humidity and extreme temperatures, are mostly used for grains. The upper Ebro Valley, the southern plateau, Andalusia, and the eastern regions produce olives and grapes. Valencia and the irrigated areas of Murcia are noted for their citrus fruits, particularly oranges. There are some semidesert areas in the valley of the Ebro, around the city of Zaragoza, and in southeastern Spain in the province of Almería. The most common trees are the olive and the evergreen oak, as well as the cork tree in the southwest. Pines and other coniferous trees are found in mountainous areas all over the country.

Fruit growing regions

Livestock. Cattle are raised in the north and in the mountainous areas of central and southern Spain. Sheep are the most common stock animals in the plateau region. Pigs are raised in most of Spain, particularly in the west and northwest. Goats are mostly raised in the south of the peninsula and in the islands. Mules and donkeys are still generally used all over Spain, particularly in the south, for work and transportation.

Wildlife. The larger species of wildlife are almost extinct in Spain, although bears and wolves can still be found in the Pyrenees and some of the northern mountain areas. The Spanish ibex (*Capra pyrenaica*) is threatened with extinction, and most are found within preserves. Other types of wildlife, such as deer or wild pigs, can be found in preserves all over Spain. Minor game, such as partridge, grouse, duck, and rabbit or hare, abound throughout Spain.

Fish. The Spanish coasts are rich in fish and shellfish, particularly in the Atlantic and the Bay of Biscay. Spanish fishermen, however, obtain most of their catches beyond territorial waters, in the North Atlantic and south of the Canary Islands.

THE HUMAN IMPRINT

Traditional regions. Because of the geographical diversity and the great variety of climates and vegetation, as well as for historical reasons, regional differences are very important in Spain. The country is divided into 15 traditional regions (two of which are the Canary Islands and the Balearic Islands). In the northern plateau are the regions of Old Castile and León. These two regions, together with New Castile, south of the Cordillera Carpetovetónica, constitute the larger region of Castile; and inhabitants of these three regions are generally called Castilians. The region of Old Castile also includes the province of Santander, between the Cordillera Cantábrica and the Bay of Biscay. The inland region of Extremadura and the coastal regions of Murcia and Andalusia in the south were also part of the historical Kingdom of Castile; but their inhabitants are not considered Castilians, and they have their own particular features. The northern seaboard is divided into three historical regions: Galicia, in the northwest; Asturias, between Galicia and Santander; and the Basque Provinces, bordering France. South of the Pyrenees are three historical regions: Navarre, in the western Pyrenees, also considered a part of the larger Basque country; Aragón, which covers most of the valley of the Ebro; and Catalonia, in the northeast. Finally, the region of Valencia covers a large section of the Mediterranean coast between Catalonia and Murcia.

A strong Moorish influence distinguishes Andalusia from Castile, while Asturias obtains its individuality from its position of near isolation in the Cordillera Cantábrica. Differences among the central regions are more subtle: for example, landscape, dialectal differences, and historical border. Even so, with the possible exception of the three Castilian regions, the inhabitants consider themselves essentially different from each other.

Regional consciousness

Human settlement. Spain is a predominantly industrial country, with less than 30 percent of the population involved in agriculture and other forms of primary production. There has been a sharp movement from the country to the cities since the 1930s, and this economic and structural change is of paramount importance in understanding modern Spain.

Traditional Spanish rural settlements vary widely from

one region to the next. Northern and Mediterranean Spain and the islands follow a pattern of small settlements, surrounded by agricultural land. On the other hand, the plateau and the valleys of the Ebro and the Guadalquivir are characterized by highly concentrated rural populations inhabiting large towns. These differences are connected with the history and the landscape of the different regions.

Most of the population lives in towns of 10,000 or more inhabitants. The size of towns does not, however, always reflect their economic activity. In southern Castile, Extremadura, Murcia, and Andalusia, large towns of 20,000 or 30,000 people may concentrate mainly on agricultural tasks, while smaller towns in the Basque country or Catalonia may be predominantly industrial. The two main urban centres are Madrid, the capital, in New Castile, and Barcelona, on the Catalonian coast. Madrid has an approximate population of 3,500,000 and Barcelona over 2,500,000; but, whereas Madrid has a high concentration within a small area, metropolitan Barcelona covers a large area, divided into several cities interspersed with agricultural and undeveloped land. Other large urban concentrations include: Greater Bilbao, in the Basque country; Greater Valencia, in the east; Zaragoza, in the Ebro Valley; and Seville, in the lower Guadalquivir Valley. There are 26 other cities with more than 100,000 inhabitants.

Urban centres

II. The people

HISTORICAL ORIGINS

Because of its position between two continents and its relative accessibility from the Mediterranean, Spain has been invaded and inhabited by many different peoples. In prehistoric times, peoples from North Africa and western Europe settled in the peninsula, including the Celts, in the north. The Mediterranean coast appears to have been settled by North Africans, whereas the Basques probably occupied a larger area than the present Basque country. Ancient Greek and Phoenician settlements were established on the eastern and southern coasts, and Carthaginians founded large colonies in the southeast (around what is now Cartagena); but only the Romans were able to conquer the entire area, subjecting it to intensive romanization. After the fall of the Roman Empire, Suebi, Vandals, and Alani razed the country; and the Goths established a kingdom in Spain between the 5th and 8th centuries. The Arabs invaded the peninsula in 711 and were only driven out of Spain in 1492. These eight centuries of Arab influence, together with the prolonged Roman colonization, have left a permanent imprint on the Spanish population.

The Roman conquest

After almost five centuries of consolidation of the Spanish state, the population has become unified racially; and, with the possible exception of the wandering Gypsies, there are no recognizable racial minorities. Somehow, the Spanish population has incorporated into its genetic pool something from each of the succeeding invaders.

Linguistic groups. *Castilian.* Pure Spanish (or Castilian) is the official language of the country; in addition to being the school, administrative, judicial, and commercial language of Spain, it plays a similar role in most Latin American countries. It is basically a Latin language, with some traces of the languages spoken in Spain before the Roman conquest and some Arab influence in the everyday vocabulary. There are, however, important linguistic minorities within Spain—in the Mediterranean regions, in Galicia, and in the Basque country.

Catalán. The most widely spoken and strongly rooted of the non-Castilian languages in Spain is Catalán, a Romance language closely related to the French Provençal or Languedoc. Catalán is the basic language of Catalonia, the Balearic Islands, and Valencia. The Valencia dialect of Catalán, however, has a strong Castilian influence. Catalán is a highly developed language, with a wealth of literary tradition in the areas of poetry, novel, and drama. There has been a revival of the language as an attempt to regain cultural and political autonomy against the increased process of centralization of the national government.

Gallego. Modern Portuguese originated in Gallego, a Romance language still spoken in the region of Galicia. Gallego was once the first literary language of the peninsula, but it has lost importance both in geographical extension and identity. It has been widely influenced by Castilian and is barely spoken in the Galician cities, although it still survives in the countryside and among the Galician fishermen. There is some revival of the language among intellectual circles of the region, but it has not reached the proportions of the Catalán cultural renaissance.

Basque. The most distinctive of the non-Castilian languages is found in the Basque country. Basque is neither a Romance language nor of Indo-European extraction. Its origins are uncertain, but it was spoken in large areas of the peninsula before the Romans conquered Spain. Over the centuries the language has diminished in importance, and, in recent times, it is found mostly in the countryside and in fishing villages and rarely heard in the two larger Basque cities of Bilbao and San Sebastián. Basque regionalism, however, has found the language a vehicle for separatist, or autonomist, ideals; and there is also a revival of interest in Basque in intellectual circles.

Basque separatism

Other dialects. Within the Castilian areas there are some dialects spoken in remote places, such as the Bable in the northwestern mountainous fringes of the plateau. There are also minor dialectal differences in vocabulary and pronunciation between the different regions. The best Castilian is reputed to be spoken in Old Castile, especially in the provinces of Burgos and Valladolid. Marked differences of pronunciation are observed in Andalusia and the Canary Islands, where the spoken language bears similarities to that of Latin America. Regional intonation and dialectal expressions can also be identified in the Castilian of Basques, Cataláns, and Gallegos. The Castilian-speaking immigrants and their descendants living in Catalonia constitute a large and relatively underprivileged minority (within a national minority), particularly in the metropolitan area of Barcelona, and they are called *els altres Cataláns* ("the other Cataláns").

Ethnic groups. *The Gypsies.* The Spanish population is basically homogeneous in race and culture, and the only racial minority within Spain is the Gypsies. While a large number of Gypsies have been integrated, there still remain a substantial number who lead a nomadic life and who can be seen on Spain's main roads, driving mule or donkey carts or begging or peddling in the cities. Their number is estimated at about 200,000. They do not identify with the values and customs of the majority of Spaniards, although they have Spanish citizenship and profess the Catholic faith. The greatest number are found in Andalusia, especially in Granada; but important groups are also found in some northern cities, such as Valladolid, in Old Castile, and Tarragona, in Catalonia.

Arab and Basque influences. Different racial features are still recognizable in some of the regions. Thus, Andalusia and the southeast still have an Arab imprint because of the long duration of Arab settlement in those areas; and in northern Spain natives of pure or semipure Basque stock have facial characteristics that distinguish them from the average Spaniard. In the other regions, the intermingling has been so complete that it is difficult to identify racial features.

Religious groups. *Roman Catholicism.* The Spanish population is basically Roman Catholic, and this is the official religion of the state. Although most Spaniards belong to the Catholic Church and are baptized, married, and buried within the church, active participation in church work is very small, with a large percentage of the population alienated from the church, especially since the tragic Civil War of the 1930s.

Religious minorities. Non-Catholic minorities are estimated at a total of 50,000 persons or about one and a half per 1,000 of the total population of the country, not including foreigners residing in Spain. Most of the non-Catholic population is Protestant; and while no reliable statistics are available, it appears that the larger sects are the Church of the Brethren (estimated membership, 5,500) and the Evangelical Baptist Church (estimated at

Spain, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census
Regions (*regiones*)				
Andalucía				
Provinces (*provincias*)				
Almería	3,388	8,774	361,000	375,000
Cádiz	2,851	7,385	819,000	885,000
Córdoba	5,297	13,718	798,000	724,000
Granada	4,838	12,531	769,000	733,000
Huelva	3,894	10,085	400,000	398,000
Jaén	5,212	13,498	736,000	661,000
Málaga	2,809	7,276	775,000	867,000
Sevilla	5,406	14,001	1,234,000	1,327,000
Aragón				
Provinces				
Huesca	6,051	15,671	234,000	222,000
Teruel	5,715	14,803	215,000	170,000
Zaragoza	6,639	17,194	657,000	760,000
Asturias				
Province				
Oviedo	4,079	10,565	989,000	1,046,000
Baleares				
Province				
Baleares	1,936	5,014	443,000	558,000
Canarias				
Provinces				
Las Palmas	1,569	4,065	454,000	580,000
Santa Cruz de Tenerife	1,239	3,208	491,000	591,000
Castilla la Nueva				
Provinces				
Ciudad Real	7,625	19,749	584,000	508,000
Cuenca	6,587	17,061	315,000	247,000
Guadalajara	4,707	12,190	184,000	148,000
Madrid	3,087	7,995	2,606,000	3,793,000
Toledo	5,934	15,368	522,000	469,000
Castilla la Vieja				
Provinces				
Avila	3,107	8,048	238,000	204,000
Burgos	5,509	14,269	381,000	358,000
Logroño	1,944	5,034	230,000	236,000
Palencia	3,100	8,029	232,000	199,000
Santander	2,042	5,289	432,000	467,000
Segovia	2,683	6,949	196,000	163,000
Soria	3,972	10,287	147,000	115,000
Valladolid	3,166	8,201	363,000	413,000
Cataluña				
Provinces				
Barcelona	2,986	7,733	2,878,000	3,929,000
Gerona	2,273	5,886	351,000	414,000
Lérida	4,644	12,028	334,000	347,000
Tarragona	2,426	6,283	363,000	432,000
Extremadura				
Provinces				
Badajoz	8,362	21,657	834,000	688,000
Cáceres	7,701	19,945	544,000	458,000
Galicia				
Provinces				
La Coruña	3,041	7,876	992,000	1,004,000
Lugo	3,785	9,803	480,000	415,000
Orense	2,810	7,278	451,000	414,000
Pontevedra	1,729	4,477	680,000	751,000
León				
Provinces				
León	5,972	15,468	585,000	549,000
Salamanca	4,763	12,336	406,000	372,000
Zamora	4,077	10,559	301,000	252,000
Murcia				
Provinces				
Albacete	5,737	14,858	371,000	335,000
Murcia	4,369	11,317	800,000	832,000
Valencia				
Provinces				
Alicante	2,264	5,863	712,000	920,000
Castellon	2,579	6,679	339,000	386,000
Valencia	4,156	10,763	1,430,000	1,767,000
Vascongadas y Navarra				
Provinces				
Alava	1,176	3,047	139,000	204,000
Guipúzcoa	771	1,997	478,000	631,000
Navarra	4,024	10,421	402,000	465,000
Vizcaya	853	2,210	754,000	1,043,000
Total Spain (peninsula, Canary, and Balearic Islands)	194,885	504,750	30,431,000	33,824,000
Ceuta, Melilla, and minor islands in northern Morocco				
Ceuta	7	19	65,000	63,000
Melilla	4.5	12	72,000	61,000
Alhucemas, Chafarinas, and Penón de Vélez de la Gomera	70	182	200	300
Total national territory	194,967	504,963	30,583,000	33,956,000
Spanish Sahara	102,703	266,000	24,000	76,000
Total Spain	297,670*	770,963	30,607,000†	34,033,000†

*Converted area figures do not add to total given because of rounding. †Figures do not add to total given because of rounding.
Source: Official government figures.

4,600). Recently, however, Jehovah's Witnesses have developed a strong proselytizing campaign, which has had its greatest success among the poorer sections of the population. Eastern Orthodox and non-Christian groups constitute a very small number. Among non-Christians, the Jewish community is the most important. The government has changed recently from a position of intolerance in religious matters to toleration of non-Catholic churches, provided that their membership and properties are registered with the state.

DEMOGRAPHY

Birth and mortality rates. The population of Spain, including the Canary and Balearic islands, totalled 34,000,000 in the early 1970s. The birth rate, at just under 20 per 1,000, is slightly above the average birth rate of Europe and North America but well below the figure for most developing countries. The long-range trend of the birth rate is one of decline, and this has been accentuated since the mid-1960s, with a marked preference for small families—notwithstanding the official position of the church and the government against most methods of birth control. The mortality rate is about 8.9 per 1,000, and the natural increase of the population is about 10.7 per 1,000, but this increase is partly offset by emigration. The natural rate of growth is unevenly divided throughout the country. The highest rates are registered in the Canary Islands (18 per 1,000) and in the southern regions of Murcia (14 per 1,000), and Andalusia (13 per 1,000). The Basque country, in the north, also has a high rate of natural growth of the population (13 per 1,000), whereas the lowest birth rates are found in the plateau, especially Old and New Castile.

Immigration and emigration. Traditionally, Spanish emigration has been to the Spanish-speaking countries of Latin America. After World War II, however, and especially during the 1950s and 1960s, there has been a radical change in emigration patterns. While there is still a trickle of emigrants to Latin America, this is offset by an almost equal number of immigrants and repatriates from that area, particularly from Cuba, Uruguay, and Argentina. On the other hand, since the mid-1950s most Spanish emigrants have chosen western Europe, the majority temporarily but some on a permanent basis. According to official figures, there are about 270,000 Spanish workers in western Europe, mostly in the Federal Republic of Germany, France, and Switzerland. This figure, however, should probably be doubled or tripled to reflect more accurately the true situation (including unregistered emigrants), and the total of Spanish workers in western Europe has been variously estimated at between 500,000 and 1,000,000. Spain has received some immigrants of its own, although nothing approaching the number of emigrants. After World War II, a number of eastern Europeans settled in Spain; and during the 1960s thousands of Cubans and Spaniards residing in Cuba emigrated to Spain as a result of the Cuban Revolution. Portuguese migrant workers have also found jobs in Spain. Finally, a

The emigrant work force

substantial number of retired senior citizens from western Europe and the United States have become permanent or semipermanent residents in Spain. In 1970 there were nearly 150,000 registered foreigners living in Spain, of whom some 45,000 held a labour permit, while many others probably worked without such a permit.

Rural depopulation

Internal migration has become particularly important since the end of the Civil War, showing the familiar industrial pattern of movement from the country to the cities. This has caused most Spanish cities to double or even triple their population since the mid-1930s, whereas the population as a whole has grown by only one-third. Many small villages, especially in Castile, have become almost uninhabited and in a few cases are completely deserted. The main centres of attraction for internal migration are the cities of Madrid and Barcelona. Madrid has attracted immigrants mostly from the centre and the northwest, while Barcelona has drawn its migrants from the east and the south. Other large cities, such as Bilbao, Valencia, Zaragoza, and Seville, have also attracted large numbers of immigrants, and this is true of most provincial capitals. The development of tourism has also brought workers to the resort areas of the Costa Brava (Catalonia), Costa del Sol (Andalusia), and the southeast (Alicante).

The distribution of the population. About 65 percent of the Spanish population is found in the coastal regions, whereas the plateau and the valley of the Ebro, with over one-half of the national area, possess only one-third of the total population. Thus, excluding the Spanish Sahara, much of which has a desert character, the average density of 175 inhabitants per square mile (or 68 per square kilometre) is not equally spread throughout the country; and while the coastal regions of the Basque country, Galicia, Catalonia, Valencia, and the Balearic Islands have population densities approaching those of industrial western Europe, the northern plateau, Navarre, and Aragón are almost deserted by European standards. One-third of the entire population of the inland regions is concentrated in Madrid, and this accentuates the semideserted character of the central areas. Although the metropolitan area of Barcelona also has a high percentage of the peripheral population (18 percent), this is more evenly divided among the coastal regions. Only some areas of eastern Andalusia and Murcia approach the low densities of the plateau. Most of the Spanish population now inhabits cities with populations of 10,000 or more, and the trend toward urbanization is increasing. Spanish cities have not yet developed extensive suburbs, and the urban population is heavily concentrated in high-rise apartments. The growth of suburbs is already noticeable, however, in Barcelona and Madrid, as well as in Bilbao.

Current problems and future trends. With open spaces still available and a moderately decreasing birth rate, Spain is not facing a population explosion as such, although some of the urban areas have achieved high densities. Increasing industrialization and urbanization, however, are causing problems of overpopulation and economic imbalance in the dense metropolitan areas of Madrid, Barcelona, and Bilbao and in the industrial area of the province of Guipúzcoa in the Basque country. Pollution has become an urgent problem in these and other industrial regions. There is increasing contamination of rivers and some of the coastal areas, and these developments are threatening both the freshwater supplies and the fish resources of the country. A peculiarly Spanish problem is posed by the massive influx of tourists (more than 20,000,000 per year), especially during the summer months. This sudden increase of population during the summer (a 60 percent increase over the normal Spanish population) is beginning to create problems of its own, by reducing the space available for recreation and oversaturating public services, such as health, roads, transportation, and public accommodations.

The tourist problem

Spanish public opinion, especially in the larger cities, did not become aware of the problems of pollution and overpopulation until the late 1960s. In November 1971 the Spanish government announced a program to fight pollution, with substantial financial appropriations for this purpose. As for the problems of overpopulation, the official position of both the Catholic Church and the government inhibits the development of an effective family planning program. On the other hand, the government has initiated a policy of industrial decentralization to limit the dangers of excessive urbanization, and there are studies in progress to develop self-supporting new towns around the larger metropolitan areas. (M.Me.)

III. The national economy

Spain has one of the fastest growing economies in Europe and in the early 1970s was in the process of transformation from an underdeveloped to a developed country. Its transitional status is reflected in its still comparatively low standard of living; even after a decade of explosive growth, the country's gross national product per head is well below that of Italy though above that of Portugal. But from about the mid-1950s, when Spain emerged from a period of economic as well as political isolation, its development, especially in trade with other countries, has been notable. In 1958 it joined the International Monetary Fund, in 1959 the Organization for European Economic Cooperation (now the Organization for Economic Co-operation and Development, or OECD), and in 1963 the General Agreement on Tariffs and Trade. It became an associate member of the European Economic Community in 1970. To complete its transition to developed status, Spain needs ever-increasing industrial investment and modernization of equipment. In the early 1970s, Spain, in common with many other countries, was suffering from a certain degree of business stagnation; and inflation, while less severe than in more developed countries, was also a problem.

SOURCES OF NATIONAL INCOME

Mineral resources. Spain has one of western Europe's largest coal-mining industries; in the early 1970s its output was over 13,000,000 metric tons a year but appeared to be declining. Pits are situated chiefly in the Asturias, but there are also some in the lower hills of Sierra Morena. Unfortunately, there is a surplus of steam-raising coals and a shortage of the coking coal needed by the booming steel industry. Supplies of the latter, therefore, have to be imported. By the mid-1960s, a radical reorganization of the industry was begun, and this has continued into the 1970s; uneconomic pits were closed, while investment was diverted to those having the greatest potential.

Discoveries of petroleum and natural gas have so far proved disappointing, though in 1970 there were reports of a promising offshore Mediterranean field. The originally rich veins of hematite (rubious iron ore) of Bilbao are now nearly worked out, but Spain is relatively rich in other minerals, most of which are exported. Mercury is obtained from the ancient mines of Almadén in Ciudad Real, and there is also a good supply of tungsten, lead, zinc, potash, pyrites, and some uranium oxides.

Spain struck one of the world's most spectacular finds of phosphate rock in the 1960s, in the desert of Bu Craa in the Spanish Sahara. The government is believed to have committed huge sums for the development of these reserves, which involves the construction of a railway line and port.

Agriculture, forestry, and fishing. Although Spain is one of western Europe's largest agricultural producers, the barrenness of much of the countryside has always made agriculture and cattle rearing difficult. With the recent rapid increase of manufacturing industry, it is not surprising that the proportion of national income provided by agriculture has steadily declined: from something over 27 percent in the late 1950s, it dropped to a little more than 13 percent by 1970. Yet, despite the relative decline, agriculture has maintained a small real growth rate, the more notable in that the proportion of agricultural labourers is constantly diminishing, from about 40 percent of the total working population in the mid-1960s to under 30 percent in the early 1970s.

The two main problems confronting Spanish agriculture are social and environmental; these involve, respectively, the *latifundia*, large estates owned by absentee landlords,

and the still more intractable difficulty springing from the uncertain climatic conditions and the general infertility of the land. The *latifundia* are mainly in the south, in Extremadura, Andalusia, and La Mancha, and although some attempt at division of them was begun in 1939, progress has not been great. More success has been achieved, however, in the amalgamation of small holdings into viable farm units.

Uncertain rainfall has a serious effect on crops, and land is mostly poor, with widespread soil erosion as a result of years of overtilling and overgrazing. Frost is frequently a formidable threat to the Valencian orange crop. With periodic widespread drought, irrigation has been an obvious necessity, and the earliest large canals and dams for this purpose, such as the Canal Imperial de Aragón on the Río Ebro, were built in the 18th century. Major irrigating projects have been carried out in the 20th century, and it was estimated in the mid-1960s that about one-third of the total river flow was controlled by about 140 major reservoirs. In the 1960s more dams were under construction, and water-storage capacity was being almost doubled.

Wheat, the most important cereal crop, accounts for about one-quarter of the value of all agricultural produce, but the government is trying to encourage growers to concentrate instead on fodder crops, such as barley. Barley is already grown in the southeast, while the wheat-producing areas are mainly in Old Castile, Aragon, New Castile, and Extremadura. The cereal lands, those which are less well irrigated, are farmed according to traditional practice, with a two-year rotation of crops. Pulses are a staple crop of the Meseta Central, while rice is grown in Valencia and the Ebro Delta. Potatoes, beans, and onions are also produced. Spain, in common with other western European countries, has a surplus of high-quality sugar beet production.

Citrus fruit and wine production

Spain's most successful adaptation to its difficult Mediterranean climate has been in citrus fruit production, which rose from 2,000,000 tons in 1961 to over 2,500,000 tons in 1971. Citrus fruit is grown mainly in Valencia and Castellón. Other areas specialize in other fruits; apples and pears are grown in the north and northwest, figs and almonds in Castellón and Alicante, and peaches and apricots in Murcia. The main crop of the *latifundia* is olives, almost the whole yield of which is made into oil. Vines are grown especially in La Mancha, La Rioja, and the hills of Catalonia, Valencia, and Málaga. As the third largest wine-producing country in Europe, Spain suffers from overproduction of inferior wines, although brandy and sherry from Cádiz and Jerez de la Frontera and dessert wines from Valencia and Málaga are in demand. Hemp, cotton, and tobacco are also grown, in comparatively small quantities.

Sheep and cattle rearing is very important, and Spain has one of Europe's largest sheep and goat populations (more than 17,800,000 sheep in 1972). More than half the cattle are concentrated in the humid north, especially dairy cows in the Cordillera Cantábrica. Sheep, concentrated on the Meseta Central, are of three main breeds: the long-wooled churro, the short-wooled merino, and the manchegan. Goats are best suited to the arid conditions of the southeast, and pigs are reared in the holm-oak woodlands of Extremadura.

Some afforestation is practiced to counter soil erosion. Fishing is concentrated mainly on the northern and western Atlantic coasts and is particularly important in Galicia, where Vigo is the main fishing port. The catch, averaging more than 1,000,000 metric tons a year, comprises mainly sardines, anchovies, hake, cod, tunny, bream, mollusks, and crustaceans.

Power and energy resources. Power shortages have often seriously restricted Spanish industrial production, especially in times of drought. Although Spain is a mountainous country in which, it might be thought, considerable hydroelectric power could be produced, the relatively low rainfall restricts capacity, and all profitable hydroelectric potential had been developed by the early 1970s. Much thermal electricity is produced by the abundant steam-raising coal, and imported petroleum is also used. Because of the impossibility of further hydroelectric expansion and the cost of imported petroleum, Spain has a relatively ambitious building program for nuclear power stations; in 1971 it already had a production capacity of 600 megawatts, with a further capacity of 1,000 megawatts under construction. Another reason for its development of nuclear power is that Spain is one of the world's major producers of uranium, which it may eventually hope to process itself instead of, as at present, importing enriched uranium.

Manufacturing. Although historical inertia and traditionalism long hampered the development of Spanish industry, the country's industrial growth rate since the mid-1950s has been one of the fastest in the world, and, in the early 1970s, industry was contributing about 35 percent of the total national output. Some manufacturing industries are long established and are now supported by heavily protectionist policies. The historical regional location of light industries is still operative: Manises, in Valencia province, is a centre for ceramic works; paper has been manufactured since Moorish times at Játiva, in the same province; Béjar, Segovia, Soria, and Burgos have long been woollen centres; and cotton manufactures are concentrated around Barcelona. Other light industries include a variety of textiles, clothing and footwear, food processing, liqueurs and wine, and light engineering. Chemical industries include the production of nitrogen fertilizers, which reached 596,000 tons in 1971. During the 1960s, massive investment was made in newer, high-technology industries.

Industrial growth

Heavy industries are largely centred in the north, where metallurgical works are found in Santander and the Basque provinces. Bilbao has been a steel centre since the Middle Ages. Madrid has engineering works as well as the Pegaso motor factory and the Hispano aircraft works. Passenger cars and commercial vehicle production had reached 528,000 vehicles by 1971, and crude oil refining more than 30,000,000 tons. Perhaps the most successful and competitive of Spain's heavy industries is shipbuilding; with an output of 875,000 gross tons in 1971 and around treble that figure on order from Spanish yards, Spain ranked as the sixth largest shipbuilding nation in the world.

Financial services. The central bank is the Banco de España, in the profits of which the government has participated since 1922. The dealings of the commercial banks are government regulated and are controlled by the Institute for Medium and Long-Term Credit. The majority of corporate borrowers, both public and private, have recourse either to these banks or to the state-owned holding corporation and public development bank created in 1941, the Instituto Nacional de Industria (INI). Largely because the public prefers to keep its savings on deposit rather than invest in securities, the banks control more than half of the issued capital of Spanish limited-liability companies. A considerable amount of private investment is also financed from abroad, and the Spanish stock market has thus remained relatively undeveloped. In the early 1970s the most important sources were West Germany, the United States, Switzerland, the United Kingdom, and France, in that order. The most important sectors of the economy thus affected were chemicals (32 percent) and real estate (13 percent).

Foreign trade. Importing considerably more than it exports, Spain has a visible trade deficit, which averaged about 125–155,000,000,000 pesetas a year in the late 1960s and early 1970s. Trade is most frequent with member countries of the European Economic Community (to which Spain is affiliated with associate status) and, after them, with member countries of the European Free Trade Association and with the United States. There is also some trade with Latin America. Food and agricultural exports, mainly citrus fruit, wines, fish products, and olive oil, still account for about 35 percent of the total bulk of exports, but the proportion is gradually declining. Raw materials exported are mainly iron ores and pyrites, while cotton goods, textiles, and clothing also have a steady market. Engineering and heavy industrial exports, such as ships, automobiles, and electrical goods,

are increasingly important. Some agricultural products have to be imported, notably cereals, animal products, and vegetable oils. Spain also imports raw cotton as well as textile products; petroleum and its products are a heavy import item, and engineering goods are also brought in.

Spain's rapid growth in the 1960s was made possible because imports were allowed to outstrip exports so greatly. The resultant gap was mainly financed in two striking ways. First, Spain became, in the mid-1960s, one of the world's most popular centres of tourism; visitors in 1971 alone numbered more than 26,700,000. Secondly, Spain exports labour to other European countries, and workers' remittances sent home contribute substantially to the country's balance of payments.

MANAGEMENT OF THE ECONOMY

Private enterprise and the role of government. Government influence over the economy is very strong in Spain. A large number of concerns are state-owned, including the railways and shipping, telephone services, radio and television, and hotels. The state also has large holdings of ordinary shares in a wide range of industrial companies, mostly through the medium of the Instituto Nacional de Industria (INI). This corporation, run by the Ministry of Industries, is the sole owner of several important companies and has significant share holdings in about 70 others, representing all facets of the industrial economy. In some industries, such as coal, basic steel, petroleum and petrochemicals, and automobile manufacture, INI's participation in individual firms has had an influence on the development of the whole sector. As in other developing countries, the national development bank plays a vital role quite out of proportion to its actual financial holdings.

Taxation and public spending. The collection of revenue by means of taxation was one of the least efficient features of the Spanish economy in the 1960s. Although incomes rose rapidly during the period, central government tax revenue did not rise proportionately and in 1969 was still well below that in other countries of similar development status. While direct taxation has maintained a growth reasonably proportionate to incomes, the system of indirect taxation seems to have been inadequately framed and has apparently failed to derive adequate profit from the boom sales of consumer goods, such as cars, which were a feature of the late 1960s.

During the 1960s, the government had only limited success in reducing the financial subsidies that had traditionally been given to certain favoured producers and which at one time had, indeed, played an important part in generating dynamic growth. But by the end of the decade much of this support, such as the subsidization of what had become surplus agricultural produce, was being directed into activities that were becoming increasingly uneconomic, while other projects that genuinely needed government help tended to be neglected. Works of national importance suffered least, and official programs for transportation and communications and for energy and major irrigation projects have been well maintained. The main area of neglect was in social and educational services; enrollment rates for primary education and literacy rates in general are among the lowest in western Europe. But money was beginning to be appropriated for educational purposes under the four-year development plan of 1968 to 1971 and in the plan for the period from 1972 to 1975 even larger sums were to be allocated.

The changing role of government investment

Trade unions and employer's associations. Spain's trade-union structure, the syndicalist system, has been an integral feature of the country's political development ever since the end of the Civil War in 1939. Broadly speaking, most organized economic concerns, comprising mainly the manufacturing and service industries and a substantial section of agriculture, are covered by "syndicates." In theory, these syndicates represent simultaneously the interests of employers, workers, and the state, and their formal structure still reflects the Fascist preoccupations of the 1930s and 1940s. As framed, the corporate structure of the syndicates would seem to allow neither workers nor employers sufficient independence to engage in viable collective bargaining.

Early in 1971 the Spanish Parliament finally approved a new, government-sponsored basic law for trade-union activities. Although the vertical and comprehensive structure of the syndicates was retained—under the control of a cabinet minister, the minister for syndical relations—the new law allowed management, technicians, and workers to set up independent associations. Workers could hold meetings and would be able to appeal in the courts against certain types of government decisions. But even in its new form, Spanish industrial and trade-union organization is still far from even approximating that of the countries of the European Economic Community. Spain's admission as a full member of the EEC would, of course, require a full organizational correspondence in this respect. The official syndicalist system, moreover, seems to have lost ground decisively to unofficial labour organizations, which have been operating illegally and successfully despite police harassment. In face of this development, the government appeared visibly divided between those who saw no alternative to maintaining and, if necessary, increasing repression, and those who were above all concerned about the serious economic effects of the increasing chaos in industrial relations.

Problems in the labour movement

Contemporary economic policies. Although it is normal for national economic policies to exhibit contradictory elements, in Spain the latent conflicts are more basic than in many other countries. From the late 1950s a group that came to be known as the technocrats was trying to modernize and internationalize the Spanish economy. The convictions underlying their policies were that adjustment would be easier and more successful in the context of rapid, overall growth and that this would be stimulated by encouraging competition. In the late 1960s, in order to exercise greater control over the management of the economy, the technocrats decided that fiscal and monetary policies must be strengthened and tightened.

The technocrats, whose power base was originally in the planning departments, were so successful that it became fashionable to believe that they had effective control of the economy. In a cabinet reshuffle in 1969 they did take over major cabinet posts, including the direction of foreign policy, but their appearance of success stemmed partly from the fact that the overall performance of the economy was at that time reasonably satisfactory. Their influence on individual sections of the economy was, however, apparent rather than real, and their successful manipulation of overall policy did not mean that they had managed to reform particular monetary institutions and practices. They failed adequately to effect the reduction or elimination of subsidies, and particular business decisions remained subjected to far greater government regulation in Spain than in other Western industrialized economies. In particular, the technocrats failed to dominate or perhaps even to influence two vital aspects of the economy: industrial relations and the relation between the central government and the various regions. Thus, although they had successfully brought Spain within reach of becoming one of the developed industrial states of Europe, they had not altered the basis of the economy sufficiently to make Spain's industry competitive with that of its neighbours.

ECONOMIC PROBLEMS AND PROSPECTS

For a variety of reasons, the 1970s presented to Spain more problems than had the 1960s. It seemed impossible, for purely statistical reasons, that the high growth rate of the 1960s could be maintained, and the average was expected to fall from 7.3 percent a year to a maximum of 6 percent. But expensive restructuring and modernization of equipment would remain imperative, especially in the coal-mining and textile industries, as well as in the agricultural sector, if the economy was to become internationally competitive. Industrial relations seemed likely to cause trouble, and the most profitable use of the taxation system had not yet been devised. Furthermore, uncertainties about political developments in Spain after the

end of Franco's rule affected both internal planning and confidence and the flow of foreign investment into Spain. (E.I.U.)

TRANSPORTATION

The Roman heritage

The first system of communications established on the Iberian Peninsula dates back to Roman time, and traces of the extensive Roman road network still remain. A new system of highways was not built until the 18th century, during the reign of the first Bourbon kings. Since their capital, Madrid, was situated in the heart of the country, the new system took the form of radial highways extending to the periphery of the peninsula. During the 19th century and into the 20th, efforts were made to consolidate the system of public works. The maintenance and improvement of highways has, since 1964, been one of the chief features of successive plans for economic and social development that have been undertaken.

The highway system. In Spain there is a total of about 50,700 miles (81,600 kilometres) of highway maintained nationally and a further 35,900 miles (57,800 kilometres) maintained by various provincial and local authorities.

The principal improvements being effected in the contemporary highway system were set forth in the Red de Itinerarios Asfálticos (Network of Asphalt-Paved Roads, or Redia) program, approved in 1966. Under this plan, some 3,100 miles of frequently travelled state highways were to be given a coating of asphalt over coal brick and made wide enough for at least two lanes of traffic. Provision was also made for heavy vehicles to travel in slow lanes built on strong ramps. By the early 1970s two-thirds of the highways concerned had been treated under the plan, which was scheduled for completion before mid-decade. In addition to the work being carried out under the Redia program, some 3,700 miles of macadamized road were being paved with asphalt in each year of the early 1970s, and a smaller number of arterial highways were also under construction in the major cities.

Motorway construction

Construction of a motorway (expressway) network was initiated in 1962, and, five years later, tolls were introduced. The motorway program provided for the construction of one kilometre of motorway for each 30,000 inhabitants. At the end of the first decade of the plan's operation, more than six major motorway links were in total or partial use, and a Third Development Plan envisaged the construction of twelve more. At the start of the present decade, there were almost 4,400,000 vehicles (including tractors) in circulation on Spanish roads. More than half of these were passenger cars (*turismos*), and the number of vehicles in operation was found to be increasing at a rate more than 10 percent per year, a situation that gave transportation planners much food for thought.

Railroads. Spain was only the ninth European nation to undertake the construction of railroads, beginning with the lines between Barcelona and Mataró (1848) and between Madrid and Aranjuez (1851). During the second half of the 19th century, the national railroad system was gradually expanded, with Madrid as its centre. The legal arrangement under which the Spanish railroads operated was one of indirect government ownership, whereby the state granted concessions to private companies or individuals for the construction and management of the railroads during a 99-year period, after which the railroads with all their installations and rolling stock would become the perpetual property of the state. This arrangement was radically changed in 1941 by the passage of the Ley de Bases de Ordenación Ferroviaria y de Transportes por Carretera (Law Establishing Basic Rules of Operation for Railway and Highway Transportation), creating the National Network of Spanish Railroads, Red Nacional de los Ferrocarriles Españoles (Renfe). The Civil War of 1936–39 had done considerable damage to the railroads, and it was said that only the take-over of operations by the state could save them from economic disaster. The greater part of the rolling stock belonging to the railroads consisted of outmoded equipment. In 1964 a major long-term modernization plan was initiated. Outstanding among the new equipment introduced was the Talgo, a light train of Spanish design, drawn by an electrically powered locomotive with two diesel motors. By the early 1970s the entire Spanish railroad network totalled some 12,000 miles of track, approximately one-quarter of which was electrified. The major cities of Madrid and Barcelona possessed metropolitan subway networks, and there were plans for the construction of a subway in Seville.

Ports and waterways. Spain has about 200 ports situated along its 2,425 miles of littoral—one-third of which is insular and two-thirds peninsular. Of these, 26 ports service more than 80,000 tons of shipping annually. The total of merchant shipping handled exceeds 140,000,000 tons annually, but the greater part of it—including almost all foreign trade—is handled by a very small number of ports, notably Cartagena, Santa Cruz de Tenerife, La Luz and Las Palmas de Gran Canaria, Barcelona, and Bilbao.

In addition to their important role in foreign trade, a good number of ports are indispensable to basic sectors of the internal Spanish economy, notably steel mills, petroleum refineries, and naval construction. In addition, the fishing industry, in which Spain held second place in Europe and about tenth worldwide in the early 1970s, is strongly influenced by the quality and the type of port facilities available.

In Spain, as in many other countries, maritime trade is assisted by various forms of state intervention, including premiums for shipbuilding and subsidies to specific maritime companies that are regarded as helpful, if not income producing, to the nation. Transport services operating between the peninsula and the Balearic Islands, the Canary Islands, Ceuta, Melilla, and various ports of the Spanish Sahara are known as "Sovereignty Lines." They are assigned to national shipping concerns on a competitive basis under government regulation, although the present major concessionaire, the Transmediterranean Corporation, has a concession granted for a 70-year period. Other companies receiving subsidies include the important Spanish Transatlantic Company, which offers services to the United States, the Caribbean, and Mexico, and also the Ibarra Company, which services the Mediterranean–Brazil–La Plata route.

State subsidies and concessionaires

Air transport. Under a 1940 law the Iberia Spanish Airlines Company was granted exclusive rights to provide regular domestic air services, but a 1960 law did away with this monopoly, replacing it with a 15-year renewable concession. Administratively, air transport is handled by two corporations: Iberia and Aviaco. Aviaco —in which Iberia is the largest stockholder—is a member of the European Air Charter Carriers Association. There are also a number of private companies that operate nonscheduled flights both within and without the country. By the early 1970s the Spanish airlines had a total of about 170 planes, half of them jets, running services in over 30 countries and employing some 18,000 persons. More than 6,500,000 passengers and about 80,000 tons of freight were being transported annually. In Spain itself there were over 40 airports, half of them exclusively for civilian use and a third open to international traffic. By the early 1970s the total annual movement through the national airports exceeded 27,000,000 passengers and was rising rapidly year by year. The busiest airport was that of Palma, which alone handled almost a fifth of the total, while Madrid, Barcelona, Las Palmas, Málaga, and Tenerife airports all handled more than 1,000,000 passengers.

IV. Administration and social conditions

CONSTITUTIONAL AND POLITICAL FRAMEWORK

The Fundamental Laws and the role of the Cortes. Considered as a political entity, Spain is officially designated as a Catholic, social, and representative state, which, in line with its historic traditions, is constituted as a kingdom, under the first article of the Law of Succession.

Spain has what is characterized as an open constitutional system that was set up in 1938 and which comprises a complicated series of laws designated as fundamental.

The system began with the Labour Charter of March 9, 1938, and was later amplified by a series of laws and amendments passed in 1942, 1945, 1946, 1947, and 1958, all culminating in 1967 in the Organic Law of the State. These Fundamental Laws regulate all the matters traditionally included in national constitutions, as well as a number of specific questions and problems peculiar to the Spanish nation. The Law of the Principles of the National Movement (1958) comprises a declaration of fundamental political principles. The rights and the duties of individuals are established in two statutes: that of Labour and that of the Spanish People. The Organic Law of the State (1967) treats of the organization of power, personified in the head of state, as well as of the government and the tribunals of justice. The Law of Succession (1947) states the prerogatives of the Council of the Realm. The system of indirect representation is treated in two of the Fundamental Laws: The Constitutive Law of the Cortes (1942, amended 1946), insofar as it affects the latter; and the Organic Law of the State, in its relation to the National Council. Direct representation is regulated by the Law of the National Referendum (1945). Any abrogation or modification of the Fundamental Laws is subject not only to the approval of the Cortes but also to a national referendum.

According to the first article of its constitutive law, the Cortes is the highest organ of participation by the Spanish people in the affairs of state. Its principal duty is that of elaborating and approving laws, subject to the overall approval of the head of state. It is made up of 561 members, called procurators, designated partly through limited election (as in the case of trade union representatives) and partly through direct election (as in the case of "representatives of the family" comprising both the heads of families and married women). There are also procurators representing the provinces (two for each province), while each municipality with more than 300,000 inhabitants also has its representative. The members of the National Council (see below *Political organization*) are also procurators, as are some two dozen ministers and presidents of distinguished organizations and an equivalent number of representatives of professional schools and other associations. Finally, there are representatives from cultural organizations and members designated directly by the head of state.

Composition of the Cortes

The term of office for the procurators is four years, but those who belong to the Cortes by virtue of their position in another office cease to be members when their tenure in such office expires. The president of the Cortes is appointed by the head of state from among three procurators proposed to him by the Council of the Realm. The Cortes has two vice presidents and four secretaries elected by the procurators themselves from among their number.

The structure of the government. *Central administration.* Under Article 6 of the Organic Law of the State, the head of state is the supreme representative of the nation, personifies national sovereignty, exercises supreme political and administrative power, and has a wide range of other important functions. He gives his sanction to the laws that have been approved by the Cortes and may, at the instance of the Council of the Realm (see below *Regional and Consultative Bodies*) and with its approval, return a law to the Cortes for further consideration. He may also initiate national referenda. He also directs the government of the kingdom through the Council of Ministers.

The role of the head of state

The head of state may propose to the Cortes the person he wishes to have succeed him, under conditions laid down by the Law of Succession. Such a proposal was made in the early 1970s, when the Cortes took the first step toward provision for a successor to Gen. Francisco Franco by approving the designation of Prince Juan Carlos de Borbón y Borbón, following whom the regular order of succession would be according to dynastic line, without the necessary approval of the Cortes.

The Council of Ministers is the highest central organ after the head of state. The government, as constituted by the Council of Ministers, may be convoked by the head of state and is presided over by the latter when he attends its meetings. The Organic Law of the State establishes the office of the president of the government, although up to the early 1970s no provision had been made for filling it, since the positions of head of state and president of the government were in fact both held by Franco. It is officially incumbent upon the president of the government to represent the government of the nation, to direct general policy, to insure the coordination of all the organs of government and administration, and to propose the names of those he wishes to have as ministers. There may be one or more vice presidents, while the regular ministers themselves number 18, excluding three ministers without portfolio, who are not heads of an administrative department but who form part of the government by virtue of their appointment. The ministers, in the early 1970s, were still both named and retired entirely according to the will of the head of state, but provision has been made for this role to be filled by the president of the government, when that office was filled.

The appointment of ministers

Regional administration and consultative bodies. Civil governors represent the central government in each one of the 50 provinces of Spain. They are appointed by the Council of Ministers after being proposed by the Ministry of the Interior and also serve as provincial heads of the National Movement (see below *Political organization*). Each ministry also normally has delegates in the provinces.

Mayors are the heads of local government in the various municipalities. In cities other than provincial capitals, they also represent the central government and are entrusted with maintaining public order as commissioned by the civil governor.

The Council of the Realm, which has precedence over other consultative bodies of the nation, assists the head of state on those matters that fall within his exclusive competency. In addition to its functions as a consultative body, it also has other duties related to certain high-level appointments, including that of proposing the names of the candidates from whom the president of the Cortes will be selected and making nominations for the Supreme Court of justice, the Council of State, the Court of Exchequer, and the National Economic Council. The Council of the Realm is composed of *ex officio* councillors and representatives of the Cortes.

The Council of State is the highest administrative consultative body and acts through plenary session, permanent committees, and its various divisions. The National Economic Council is the consultative, advisory, and technical organ in matters of importance that affect the national economy.

Local administration. The provinces, together with the municipalities, form what is called the local administration. The Spanish provincial system was definitively established, after several unsuccessful attempts, in 1833. By the 1970s the national territory was divided into 51 provinces, including the African province of the Spanish Sahara. Each province has its provincial House of Deputies with a president. The former is constituted as a corporation by the president and the provincial deputies, of whom there is one for each judicial district elected from among the mayors and city councilmen by electors of the municipal governments; these are in addition to other deputies who represent corporations and various provincial economic, cultural, and professional bodies, such as the Trade Union Organization.

The provincial and municipal hierarchy

Spain has about 9,000 municipalities, the government and administration of which are handled by mayors and municipal councils. The mayor is president of the municipal corporation and at the same time represents the state administration. In the provincial capitals and municipalities of more than 10,000 inhabitants, the mayor is appointed by the Ministry of the Interior. In the case of other municipalities, the appointment is made by the civil governor of the province with the approval of the Ministry of the Interior.

Municipal government comprises the mayor-president and the councilmen, whose number varies from three to 34, depending on the number of local inhabitants. One-

third of their number is made up of representatives of the heads of families; another third of members of labour organizations from within the limits of the municipality; and the final third of the economic and professional organizations located therein.

Political organization. *The National Movement.* The law of the principles of the National Movement officially characterizes the latter as the Spanish people united by law and order. It is open to all Spaniards and avowedly seeks to serve the nation by promoting political life and fostering harmonious living on a well-ordered basis. The chief organs of the movement are its national leadership and, as a collective body, the National Council. The president of the government is also the head of the National Movement and president of its National Council. The secretary general, appointed by the head of state at the instance of the president of the government, holds the rank of minister.

The Trade Union Organization. Spaniards participating in labour and production constitute what is officially known as the Trade Union Organization, which is made up of industrial, agrarian, and public utility unions. These unions have the status of public corporations with a representative basis and enjoy juridical standing in their respective spheres of competence. Within these spheres associations of management, technicians, and workers have been established. These groups, it is officially held, organize for the defense of their particular interests and also in order to take part in the wider life of the community.

The system of representation. The Spanish system of representation is officially defined as an organic democracy, operating through representatives of families, municipalities, trade union organizations, corporations, and professional schools. On the highest level, it functions through the Spanish Cortes.

Organic democracy defined

Besides the elections of family representatives to the legislature (two for each province), there are also direct municipal elections to elect councilmen who represent the family; elections of representatives of family organizations to the National Council; and elections of members of provincial councils and local units of the National Movement.

All men and women who are over 21 years of age are also entitled to vote in national referenda held to consider important matters. Finally, the governing bodies of the trade union organizations, with the exception of the secretariat, are elected by free and secret suffrage.

Justice. Spanish judicial structures alone possess the faculty of passing judgment and applying the laws, the executive being forbidden to interfere in judicial affairs. Each judicial structure, of whatever rank, is officially the sovereign judge in its own field of competence and is subject only to the law. The judges are experts who must show their competence in a competitive examination before being granted admission to their respective corps. The fiscal magistrates, or public prosecutors, number over 270, while those of the judicial branch number almost, 1,000. Promotions and the filling of vacancies takes place according to a seniority system.

The role of the Supreme Court

In the administration of justice, the Supreme Court is the highest judicial body. It has its seat in Madrid but possesses national jurisdiction. The Supreme Court may function in plenary session, as a disciplinary court, as a judicial council, and as a court of justice through its six divisions. On a lower level, there are 15 territorial courts handling civil cases of various kinds and 50 provincial courts that try criminal cases.

The jurisdiction of the petty courts (*juzgados*) extends to the "judicial districts", of which there are almost 500. These courts try civil cases and draw up criminal indictments. The municipalities, depending on their size, are seats of the municipal courts (almost 240), district courts (almost 500), and petty courts (nearly 8,200). The labour magistracy has special jurisdiction over disputes arising between employers and employees and matters involving social security and industrial accidents. There is also a juvenile court system and a Court of Exchequer (Accounts Court) responsible for the auditing and verification of state accounts.

The Spanish penitentiary system consists of two general types of prison—preventive prisons and central, or corrective, prisons. The former comprise the provincial prisons and district prisons and also houses of detention, workhouses, farm colonies, and so on. The central prisons are of two types, each numbering less than a dozen: the common central prisons and the special prisons. The total prison population of Spain in the early 1970s was just under 14,000 and was claimed to be one of the lowest for any nation in Europe.

The armed forces. The Spanish army is organized into 11 captaincies general, while the navy has three departments, and the air force is organized into three regions. Military service is obligatory for all male Spaniards who are not exempt by reason of physical defects. Active duty is ordinarily for 18 months but may be shortened when circumstances warrant.

The Junta of National Defense

The three forces are unified in the Junta of National Defense, which comprises the ministers of the military departments, the chief of the supreme staff, and the chiefs of staff of the army, navy, and air force. The supreme commander is the head of state.

SOCIAL SERVICES AND SOCIAL CONDITIONS

Educational services. The Spanish educational system, based on the General Law for Education (1970), is divided into the following units: preschool education; general basic education; *bachillerato* (see below); university education; professional formation; and continuing education for adults. All these types of education are given in both public and private institutions.

The General Law for Education, which was directed toward a complete reform of the educational system, is being put into effect gradually and is expected to be fully in force by the late 1970s or early 1980s. Associated reforms include free and obligatory basic general education until age 14; removal of rural–urban differentials in educational opportunity; equal access to the university; and university autonomy and the creation of new universities. Aid to students is also given in a broad variety of scholarships, loans, and other subsidies.

Preschool education. Preschool education comprises games, language activities, observation of nature, and related activities. It is not obligatory and is offered without charge in public schools. There are nursery schools for children between two and three years of age and kindergartens for those between four and five years old.

General basic education. Designed for children between six and 13 years of age, general basic education is divided into two stages: in the first (for children between six and 10), the studies are of a general nature; in the second (for children from 11 to 13), basic preparation for the *bachillerato* is given, and the pupils also receive training in practical activities. In both public and nonpublic institutions, general basic education is obligatory and without charge. Upon the successful completion of this program of studies, the pupils receive a diploma—the basic baccalaureate—signifying that they are school graduates. By the early 1970s, there were over 135,000 schools offering this kind of education to well over 4,700,000 pupils.

The bachillerato. The *bachillerato* program of studies, which follows the general basic education program, covers a three-year period for pupils between 14 and 16. The curriculum includes required subjects, electives, and a measure of technical or professional training. The title of *bachiller* enables the student to enter second-level professional schools or the pre-university course required before admission to the university proper. More than 2,800 schools grant the *bachillerato:* they have a total enrollment of well over 1,500,000 students.

The *bachiller* and entry to higher education

University education. University education is preceded by an orientation course that lasts for one year and that is open to those who hold the *bachillerato* or who have completed their second-level professional studies. The course is planned and directed by the university but is taught in schools that offer the *bachillerato.*

University education proper may be undertaken in university schools, which generally have a single three-year

program of studies, or in faculties or advanced technical education schools. In the latter case, there are three cycles or programs of study: a first devoted to basic disciplines over a three-year period; a second, of specialization, which lasts for two years; and a third, of specialization and preparation for research and teaching.

Students who have completed their studies in a university school or the first cycle of a faculty or technical school obtain the diploma that enables them to work in their field of competency. Those who complete studies through the second cycle receive the degree of licentiate (*licenciado*), engineer, or architect, allowing them to practice professionally or to continue their studies in the third cycle. The completion of studies in the third cycle, together with the writing and approval of a thesis, leads to the degree of doctor in the field of specialization.

In the early 1970s Spain possessed a dozen state universities, three autonomous ones, and four under private auspices. There are eight faculties (university "colleges" or "schools"). Many of the universities are of ancient foundation. The total number of students in attendance was more than 157,000 by the early 1970s.

Professional training. Professional education is given on three levels of varying length, according to the field of specialization, but always covering a period of more than two years for each level. To begin studies upon the first level, it is necessary for the student to have completed his general basic education. This level of studies must be undertaken by those who do not hold the *bachillerato.* Admission to second-level professional courses is granted to those who hold the *bachillerato* or who have completed first-level professional courses. Courses in professional formation on the third level are open to those who have completed the first cycle of studies in a university faculty or advanced technical school; university graduates; and those who have completed their second-level professional courses together with the necessary complementary courses.

Other institutions. A variety of other educational institutions also exist in Spain. These include schools of fine arts, applied arts, and artistic crafts; the schools of social work, and of teachers of physical education. Other centres, which used to exist independently but are now gradually becoming affiliated with the universities, include the schools of diplomacy, public administration, jurisprudence, and of archives and libraries. Twelve labour high schools have their own distinctive character. They confer the *bachillerato laboral* and prepare their students for a wide range of industrial occupations.

Social Security. The law concerning Social Security (1963) is based on what is technically known as unification of the risk factor—that is, every case of inadequate salary or absence of income is considered in the same way, whatever may be its cause. Likewise, all payments are made on 75 percent of the basic salary. The Ministry of Labour administers the Social Security program through the National Institute of Welfare and the Mutual Labour Insurance organizations, together with others that operate in certain specified areas, notably farm labour and maritime workers. The result of the 1963 law was the unification of an already existing group of programs for the insurance of industrial workers into one single program called Social Security. The creation of an obligatory health insurance program has meant that a type of insurance that formerly benefitted only a specific sector of the population has been expanded to insure the extension of its benefits to a much larger number.

Administration of the Social Security program

In order to provide assistance to its beneficiaries, the obligatory health insurance scheme operates over 60 centres scattered throughout the country, and five times that number were planned for the 1970s. There were also almost 500 outpatient clinics.

Housing. The Ministry of Housing, created in 1957, brought together under one head such organizations as the Office of Architectural Planning and the National Housing Institute that had formerly been in charge of housing inspection and the construction of new units. Not included was the Service for the Reconstruction of Devastated Areas, which had been established for the rebuilding of the country following the destruction caused by the Civil War of 1936–39.

The first step taken toward the creation of a housing policy by the present authorities, however, was the passage, in 1939, of the Law of Protected Dwellings, the purpose of which was to promote large-scale construction of housing for the poorer classes. The instrument for implementing this policy was the National Housing Institute. Contemporary housing policy also includes such measures as housing provided on credit for families of limited income, who also enjoy tax exemption and similar benefits.

To control land speculation and the resulting scarcity of building sites and, hence, the high cost of acquiring them, the Ministry of Housing has adopted the policy of creating new cities or new urban areas rather than choosing to undertake urban renewal or land expropriation projects. During the 1945–70 period, almost 1,500,000 homes were built under state direction, and building has continued—at an accelerating rate—into the 1970s. The National Institute of Colonization has also undertaken the building of more than 200 new towns.

Control of land speculation

Public order. Because of its long tradition and special characteristics, the Spanish Civil Guard is singled out from among the various kinds of security police as a typically Spanish institution with its origin in the earlier *hermandades* (literally "brotherhoods"), a reference to a kind of rural police force established by the Catholic kings. It is organized along military lines for the accomplishment of its duties as a rural police force. In addition to watching over and preserving public order particularly in rural areas, it has several other specific tasks, including the patrolling of frontiers and the prevention of smuggling. The Civil Guard has also proved to be efficient in carrying out the duty most recently entrusted to it, namely, that of serving as highway traffic policemen and providing assistance for motorists in distress.

Organizationally, the Civil Guard forms part of the general administration of the Ministry of the Interior and has as its commanding officer a lieutenant general of the army.

The maintenance of public order in general is the duty of the General Security Administration, to which are attached the General Police Corps and the Armed Police Corps.

Social conditions. *Income and cost of living.* The policy of the present authorities with respect to salaries was originally established with the promulgation of the National Labour Regulations, which were intended to prevent the growth of any economic activity unless it was able to provide those who worked at it with a salary sufficient to maintain a decent standard of living. The excessively general character of these regulations, however, made the application of them in specific instances extremely difficult. This situation was remedied by the passage of the Law of Collective Agreements (1958), which served to establish a basis for collective bargaining between workers and employers. By the early 1970s more than 80 percent of the workers who were not self-employed were partners to such agreements.

The Third Plan of Development (1972–75), although it did not set a definite figure for the minimum wage to be paid, reviewed the matter annually in its relation to the cost of living and the rate of economic growth, both of which influence the salary adjustments reached through the agreements described.

The increase in consumer spending on the part of the Spanish family—which has now attained a level comparable to that of developed countries—is, Spanish authorities feel, one of the most favourable indicators of the real advances being made by the national economy. In recent years family consumption has shown a decided tendency toward a change from spending on food and clothing to greater expenditures for housing and services. The consumption of durable goods depends on the growth rate of the gross national product: in years when this exceeded 7 percent, there was a marked expansion in consumption, with a corresponding reduction during the years in which a slower growth rate holds.

Family consumption patterns

Generally speaking—and there are notable exceptions—the rate of increase in consumer spending during recent years is considered by the government to be satisfactory. There has been a levelling off of differences in the standard of living between the rural and urban population, particularly with respect to services and basic goods.

Health factors. Official public health services are under the administrative control of the National Department of Health, a division of the Ministry of the Interior. Its work is complemented by that of numerous other institutions, including the National School of Health, the National Anti-Tuberculosis Association, the National Cancer Institute, and so on. State, local, and private groups collaborate actively in the field of preventive medicine. In 1960, infant mortality during the first year of life—always a significant gauge of national health standards—was 43.7 per 1,000. By 1970 it had declined to 27.8 per 1,000. The ratio of 14 doctors to each 10,000 inhabitants is also relatively high when considered in relation to the low per capita income of the nation. Industrialization has caused an increase in the absolute number of industrial accidents but, relative to the present population, the number has decreased.

Recent years have witnessed important advances toward the eradication of malaria and tuberculosis. Finally, it may be noted that the life expectancy of the Spaniard increased from 69.7 years in 1960 to about 70.5 years in 1970.

To maintain its progress, Spain has about 1,650 establishments offering medical and surgical care. Public and private institutions, taken together, have more than 160,000 beds.

Social and economic divisions. Social conditions in general and social classes in particular have changed greatly in Spain over the last 30 years, a fact admitted by all observers. Growing economic prosperity has, in the Spanish view, brought an increase in the number of working class entering the middle class. At the same time, it is held that economic differences are disappearing within the various sectors of the middle class itself. Opportunities for university and technical education are increasing and are expected to play an important role in bringing about a fuller equalization of economic status among the different social groups. The figure of outstanding prestige today is undoubtedly the technical expert, whose work in both industry and administration is becoming increasingly important.

V. Cultural life and institutions

THE CULTURAL MILIEU

The Romanization of the greater part of the inhabitants of the Iberian Peninsula brought with it a diffusion of classical culture that was to influence permanently the cultural development of Spain as a nation. Specifically, the deep assimilation of the cultural ideals of Rome enabled them to survive the barbarian invasions that followed Rome's collapse.

During the ensuing centuries of Muslim domination and the entire period of the Reconquest (the expulsion of the Africans from Iberia), Spain continued to develop her unique culture by incorporating elements of both Islāmic–Oriental and Jewish civilizations. This amalgamation of cultures is, perhaps, the explanation for the development of the characteristic Spanish culture, which, while not removed from what is peculiarly European, does not correspond to it fully.

Serving as a meeting place for two distinct cultural worlds, the Islāmic–Oriental and the western European, Spain has performed an important task in the dissemination and extension of culture. The classic works of Greek antiquity were translated into Latin and later into Castilian; and a knowledge of both Arab and Jewish culture was passed on to other European nations. The Castilian tongue, which was to become the national and official language, acquired vigour and maturity and soon became outstanding among all the languages of Europe. So strong did Spanish culture become that the term Golden Age—which, in Spain, comprises both the 16th and 17th centuries—can be applied not only to the political, literary, and artistic creations of the period but also to its outstanding achievements in theology, philosophy, law, and science.

Traditional Spanish culture, however, being deeply religious in character and specifically Catholic in outlook, found itself farther and farther removed from the cultural life of the rest of Europe, as the latter (from the 18th century onward) assumed an increasingly secular quality. The critical nature of this stage in Spain's development is apparent in the conflict that arose during the 19th century between the nationalistic and the more European-oriented sectors of her society. Spanish culture has also had enormous impact in the New World.

In the prevailing Spanish viewpoint, Christian inspiration, a humanistic outlook that fosters appreciation of the worth of the individual, a certain eclecticism, and a striving toward universality of spirit are seen as characteristics of Spanish culture in every age. Contemporary Spain, it is felt, finds itself open to the most recent intellectual currents of the rest of Europe, and indeed further afield, but does not as a result intend to relinquish her own characteristic modes of thought and behaviour. It would seem that the mission of Spain continues to be that of creating her peculiarly national forms of Western culture.

THE MAJOR ARTS

Architecture. Architecture that finds its inspiration in classic sources is represented in Spain today by the work of several outstanding figures. They include Luis Gutiérrez Soto (who designed the Air Ministry in Madrid); Luis Moya (responsible for the Labour University of Gijón); and Pedro Muguruza (designer of the magnificent war memorial, the Valley of the Fallen, near Madrid). The major contemporary architects

An intermediate position between tradition and modernity is held by Miguel Fisac, who has worked chiefly in Madrid, and Juan Antonio Coderch, who has been well known in Barcelona.

Younger architects accorded acclaim by Spanish critics include Vázquez Molezun, who, together with José Antonio Corrales, designed the Reader's Digest Building in Madrid; Javier Carvajal, the designer of the Spanish Pavilion at the 1964 New York World's Fair; and Javier Sáenz de Oiza, architect of the Torres Blancas Building in Madrid. Felix Candela has also become internationally known for his work outside of Spain, especially in Mexico.

Sculpture. José Planes, Cristino Mallo, and especially Juan de Avalos have worked along the traditional lines of representational sculpture. A forerunner in the use of the nonrepresentational style was Angel Ferrant, who also worked with representational sculpture, as did Carlos Ferreira and Jorge de Oteiza. The work of Pablo Serrano is significant for its abstract style, and Eduardo Chillida gained world renown in a very short time for his sculptures in metal.

Painting. Three of the outstanding painters of the 20th century have been Spaniards: Pablo Picasso, Salvador Dalí, and Joan Miró; however, they all spent much of their working lives outside Spain. In addition, the list of contemporary Spanish painters is long. The best known artists who do representational painting more or less traditional in style include Francisco Arias, known best as a landscape painter, and also Agustín Redondela, Alvaro Delgado, Juan Antonio Morales, all portrait painters. Various types of stylized painting have been done by Antonio Saura, Juan Ponc, and Juan Brotat, while the better known pure abstractionists include Pablo Palazuelo, Rafael Canogar, Luis Feito, Antonio Tapies, Modesto Cuixart, Juan José Tharrats, and Manuel Millares. Famous 20th-century painters

Music. The internationally significant generation of musicians to which Manuel de Falla, Joaquín Turina, Conrado del Campo, and Oscar Esplá belonged was succeeded by a generation of composers born at the turn of the 20th century. They include Federico Mompou, whose work follows the French impressionistic style; and Ernesto Halffter, often considered to be the truest of the disciples of Falla and who, like his master, has sought to combine both aristocratic and popular elements in his

Innovative composers

music. Composers whose work reveals a tendency to greater innovation include Xavier Montsalvage and Joaquín Homs. The same tendency is even more clearly expressed in the works of Cristóbal Halffter, Luis de Pablo, and Carmelo Bernaola, three figures of international renown in the world of avant-garde music. Among guitarrists, Andrés Segovia, Narciso Yepes, and Regino Sáinz de la Maza all enjoy a worldwide reputation, as do such vocal artists as Victoria de los Angeles.

Literature, drama, and film. The illustrious traditions of Spanish literature continue to influence contemporary writers, and the *Don Quixote* of Cervantes has been characterized as the 17th-century prototype of the modern Spanish novel.

The Civil War (1936–39) represented a watershed in Spanish arts as well as politics, driving many artists in disparate fields into exile while substantially altering the focus for those who remained or emerged subsequently. Among its effects was the shift in predominance from poetry to the novel, though the Generation of 1927 movement in poetry, especially through such pre-eminent figures as Vicente Aleixandre and Dámaso Alonso, continued to influence a host of postwar poets, including Salvador Espriu, Gabriel Celaya, Blas de Otero, and José Hierro. The best known of 20th-century Spanish poets outside their own country, however, remained the lyric poet Juan Ramón Jiménez, winner of the Nobel Prize for Literature in 1956, and Federico García Lorca, who is perhaps even more renowned as a dramatist. In both poetry and drama Lorca echoed an earlier Spain. His poetry is alive with the labyrinthine and sensuous imagery of a Moorish Spain of several centuries past. The few plays he wrote before he was murdered during the Civil War are both richly poetic and emotionally elemental in their actions, harkening back to the works of Lope de Vega and Pedro Calderón de la Barca, which helped the Spanish stage of the 17th-century Golden Age rival that of Elizabethan England in vigour and popular appeal.

Major 20th-century novelists include Ramón Pérez de Ayala, who touched his philosophical themes with an Asturian humour; Gabriel Miró, who was occupied with more poetic themes; and Ramón José Sender, who returned to greater realism and social commentary; as well as Max Aub and Francisco Ayala. During the early 1940s the postwar novel began to emerge in the work of such writers as Camilo José Cela and Carmen Laforet, who influenced a younger group including José María de Ginorella, Jesús Fernández Santos, Juan Goytisolo, Luis Goytisolo-Gay, and Juan Marsé. Criticism, previously led by José Ortega y Gasset, champion of a pure, depersonalized art, passed to such younger observers as José Luis Cano and José María Castellest.

In the film, Spain is most widely known for the work of the expatriate director Luis Buñuel, many of whose often heavily symbolic films are regarded as among the major examples of cinematic art. Within Spain, Juan Antonio Bardem and Luis Garcia Berlanga were among those who renewed the nation's film industry in the postwar period, while Carlos Saura assumed major rank in subsequent years.

TRADITIONAL SURVIVALS

The fiesta. The form of popular tradition that remains most intensely alive in the Spain of today is that of the fiesta, or popular festival, which never fails to have a certain amount of religious significance even though it may be essentially nonreligious in character. After having been preserved for centuries, the fiestas have now been incorporated into the activities carried on by the Ministry of Information and Tourism and consequently can anticipate not only a larger audience but also a continuance of their existence. Each locality, no matter how small, has its own festival, celebrated in honour of a saint for whom the town feels a particular devotion. The list of local festivals is, as a result, almost interminable, but three outstanding ones are the Fallas of Valencia, the April fair of Seville, and the festivities honouring San Fermín in Pamplona. Although the festivals are found all over Spain, it is interesting to note that the coastal regions, despite the fact that they are more exposed to foreign influences, are precisely those which have best preserved their typical festivals.

The bullfight. The bullfight, too, may be considered something of a hybrid celebration, since it possesses certain characteristics of a sport as well as those of a popular festival. It is customarily regarded, however, as the most typical of Spanish festivals and as the Spanish national fiesta par excellence, even though bullfighting is found in other countries, particularly those of Hispanic America. The tradition of bullfighting can be traced back to very ancient times, but the present style of fighting the bull on foot dates back only to the 18th century.

Other traditional features. There are seemingly innumerable other forms of traditional activity, and many objects still used in daily living also have their roots deep in the past. Some change is taking place, but other activities continue to exist as they have in the past, notably in the remote rural areas. Spanish houses, for example, reflecting as they do both the varying regional climates and the exigencies of a pre-industrial society, still continue to be constructed more or less as they were many years ago. On the other hand, the traditional or typical regional garb has undergone a slow evolution as to style, although certain costumes are still considered characteristic of the regions in which they originated.

Traditional arts and crafts, grouped together under their official name of *artesanía* (handicrafts), have become the object of government protection, designed to insure their continued existence in the face of the increasing mechanization of industry. Organizations devoted to this end include the Artisans' Trade Union, whose efforts have been complemented by the work of the National Handicraft Enterprise, a branch of the National Industrial Union. The promotion of handicrafts is also carried on by the Sección Femenina de Falange (Women's Division of the Falange Party), an organization which has also helped keep alive popular traditions in folkloric music, song, and dance through its performing groups, Coros y Danzas (Chorus and Dances).

POPULAR CULTURE

In Spain, cultural affairs, as distinct from educational matters, are handled by the Ministry of Information and Tourism through its Division of Popular Culture and Public Spectacles. The latter subsidizes various organized groups, such as Casas de Cultura—the Houses of Culture, which contain libraries, lecture rooms, and exhibition halls, and also give theatrical and musical performances—as well as sponsoring cultural events of many types. One of the moving forces behind these activities is a desire on the part of the government to make it possible for outlying sections of the nation to enjoy at least in part the cultural opportunities to be found in the large cities.

In the realm of theatre, the diffusion of culture is carried out through the national theatres, the festivals of Spain, and the theatre campaigns held in the provinces. The national theatres are: the Teatro Español, which stages premieres and also has a large repertory of both national and foreign plays; the María Guerrero Theatre, which presents modern and contemporary plays; the Teatro de la Zarzuela, which produces lyric theatre; and the Teatro Nacional de Cámara y Ensayo (the National Experimental Theatre), which specializes in the production of avant-garde plays. All of these theatres are located in Madrid.

The festivals of Spain, specially established to further the cultural decentralization of the country, have enabled a great number of people who would not otherwise have had such an opportunity to attend theatrical performances, dance exhibitions, and concerts. More than 100 such festivals are held annually throughout Spain.

Annual festivals that enjoy international prestige are the Musical Festival of Granada and the Musical Fortnight of San Sebastián. The International Theatre Festival, located in Madrid, also attracts highly qualified foreign theatrical companies. There are also other special festivals (such as that of the opera, also celebrated in Ma-

drid) held to meet needs that the cultural activities of private individuals have failed to fulfill.

This type of cultural promotion can only be carried out effectively if it is pursued on a continuing national basis. In 1965, therefore, the Ministry of Information and Tourism set up a national network of what were called Teleclubs. The ministry supplies to these centres (which numbered over 4,000 by the early 1970s), free of charge, a television set, daily and weekly newspapers, books, records, and various audiovisual aids. The Teleclubs are directed by a specially trained supervisor. Spanish publishing houses have also made a substantial contribution to the diffusion of popular culture by making high-quality, but moderately priced, books available on a mass basis. Annual book fairs are also becoming increasingly popular. The efforts of the Division of Popular Culture to raise the cultural level of the population are also furthered, it is held, by a variety of nation-wide campaigns.

The role of Teleclubs

CULTURAL INSTITUTIONS

The Academies and related bodies. Spanish cultural institutions have deep and rich roots in the nation's past. The oldest of the Spanish academies is the Royal Academy of Language, usually called the Royal Spanish Academy, which was established by King Philip V in 1713. Its primary purpose is "to cultivate and to set standards for the purity and elegance of the Castilian tongue." An essential aspect of this task involves the publication by the Academy of a Dictionary of the Language and other works, as well as the granting of prizes. Spain also possesses the internationally known Royal Academy of History; the Royal Academy of Fine Arts of San Fernando; the Royal Academy of Exact, Physical and National Sciences; the Royal Academy of Moral and Political Sciences; and the Royal National Academy of Medicine. All the academies form part of the Institute of Spain, founded in 1938 as a coordinating organization.

The Royal Spanish Academy

The Higher Council of Scientific Investigation, created in 1939, directs research activities of some 225 scientific centres located throughout the entire nation. The Institute of Hispanic Culture is a corporation dedicated to the preservation of the cultural community of *hispanidad* (*i.e.*, those of Spanish origin). It has as its chief purposes the study, defense, and diffusion of Hispanic culture and the promotion of mutual understanding among Hispanic peoples. The institute maintains cultural exchange departments, possesses a commissariat in charge of expositions and also the important Hispanic Library, the collection of which numbers more than 300,000 volumes. There are over 40 institutes of Hispanic culture located in 18 countries of North and South America. These collaborate with the parent institute in Spain, which has almost 20 affiliated Hispanic-American cultural associations within the country itself.

The famous athenaeums (which carry on varied cultural activities) are located in 20 different Spanish cities. Founded in the 19th century as private political clubs, they have since been gradually transformed into groups influencing the artistic and literary life of their home city. The Casas de Cultura, through lectures, expositions, concerts, and libraries, perform analogous functions in smaller towns.

Art galleries and museums. The great national art gallery is the world-famous Prado in Madrid. Installed in a magnificent building, it holds a unique place among the great museums of the world. Its collections and galleries have been enlarged in recent years by the acquisition of many of the Spanish paintings that were done in the 19th century.

Other important museums include the Museo de la Real Academía de Bellas Artes de San Fernando, Madrid, the Museum of Ancient Art in Barcelona, and those of fine arts in Bilbao and Seville. Collections devoted to specific aspects of the art of Spain are found in other museums: thus, religious sculpture is housed in the National Museum of Sculpture in Valladolid; and the art of El Greco is recognized in his museum and house in Toledo.

The Army and Naval Museums in Madrid are interesting from the historical viewpoint, as is the Maritime Museum in the royal arsenals of Barcelona. In addition to numerous provincial and municipal museums, there are others belonging to the cathedrals, dioceses, and parishes. The various museums of the National Patrimony, housing collections that form part of the former royal patrimony, are also very significant. Finally, the country is rich in private museums, including the Cerralbo in Madrid and the Marés in Barcelona. New museums include the Sephardic Museum in Toledo; the Museum of Costume and the Picasso Museum in Barcelona; and the Museum of Abstract Art in Cuenca.

It is notable, however, that in a country that possesses great artistic and historic treasures and in which art, particularly religious art, is alive and active *in situ*, the activity of museums of necessity plays only a secondary role.

Libraries and archives. The most important of Spain's famous historical libraries is that of El Escorial, which served as a model for other European libraries up until the 19th century. The Royal Library, established in the 18th century, was in the following century transformed into the National Library of Madrid. It has very valuable collections that originally belonged to private libraries or to monasteries that were confiscated by the government. The Private Royal Library, or Palace Library, the libraries belonging to the royal academies, that of the Higher Council of Scientific Investigation, and the public libraries of the provinces form a broad library network throughout the entire country.

The Royal Library

In addition to a great number of smaller archives—regional, local, special, ecclesiastical, and private—Spain has four great general archives in which are assembled the historic documents of the nation. They are: the National Historical Archives, in Madrid; the General Archives of Simancas; the General Archives of the Indies, in Seville (which contains documents relating to the government of the Spanish overseas empire); and that of the Crown of Aragon (of supreme importance for the study of medieval history), in Barcelona.

Other cultural bodies. The National Orchestra, which was founded in the period immediately following the Civil War of 1936–39, has devoted itself chiefly to the promotion of interest in symphonic music. The Symphony Orchestra of Spanish Radio and Television, which gave its first concert in 1965, has, in its shorter existence, been able to reach a wide audience through the use of the communications media. Both these orchestras are located in Madrid, as are two others with a long and brilliant history: the Symphony and the Philharmonic. Outside the capital, the principal Spanish orchestra is the municipally supported City Orchestra of Barcelona.

Spain has only one theatre for opera that has a regular season for its performances, the Lyceum Theatre in Barcelona, a private enterprise. The Festival of the Opera, however, which, in some respects, corresponds to a short opera season, is held every year in Madrid. Similar festivals, also of brief duration, are held in several of the provincial capitals.

The March Foundation (founded 1955) is outstanding among foundations under private auspices. Its cultural activities include the giving of prizes, grants for research, scholarships for study in Spain or abroad, and annual contributions to the study of literature and the fine arts, as well as subsidies to institutions and enterprises interested in culture. It also publishes books.

The influence of private foundations

The number of literary prizes established has increased notably in recent years; the 17 national prizes for literature remain, however, as the oldest and most distinguished. Other important prizes include the "Nadal" and the "Planeta" for the novel; and the "Adonais" for poetry.

COMMUNICATIONS MEDIA

The press. The promulgation of the Press Law in 1966 led to a profound change in the journalistic world of Spain. The stationary situation that prevailed under the press regulations established in 1938 had, indeed, shown a certain tendency toward change in recent years, but it was not until the passage of the new law that press moved

into a phase of extraordinary activity. The three basic principles enunciated in the new law—liberty of expression, liberty of enterprise, and free choice of a director—are said to have changed the relationship that formerly existed between the state and the press. Now that government controls have been reduced and the exercise of them is subject to juridical norms, the Spanish press, it is generally acknowledged, has acquired a quite remarkable dynamism, and the number of publications has increased. By the early 1970s, almost 100 morning papers were published in Spain, along with several dozen evening ones.

Radio broadcasting. The first radio broadcasts made in Spain were carried out by an experimental station in Madrid in 1923, and in the following year, the first transmitters—privately owned and operated—were installed in Barcelona and Madrid.

After the Civil War had ended in 1939, a national radio broadcasting service was established, centred on the group of transmitters located in Arganda, in the province of Madrid. In 1951, when the Ministry of Information and Tourism was created, the government-operated radio and television stations were placed under its direction. Technical improvements were made, and a national plan for radio broadcasting was worked out.

In 1960 the first station of Radio Peninsular devoted entirely to the broadcasting of music was set up, and, since then, stations of a similar type have been established in several of the provincial capitals. The first FM transmitter was installed in Madrid in 1957 by Radio Nacional de España, to be followed by several others of the same type. All private transmitting stations broadcast simultaneously in FM and AM.

State and private broadcasting

Before 1936 the only radio broadcasting done in Spain was under private auspices; privately owned stations now coexist with the official network. Almost all the privately owned transmitters, of which there are almost 200, are organized into the SER Chain (Spanish Radio-broadcasting Society). They are either regional or local stations. The transmitters that belong to the National Movement and to the Catholic Church, as well as those operated by private persons, are not under state control.

Television. The state has a monopoly over television programming, and the first daily program was set up in Madrid in 1956. By the 1970s the Spanish television network covered the entire national territory through its "First Program," utilizing microwave links. The "Second Program," initiated in 1965, operates on the uhf band. It should be noted that the complicated topography of the country makes television broadcasting on a national scale difficult.

In 1964 a production centre for television programming was installed and inaugurated in Prado del Rey. Its purpose is to give general direction to the many hours of television programmed on the Spanish network. By the early 1970s there were five such centres in existence, with a total of 17 studios, but colour television in Spain is still in the developmental stage, although well over 40 percent of Spanish families possessed television sets.

FUTURE PROSPECTS

In the Spanish viewpoint, the expectations for normal political development for Spain in the years ahead are based on a continuation of current trends. The geographical location of Spain—with links to the European, Mediterranean, African, and American worlds—will doubtless continue to exert a powerful influence. Irreversible trends toward modernization and industrialization, already in evidence, will, it seems, lead to further links with the European community. The future of the country will depend heavily on the successful blending of a venerable tradition with the technological and industrial traditions that have already come to dominate northern Europe but from whose influence Spain has hitherto been isolated by virtue of its marginal position. (E.L.O.M.)

BIBLIOGRAPHY

General works: The *Anuario Estadístico de España*, published annually by the Spanish Government, contains general statistical information about the country. The *Enciclopedia Universal Ilustrada Europeo-Americana*, 70 vol. (1908–30, and supplements), provides detailed information about Spain, its regions, and its towns, especially in vol. 21, *España*, and under "España" in the supplements. PASCUAL MADOZ, *Diccionario geográfico-estadístico-histórico de España y sus posesiones de ultramar*, 16 vol. (1846–50), is still a valuable reference book. STEPHEN CLISSOLD, *Spain* (1969), is an informative book about the country and its regions, the history of Spain, and its present political and economic situation. JAMES MICHENER, *Iberia: Spanish Travels and Reflections* (1968), provides an impressionistic and largely subjective vision of the country.

The Land: General information about Iberian geography may be found in books about the Mediterranean area, especially JAMES MCINTOSH HOUSTON, *The Western Mediterranean World: An Introduction to Its Regional Landscapes* (1964); and MAX SORRE, *Peninsules mediterranéennes*, vol. 1, *Généralités. Espagne et Portugal* (1934). MANUEL DE TERAN (ed.), *Geografía de España y Portugal* (1952–58), dedicates its first four parts to Spain and is the most complete work on the subject. HERMANN LAUTENSACH, *Iberische Halbinsel* (1964), constitutes a comprehensive geography of the Iberian Peninsula. WILLIAM BAYNE FISHER and HOWARD BOWEN-JONES, *Spain: An Introductory Geography* (1966), is a general manual. EDUARDO HERNANDEZ-PACHECO, *Fisiografía del solar hispano*, 2 vol. (1955–56), gives an extensive account of Spanish physical geography. ADOLPH SCHULTEN, *Iberische Landeskunde. Geographie des antiken Spanien*, 2 vol. (1955–57), is an exhaustive scholarly work on the geography of the Iberian Peninsula in early historic times.

The people: The FUNDACION FOESSA, *Informe sociológico sobre la situación social de España* (1970), is a detailed survey of Spanish society, with extensive statistical information. JULIO CARO BAROJA, *Los pueblos de España. Ensayo de etnología* (1946), is a well-documented account of the development of the Spanish people, from prehistoric times to present-day ethnological regions.

The Economy: A general survey in English of the Spanish economy is GEORGE HILLS, "Spain Under Franco: The Economic 'Miracle' 1959–67," in his *Spain* (1970). The Organization for Economic Cooperation and Development publishes occasional economic surveys of Spain in several languages, which contain much basic information. RODNEY H. MILLS, *The Spanish "Miracle": Growth and Change in the Spanish Economy, 1959 to mid–1965* (1966), is a staff economic study for the Board of Governors of the Federal Reserve System. BUSINESS INTERNATIONAL, S.A., *Spain in Transition: Patterns of Growth, 1968–1971* (1968), is a survey for business readers. RICHARD HUMBERT, *Spain* (1970), is a similar study published by the United States Bureau of International Commerce.

Transportation: Las Obras Públicas (annual), an account of the activities of the Ministry of Public Works for the year; INSTITUTO NACIONAL DE ESTADISTICA, *Primera Encuesta Nacional sobre transporte de mercancías por carretera* (1970), contains the results of a survey made during 1969 on highway goods transportation; JOSE IGNACIO RAMOS TORRES and MARCOS MARTIN BLANCO, *El Transporte en España: Organización y magnitudes económicas* (1966), a study of the Spanish transportation system, with emphasis on economic and legal aspects; COMISARIA DEL (II) PLAN DE DESARROLLO ECONOMICO Y SOCIAL, *Transportes* (1967), and official analysis of the evolution, problems, objectives, investments, and operations of the Transport Division.

Administration and social conditions: Anuario Político Español (annual), a nonofficial publication, political in nature, that examines social and economic questions; MINISTERIO DE EDUCACION Y CIENCIA, *La Educación en España: Bases para una política educativa* (1969), the so-called "White Book," which outlines in detail the measures to be carried out in compliance with the General Education Law of 1970; INSTITUTO NACIONAL DE ESTADISTICA, *Encuestas de presupuestos familiares* (1970), contains the results of surveys on Spanish family budgets made in 1967 and 1968; RODRIGO FERNANDEZ-CARVAJAL, *La Constitución Española*, 2nd ed. (1969), a well-ordered explanation of the basic principles of the Spanish constitution; SERVICIO INFORMATIVO ESPANOL, *Leyes fundamentales del Estado* (1967; Eng. trans., *Fundamental Laws of the State: The Spanish Constitution*, 1967); *Informe económico* (*Economic Report*), a general review of the economy issued annually by the Bank of Bilbao, containing interesting information on prices and salaries; FUNDACION FOESSA, *Informe sociológico sobre la situación social de España* (1970), detailed, if dense, itemized statistical information; MINISTERIO DE LA GOBERNACION, *Legislación electoral española* (1971), a legislative handbook on the Spanish electoral system; RAMON MARTIN MATEO, *Manual de derecho*

administrativo (1970), a complete summary of the Spanish system of administration; *Plan Nacional de la Vivienda: 1961–1976* (1962), an explanation of the 15-year National Housing Plan; "La Vivienda," in *El Gobierno informa*, vol. 2 (1964), an official document of housing program history.

Cultural life and institutions: *Anuario de la radio-televisión española* (annual); CARLOS AREAN, *1971: Balance del arte joven en España* (1971), a brief treatment of the most recent tendencies in Spanish art, architecture, painting, and sculpture; ANIBAL ARIAS RUIZ, *La televisión española* (1970); JUAN ANTONIO GAYA NUNO, *Historia y guía de los museos de España*, 2nd ed. (1968), an excellent overall view of Spanish museums; and the same author's *La pintura española del siglo XX*. (1970); JUAN JOSE TOMAS MARCO, *Música española de vaguardia* (1970); DIRECCION GENERAL DE BELLAS ARTES, *Museos y colecciones de España* (1969).

(M.Me./E.L.O.M.)

Spain, History of

This article traces the history of the Iberian Peninsula from prehistoric times to the present. Portuguese history from the establishment of the independent monarchy in the 12th century AD is treated in PORTUGAL, HISTORY OF.

This article is divided into the following sections:

I. Spain to *c.* AD 500
- Pre-Roman Spain
- Roman Spain
- Visigothic Spain to *c.* 500

II. Christian Spain, *c.* 500–1479
- The Visigothic kingdom
- Christian Spain from the Muslim invasion to *c.* 1260
- Christian Spain, *c.* 1260–1479

III. Muslim Spain
- Political history
- Social and economic life
- The culture of Muslim Spain

IV. Spain from 1479 to 1700
- United Spain under the Catholic kings
- Spain under the Habsburgs

V. Spain since 1700
- The early Bourbons, 1700–53
- The reign of Charles III, 1759–88
- Charles IV and the French Revolution
- The French invasion and the War of Independence, 1804–14
- Ferdinand VII, 1814–33
- Isabella II, 1833–68
- The Revolution of 1868 and the Republic of 1873
- The restored monarchy, 1875–1923
- Primo de Rivera (1923–30) and the Second Republic (1931–36)
- The Civil War
- Franco's Spain, 1939 to present

I. Spain to c. AD 500

PRE-ROMAN SPAIN

Prehistory. The earliest human fossils found in the Iberian Peninsula are those of Neanderthaloid men, dating from about 200,000 BC, whose remains have been discovered at Gibraltar, Játiva (Valencia), and Bañolas (Gerona). But utensils found in Algeciras, Madrid, the Provincia de Soria, Lisbon, and elsewhere point to the existence of even earlier habitation. This indirect evidence suggests that the peninsula was inhabited as early as 500,000 BC. The most important sites are the camps of elephant hunters found at Madrid, Torralba, and Ambrona (Soria), where many animal bones have been discovered. The Neanderthaloids and other pre-*sapiens* disappeared with the last glaciation, when Old Stone Age (Paleolithic) cultures advanced in the peninsula.

The Paleolithic period

One of the most remarkable achievements of the Upper Paleolithic was in the realm of art. The famous Paleolithic cave paintings found at such sites as Pindal, Peña de Candamo, Castillo, El Pendo, Pasiega, Ribadesella, Altamira, and elsewhere in the peninsula represent a broad historical period extending roughly from 25,000 to 10,000 BC. The animals depicted in the paintings are cold-weather varieties—bison, bulls, horses, goats, wild boars, bears, lions, mammoths, reindeer, rabbits, and seals.

The transition to the Neolithic Period coincides with a climatic change the full effects of which came about gradually (beginning around 10,000 BC) and drove the food animals northward. The archaeological deposit at Muge, near Lisbon, undoubtedly belongs to the beginnings of this age, which, in its early stages, is sometimes called Mesolithic. Deposits of great quantities of ocean and river shells, Crustacea, and the remains of birds, fish, oxen, and swine attest to the diet of these early Neolithic men. In Santimamiñe (Vizcaya), a similar economic pattern is found. There are also two important cave sites near Valencia: at Las Cuevas del Oro and at Dos Aguas. At the height of the Neolithic Period, approximately between 4000 and 3000 BC, rock painting flourished in the region of the Levantina, appearing in the open air, sheltered by rock overhangs. Many isolated animal figures are depicted, but human beings are also seen, comprising narrative scenes of hunting, herding, dancing, fighting, and even honey gathering from a hive located on top of a rock. The figures are small and exhibit considerable movement and vitality. The men are shown using bows, and the women usually appear wearing only a long, bell-shaped skirt. The Neolithic Period in western Europe, however, lagged far behind that of the eastern Mediterranean. In the West, cave dwelling persisted, and although agriculture and husbandry were known, their practice was only minimal. Settlements were quite rudimentary, being no more than simple groupings of shelters. This was an exclusively coastal culture and did not extend inland.

The age of metals

The Neolithic came to an abrupt end around 2500 BC with the beginning of the use of metals in the south. It was at this time that the first large population centres were formed, along with monumental burial sites; it was at this time also that agriculture, cattle raising, mining, and metallurgy (beginning with copper, then bronze) underwent a period of extraordinary development. It appears that these arts may have been brought from the Aegean, not, however, as the result of an invasion but apparently as the result of a sporadic penetration of "technicians." Concurrently, maritime trade flourished, thereby explaining the presence of artifacts from all over the Mediterranean, among them ivory, turquoise, kallaite, obsidian, and amber (perhaps from the Baltic). The new economy brought forth a new social organization, founded in a type of monarchy, the support of which came from an aristocracy based on wealth, who lorded it over a more or less enslaved population. This culture, the Almería, ranged over the south of the peninsula, from Algarve to Cabo de la Nao and from the Strait of Gibraltar to the Tagus River, extending inland in the Levantina and the Meseta Central. The high point in the development of this culture was reached between 3000 and 2000 BC. The best known centres were those in the southeast (Almizaraque and Los Millares) and in the southwest (Vilanova de San Pedro and Zambujal). Settlements with round or oval houses were built on fortified hilltops, indicating a life of constant danger. In Los Millares, the people of the Almería culture raised wheat, barley, millet, rye, lentils, chick-peas, grapes, and olives; as domestic animals they had sheep, swine, oxen, horses, and dogs. Among their usable plants were linen and cotton for fabrics and esparto for baskets, shoe soles, ropes, and nets. At the same time, copper was being used. Contemporary with Los Millares was the mining centre at Almizaraque, where copper and silver were processed, as evidenced by crucibles and slag heaps found next to the mines. This was the culture that built large megalithic burial vaults with corbel-arched ceilings. These tombs are entered from the outside by a tunnel and are covered by a cone-shaped mound. Tombs of this type are found at El Romeral (Antequera), Matarrubilla (Seville), and Soto (Huelva), and there are partial remains at Alcalar (Algarve). In contrast, the dolmen at Viera (Antequera) consists of a rectangular chamber, and the famous "cave" at Menga (Antequera) is slightly oval. The latter is the most remarkable monument of its kind. In its construction, 31 enormous stone slabs were used, weighing a total of 1,600 tons. The flat roof was supported by three thick pillars. The dolmen culture spread through the rest of the peninsula and through Europe, giving rise to distinct dolmenic "provinces" in the central and Atlantic zones. These cultures also reached Italy, Corsica, Sardinia, and

North Africa. Accompanying this expansion was an excellent ceramics industry characterized by campaniform, or bell-shaped, utensils, incised with geometric designs.

El Argar culture

Toward the beginning of the 2nd millennium, in the same region (Almería), El Argar culture appeared, in which the metallurgy of bronze, silver, and gold notably flourished. This industry was no longer in the hands of foreigners but constituted instead a native enterprise that produced for domestic consumption (weapons, jewelry, utensils) as well as for trade abroad. The result was a lively trade with other centres in the eastern Mediterranean. But there were substantial differences from previous cultures: dwellings had a rectangular floor plan, and the dead were buried individually in baskets or large urns beneath the floors of the houses, not in collective vaults, as in the previous epoch. Villages continued to be built on hilltops and were well fortified. Governed by kinglets or princes who were supported by an aristocracy, this culture, which at first was limited to the southeast, spread throughout the southern, the central, and the Levantina regions and reached the Balearic Islands. Its apogee came roughly between 1700 and 1000 BC. It was around this time that the kingdom of Tartessus appeared, along with the first Phoenician and Greek colonies.

Tartessus. Near the beginning of the 1st millennium, Spain entered written history. Passages in the Bible and in Greek and Latin texts refer to Tartessus (biblical Tarshish), an important centre of trade, especially in tin. The Tartessians were by this time sailing the Atlantic, southward along the African coast as well as northward in search of tin in the northeastern part of the peninsula and in the Cassiterides Islands (Scilly Islands). With this supply of tin and with the copper found around Huelva, a strong bronze industry developed that was mentioned by the Greeks. The Tartessians had a writing system, and while inscriptions have survived, they have not yet been translated. They also had a literature, but nothing of it remains. Their government was a hereditary monarchy, and while the names of several of the Tartessian kings are known, only one, Arganthonios, has historical status. Arganthonios, who reputedly lived 120 years, died in the middle of the 6th century BC, shortly before the Battle of Aleria (Alalia *c.* 540), whereby the Carthaginians gained control of the Mediterranean. His friendship with the Greeks and with the Phocians, to whom he loaned money to fortify their city against the Persians, brought reprisals against Tartessus by the Carthaginians. The exact location of Tartessus has yet to be established, and even the ancient writers themselves were in doubt about it. All that can be said at present is that it was undoubtedly situated along the coast between Huelva and the Strait of Gibraltar, perhaps in the direction of the mouth of the Guadalquivir or the Guadalete rivers.

Arganthonios

Phoenician colonization. Textual sources reveal that Phoenicians from Tyre (Sūr, Lebanon) arrived on the peninsula shortly before 1100 BC, the traditional date of the founding of Gadir (Cádiz). This date, established by the Greek philosopher Poseidonius in Cádiz itself, is historically acceptable, though some historians place the Phoenician landing in the 8th century. The Phoenicians were attracted by the thriving trade in metals at Tartessus, the flourishing agriculture of Andalusia, and the rich fishing business of the coast that stimulated such allied industries as preserving, salt evaporation, and the extraction of purple dye from shellfish.

The Phoenician colonies founded after Gadir were Baria (Villaricos), Abdera (Adra), Sexi (Almuñécar), and Malaca (Málaga), all of which minted their own coins, although not until relatively late. Many processing and preserving plants for fish were established along the entire southern coast, and scattered agricultural centres were established inland. Ibiza was founded as a fishing colony by the Carthaginians in 654, though the dates of the founding of the rest of the colonies are not known. There is no definite information about Cádiz from before the 6th or 5th centuries. Hoards of jewelry and a human-shaped sarcophagus in the Phoenician style have been found in graves there, but nothing remains of the great Temple of Melqart (Hercules), and many other monuments have been lost to the destructive action of the sea along the Cádiz coast. At Baria many artifacts have been found, along with great numbers of painted ostrich eggs. Ibiza's many underground chambers at Puig des Molins have yielded terra-cotta objects, necklaces of glass paste beads, and Greek vases. Important statues of Tanit have been found in the temple at Cuyram, and terra-cottas of a very primitive type have been uncovered at Isla Plana. But the most spectacular finds have been made at the necropolis at Sexi, where Egyptian alabaster amphorae, dating from the 10th century, have been found in tombs from around 700 BC. Inland, engraved combs have been found in tombs at Carmona and Osuna; also important are the treasury found at La Aliseda (Cáceres) and at Sines and the many series of bronze pitchers found all over the western part of the peninsula, from the Strait of Gibraltar to the interior of Castilla (Castile). All the artifacts date from the 7th century BC.

After the fall of Tyre to the Assyrians, the Phoenician colonies were taken over or protected by the powerful Phoenician daughter city of Carthage in North Africa. The Carthaginians quickly turned the resources of the peninsula to the task of creating a staging area for attacks against Rome. Hamilcar landed in Spain in 237, taking possession of the entire south and southeast of the peninsula, where he founded Akra Leuke (Alicante). Shortly thereafter, Hasdrubal founded Carthago Nova (Cartagena), and Hannibal readied his attack against Rome after assuring the subjugation of the native peoples of both mesetas. In 219 Hannibal's successful siege of Saguntum (Sagunto), an ally of the Romans, precipitated the Second Punic War (see PUNIC WARS). The Carthaginian conquest of Spain provided the Carthaginians not only with the silver necessary to finance the wars but also with a nearly inexhaustible supply of mercenary soldiers. Hannibal crossed the Alps in 218, attacking the Romans on their home ground. But even as Hannibal ravaged the Italian peninsula, the Romans attacked the Carthaginians in Spain. In 218 the Roman general Publius Scipio landed at Emporia and achieved a temporary success. In 210, Scipio Africanus, the son and namesake of Publius Scipio, took command and won a series of brilliant victories. Scipio first took Tarraco, then Carthago Nova (209), and after a victory at Ilipa (Alcalá del Río, near Seville), he overran Cádiz (206), marking the end of Carthaginian domination in Spain.

The Punic Wars

Greek colonization. Thanks to the greater number of written accounts, much more is known about the Greek colonization. In the 9th and 8th centuries, Greeks were sailing the coasts of the peninsula, and their explorations paved the way for the subsequent arrival of the Phocians. According to the Greek historian Herodotus, the first Greek to reach Tartessus was Colaeus of Samos in the mid-7th century. As early as the 7th century, the Phocians founded Hemeroscopium (near Denia) and Mainake (near Málaga). These colonies maintained close contact with the Tartessians, whose king, Arganthonios, protected and aided them. It was probably also at this time, or shortly thereafter, that Meidakritos, Euthymenes, and Euphemus of Caria sailed along the Spanish coast. Toward 600 BC, the Phocians founded the city of Massilia (Marseille), opening the way for new settlements in the northeast. Emporiae (Ampurias) appeared at this time, and the Rhodian colony of Rhodae (Rosas) was populated by Phocians. Both cities minted their own coins, beginning in the 5th century. The Battle of Alalia (*c.* 540) brought not only the fall of the kingdom of Tartessus but also the collapse of the Greeks' two southernmost colonies, Mainake and Hemeroscopium. Later, with the arrival of Greeks from Sicily and Magna Graecia (southern Italy), contacts with the northeastern coast of Spain were re-established, and there followed the founding of Alonae (Benidorm?) and other trading centres. Of these colonies, only one, Emporiae, has been identified and excavated. Artifacts found there date from the end of the 7th century, and among them are abundant ceramics, representing all Greek styles, along with terra-cottas, figured bronzes, and coins. The most outstanding piece of sculp-

ture is a large marble statue of Asclepius, dating from the 4th century. Few Greek inscriptions, however, have been found. Only the above-mentioned places are referred to as colonies, but textual evidence reveals a considerable number of Greek place names, all of them along the coast, which must refer to trading posts or even smaller population centres. Active trading flooded the Spanish coasts with Greek products. From the Pyrenees to the mouth of the Tagus, great numbers of Greek ceramics have been and continue to be discovered, ranging in date from the 7th century to the Hellenistic era. Many bronze statues, some of them quite large and dating as far back as the 6th century, have also been found, as well as two ancient Greek helmets, also made of bronze. Finally, coins from throughout the Greek world, from Massilia to the Crimea, have also been discovered. These coins span the entire period from the earliest mintings up until the Alexandrine era, with those from Greece, Sicily, and southern Italy in particular abundance.

The Iberians. The broad designation of Iberian culture is generally applied to the civilization that grew up in the south and east of the peninsula during the five or six centuries prior to the birth of Christ. The designation Iberian is derived from the usage of contemporary classical writers. From Neolithic times forward, there were no cultural changes that necessarily imply any change in the ethnic composition of the peninsula. The only important outside influences that brought about any change are those from the East, beginning with the Punic and, later, the Greek settlers. It may be assumed, therefore, that the Tartessians and other peoples who inhabited the Mediterranean coast of Spain were descended from the Neolithic and Chalcolithic peoples of the peninsula, whose direct ancestors may have been closely linked to prehistoric peoples who inhabited "white Africa" from the Red Sea to the Atlantic and the Canary Islands. During the 1st millennium, the ancient Iberians, the creators of the great cultural centre of Tartessus, were subject to the pressure of the Greek and Phoenician–Carthaginian colonization. These maritime influences gave rise in later years to a singular and progressive culture that may quite correctly be called Iberian.

Origins of the Iberians

Iron was in widespread use among the Iberians, who developed a complex economy based on agriculture, stock breeding, fishing, industry (mining, metallurgy, weapons, jewelry, preservatives, textiles, ceramics, thrown pottery), and including general commerce, both domestic and foreign. Cities grew apace, and almost all of the Iberian cities mentioned by ancient writers were highly developed before the Roman conquest. Houses were rectangular in shape. Cremation was universally practiced, and cemeteries often contain rich furnishings, such as Greek vessels, Punic artifacts, damascene weapons, and native pottery. Political organization was monarchical, with a king, princes, and dukes assisted by a council of elders. But there were also popular assemblies in times of stress. In general, these kings were politically independent, although sometimes a single king controlled many cities in military alliance. The Iberians possessed written laws, apparently in verse form, and were highly accomplished in several of the arts, including music and the dance.

Cities were built on well-defended hillocks; individual houses, with straw roofs, were made from rubblework and molded adobe bricks. To judge from the large stone cellars that have been found in Andalusia—in Tutugi (Galera, Granada), Tugia (Toya, Granada), Basti (Baza, Jaén), Almedinilla, Iponuba (Baena, Córdoba)—there must have been public buildings of monumental dimensions, although nothing remains of them. Stone sculpture shows a strong early Greek influence, attested to by the sphinxes found at Agost and El Salobral, the androcephalic bull from Balazote (Albacete), and griffins like the one found at Redován. Temples, particularly those at Albacete, have yielded hundreds of large human figurines, both male and female, in both standing and sitting positions. In general, these are votive statues, used to ward off evil. Among them, the most important are several heads, parts of full-length figures, and the "Gran Dama Oferente" from the Cerro de los Santos. Above all, there is the famous bust of the "Lady of Elche," apparently of late date, perhaps from the 1st century BC, adorned in gold.

The "Lady of Elche"

It is known that large-scale mural painting was practiced, depicting scenes of hunting, fishing, warfare, and civil ceremonies, but only scant indications of its existence have survived. Some idea of what it was like may be seen in the painted ceramics from Azaila, Numantia, and Liria. This thrown pottery, of excellent workmanship and exhibiting a surprising variety of forms, first appeared with decorations consisting of simple straight-line and circle designs and geometric motifs. But in its final period, corresponding with the wars of resistance against the Roman conquest in the 2nd and 1st centuries BC, this ceramic art was characterized by the depiction of narrative scenes embracing a great variety of decorative motifs, ranging from war and large game hunting to domestic or popular scenes, such as processions, dances, and the earliest known bullfights. These scenes were rendered in a somewhat childlike linear style but with considerable narrative strength. Iberian ceramics were exported to southern France, Liguria, central Italy, Sicily, and North Africa in the 1st century.

Iberian silverwork and goldwork was exceptionally rich; silverwork predominated in the fashioning of vases, paterae (earthenware drinking saucers), and jewelry. Decorative patterns were inlaid on weapons, dishes, and the like, and at times Iberian inscriptions were added. The most important known Iberian treasures are those of Tivissa (Tarragona), which include a beautiful set of dishes, its various pieces nielloed, or engraved, with religious themes. Other Iberian treasures have been found at Abengibre, Salvacañete, Mogón, and Santisteban. Outstanding pieces of goldwork were found in the treasure of Santiago de la Espada (Jaén) and are similar to the jewelry worn by the "Lady of Elche."

The Iberians developed early a system of writing. The Iberian alphabet has an older (6th–5th century) Western phase called Tartessian and a derived Eastern phase called Iberian. There are both Greek- and Phoenician-derived letter forms in the 28-character alphabet, most of which is of unknown origin. Many inscriptions exist, but few words—except place-names on the coinage, struck by many cities in the 3rd century—can be read. No bilingual inscriptions have been found. The Iberians retained their writing system until the Roman conquest, when the Latin alphabet came into use.

The Iberian language

The Celts. While the Iberian culture of the south and the east flourished under the influence of colonizers from the eastern Mediterranean, in the central and western parts of the peninsula, the cultural level was still quite primitive and rooted in the Copper and Bronze ages. But the 9th through the 8th centuries witnessed the gradual and peaceful penetration of the peninsula by peoples from across the Pyrenees from central Europe. These peoples, named Celts by Greek and Roman writers, were primarily herdsmen and agriculturalists—and they were able to smelt iron. In spite of constant reinforcement from new migrations, these people were often absorbed into the native population. This gave rise in a mixed people known as the Celtiberians. In other cases, the native cultural identity was erased almost in its entirety and was supplanted not only by Celtic customs and beliefs but also by the Celtic language, as is attested by the vast number of place-names and family names in certain areas. This new culture never reached the level of the coastal Iberian culture. The Celts did not have their own system of writing; they did not strike coins; nor did they practice sculpture or painting. The Celts did, however, excel in pottery techniques and in ornamental design, characterized by elaborate geometrical patterns. Their weapons and jewelry were exceptionally well made, always from gold (bracelets, torques), and they had mastered such difficult techniques as damascene. The round-shaped dwelling was replaced by the rectangular floor plan. But in the northwest (Galicia, Asturias, and the north of Portugal), the Celtic culture remained virtually unaffected by the native culture it encountered. The houses con-

tinued to be built with a circular floor plan, and the dead were buried beneath the floor of the dwellings themselves, reminiscent of the prehistoric cultures found at Los Millares and El Argar, and they continued in this area, at least with respect to the round floor plan, until well into the Middle Ages and are even found surviving today, although only in remote areas. This "Round House Culture" is also referred to as the "Castro" culture, after the arrangement of towns on fortified hilltops.

ROMAN SPAIN

The conquest. Scipio's landing in Emporia (218) during the Second Punic War began the Roman conquest of Spain. The Carthaginians were expelled by 206, and over the next two centuries the Romans, against a grudging and stubborn resistance, extended their hegemony over the peninsula. It is significant that in the same year that Cádiz fell, Italica, the first Roman settlement, was established there.

Native uprisings

The harshness of the Roman conquest provoked a large-scale uprising among the Lusitanians (Lusitani) and the Celtiberians that threatened Rome's hegemony in the peninsula. But Rome's vast resources and its superior command ended, first, with the assassination of the rebel leader Viriathus (Viriato) in 139, and, later, with the siege of Numantia, which fell only after a lengthy resistance that required Rome to use 60,000 men against a city of only 4,000 (133). The presence of Scipio Africanus, the adopted son of the Scipio who defeated Hannibal, was even called for before the city fell at last. In 138 the Romans arrived in Galicia (Gallaecia), in 121 they landed in the Baleares (Balearic Islands), and in 105 they drove out the Cimbrians and the Teutons who had made their way into the peninsula.

Early in the 1st century BC, political wars between Marius and Sulla in Rome brought about the revolt of Sertorius, who was on the verge of creating in Spain an independent state, having organized the native populations along Roman lines and having fought successfully against Metellus and Pompey. But the war became lengthy, and rivalries began to erode the original enthusiasm for the cause. The assassination of Sertorius allowed Pompey the final victory in this civil war in 72. The wars between Caesar and Pompey also had their repercussions in Spain. The Battle of Ilerda (Lérida), in 49, gave Caesar his decisive victory and put all of Spain into his hands. But Pompey's cause had many followers in Baetica (Pompey had died in 48), and Pompey's sons Sextus and Gnaeus took refuge there after the defeat at Thapsus (46). With the remaining members of the defeated armies in Spain and many partisans, they were able to build a powerful army. But the sudden arrival of Caesar put a quick end to this dangerous faction in the Battle of Munda (45), where Caesar won a decisive victory. Gnaeus died while fleeing, and Sextus succeeded in escaping from Spain. The armies in these civil wars were made up for the most part of Spaniards, either as legionnaires or as auxiliaries. This fact, along with the great zeal among the adherents to either side of the issues, is evidence of the intense Romanization that had taken place in much of the country.

The Cantabrian War

The final stage of the Roman conquest was the Cantabrian War (29–19 BC) launched by the emperor Augustus. The assault on the Cantabrian mountains was carried out from three bases of operations: Segisama, for the eastern sector (Santander); Asturica, for the central (Asturias); and Bracara Augusta, for the western (Galicia). Seven legions participated, which, along with auxiliaries, amounted to some 70,000 men. The Emperor himself accompanied these forces, aided by the generals Agrippa, Carisius, and Antistius. A naval squadron kept watch on the coast. After the Cantabrian War, the whole of Spain was under Roman domination, and peace was interrupted only by events occurring outside the country.

One such occurrence was the revolt of Galba against Nero in AD 68. The uprising was begun by Vindex in Gaul, but the central figure was Galba, at that time governor of Tarraconensis. From that time on, the Roman legions took more responsibility than the Senate for naming the emperor, and they no longer took into account the place of the emperor's birth or his social class. The 2nd-century Roman emperors Trajan and Hadrian, both born in Spain, were the first two emperors of provincial origin to be elected by the Senate. An event of great importance during this time was the Moorish invasion of 172–175. Armies of Moors crossed the Strait of Gibraltar, invaded Baetica and Lusitania, and laid siege to Singlia Barba and Italica. At the end of the 2nd century, struggles took place for succession to the throne after the death of Commodus. Supporters of Albinus rallied part of the auxiliary forces of the Seventh Gemina Legion (Legio VII Gemina), then quartered in what is now León, against Septimus Severus. Around the middle of the 3rd century the Franks entered Spain, took Tarraco, and, after crossing the Strait of Gibraltar, drove into Mauretania, where they were dispersed.

Administration. The Romans first divided provincial Hispania into two sections, Ulterior and Citerior, but this arrangement underwent certain alterations by Augustus,

From F. Lot, *The End of the Ancient World;* Routledge & Kegan Paul Ltd., London (published in the U.S. by Barnes & Noble, Inc.)



Spain at the end of the 4th century AD.

The provinces

who created three provinces: Hispania Tarraconensis, Hispania Baetica, and Hispania Lusitania. Their respective capitals were in Tarraco (Tarragona), Corduba (Córdoba), and Emertia Augusta (Mérida). In the late 3rd century, Diocletian made Spain into a single diocese and subdivided Tarraconensis into three new provinces: Tarraconensis, Carthaginiensis (along with the Baleares), and Gallaecia. In addition, he joined to the Diocesis Hispania the new Provincia Mauretania Tingitana (Tingis: Tanger), which took in all of northern Morocco. Of the three Augustan provinces, Baetica, which was considered pacific, had senatorial status and was administered by the Roman Senate through a proconsul. There were no occupational troops in such a province. The others were "imperial" provinces and were thus governed through delegates appointed by the emperor and were controlled by occupational troops.

Justice was administered through *conventus juridici.* Baetica had four such administrative districts: Astigitanus, Cordubensis, Gaditanus, and Hispalensis. Tarraconensis had seven districts: Tarraconensis, Caesaraugustanus, Carthaginiensis, Cluniensis, Asturicensis, Lucensis, and Bracaraugustanus. Lusitania had three: Emeritensis, Scalabitanus, and Pacensis. There were thus a total of 14.

Roman forces in Spain

Complete information is not available concerning the exact numbers and the composition of the forces that Rome sent into Spain during the conquest. Some 30,000 men, not including the native auxiliaries, took part in the battles of Viriathus and Numantia. With Augustus' reign, the record is more precise. In the Cantabrian War about

40,000 legionnaires were used; when joined by auxiliaries their number reached a total of about 70,000. With the wars over, the legions were slowly withdrawn, so that around the middle of the 1st century AD only one legion remained. It was later replaced by another one whose permanent base (garrison) was León.

The peninsula supplied the Roman Army with great numbers of auxiliaries in the first century of the empire. Somewhat more than 100 units, alae and cohorts, were composed of men recruited from the least Romanized regions of Spain—Lusitania, the far north, and the northern *meseta*. The majority of them saw action on the Rhine and Danube fronts, although they also fought in Britain, North Africa, and the East.

The decisive factor in the Romanization of Spain was the extension of Roman citizenship. At first colonists of Roman or Italian origin living in Roman colonies in Spain were considered citizens. These colonies were established chronologically as follows: seven were created before Caesar, six during his reign, and three between Caesar and Augustus. Fourteen colonies were then founded by Augustus himself, and four between Augustus and the Flavians. Geographically, the colonies were distributed in the following manner: 12 in Baetica, 16 in Tarraconensis, and five in Lusitania; that is, a total of some 33 colonial seats of government, including some that did not receive the official designation of "colony." It must be noted that the colonization did not follow a studied plan but was determined instead to a great extent by the availability of suitable expropriated territory. In the area where the most conflicts developed (the civil war between Pompey and Caesar, in particular), more punitive action was taken against certain cities, and, consequently, there was in these places more *ager publicus* ("public land") to be distributed among the Roman legionnaires. The total number of Roman colonists may have reached 60,000. But since transient natives of the peninsula could in various ways also obtain Roman citizenship, the total increased by a considerable number. At the beginning of the empire, there were already 24 municipalities of Roman citizens and 48 cities under the *jus Latii* (Latin right). When these figures are added to those for the 33 colonies, there is a total of 142 cities under Roman law, with a population that may have reached 250,000 to 300,000.

Cities

Beyond this there still remain, however, nearly 300 cities and great numbers of non-Romanized people who are not accounted for. The complete Romanization of the Hispanic territory was not achieved until the Edict of Vespasian (74), which extended rights of the empire to all those Spaniards who did not previously have them. The rest of the empire had to wait until the Edict of Caracalla (212), which made Roman citizens of all persons in the empire.

These political advances were accompanied by great steps forward in the process of urbanization. In Tarraconensis, the western borders of which were very little urbanized at the time of Augustus, there were 293 settlements, 179 of which had a capital and 114 of which were rural. A century and a half later, the number of cities had risen to 248, an increase of 69; and the number of rural localities had fallen to 27, a decrease of 87.

Society and culture. Native languages (Iberian and Celtic), and the Greek and Phoenician that were still spoken in what had been their colonies, had virtually disappeared by the end of the 1st century BC, at least in Baetica and in the important cities and Romanized provinces. Languages spoken in the more remote centres of population were much longer in disappearing, and Basque survives even into the 1970s as an example of one of the pre-Roman tongues of the peninsula. It can be said, however, that by the time of Caesar, Latin was the universal language for the cultivated citizens of the peninsula, and there soon developed a native Hispano-Roman literature.

Religion in the Roman period

With regard to religion, in addition to the gods and minor deities of the Roman pantheon, who were quickly identified with the gods of the native Hispanic peoples, there were also adherents of many exotic religions originating in Persia, Asia Minor, Egypt, and North Africa. The most widely practiced religion appears to have been the worship of Isis, followed by the worship of the Magna Mater. Several of these beliefs were brought into Spain before the Roman conquest (Melqart, Tanit-Caelestis, Magna Mater). The majority of the remaining ones were introduced during the empire.

There is no evidence as to how Christianity was first introduced into Spain. It is quite possible that St. Paul visited Spain at some time between 63 and 67, but it is difficult to distinguish historical fact from legend concerning the "seven apostles." It is certain, however, that in the 2nd and 3rd centuries there were Christian communities throughout the peninsula. The most reliable documents on martyrs speak of persecutions of Christians in various locations under Prudentius. Shortly before (evidently), the Edict of Milan (313) called together the Council of Iliberri (Granada), which was attended by 19 bishops and 24 presbyters. The Nicene Council (325) was directed by Osio, the bishop of Corduba. The emperor Theodosius I (379–395), a Spaniard, made Christianity the only tolerated religion in the empire. The ascetic Priscillian spread the Manichean heresy throughout the western part of the peninsula during the second half of the 4th century, and the heresy persisted into the 6th century. Besides Priscillian, other Christian writers of note were Osio, Juvencus, Prudentius, and Orosius.

Roman remains

Archaeology has brought to light numerous monuments from this era—basilicas, amphitheatres, aqueducts, baths.

Monuments to Roman construction remain in many parts of the peninsula. Roman walls are to be found in Barcelona, Tarragona, Zaragoza, Idanha, Evora, and elsewhere. Bridges stand at Alcántara, Mérida, Piedras Albas, Alconétar, Chaves, Villa Formosa, Bibey, Salamanca, and Villa del Río, and aqueducts in Segovia, Mérida, and Tarragon. The entire peninsula is crossed by Roman roads, the principal ones going from Cádiz (the Via Augusta) along the Río Guadalquiver and continuing along the Levantina coast, from Cartagena north through Valencia and Tarragona to Narbonensis. Another road follows the Ebro from Tarragona to Galicia. Roman lighthouses are found at La Coruña and Caepionis Turris (Chipiona). Sites of military encampments are at León, Numantia, and elsewhere. Castles were built at Lousa and Benavente. Carmona has a necropolis, and there are mausoleums remaining at Tarragon, Fabara, Sádaba (two), Zalamea la Real, and Jumilla. Roman palaces and villas, many with mosaics, are found at Clunia, Milreu, Italica, and many other sites.

Roman art is also abundant throughout the peninsula. Principal centres for sculpture were Barcelona, Tarragona, Mérida, Italica, and Córdoba, while the centres for portrait work were at Italica, Córdoba, Cádiz, Tarragona, and Barcelona. There remain about 40 pagan sculptured sarcophagi, originating primarily in the south, the coast of the Levantina, and the central region. Mosaic work is common in all of the peninsula, with the exception of the northwest and the north.

Spaniards of great importance in the culture of Rome were writers such as the Senecas, Lucan, Mela, Martial, Quintilian, Columella, Canius Ruffus, Decianus, and Moderatus. Trajan, Hadrian, and Theodosius were Spanish emperors, and the Balbos and Lucius Licinius Sura were statesmen of note.

VISIGOTHIC SPAIN TO C. 500

There were many precedents to the great invasions by Germanic tribes of the beginning of the 5th century. Incursions by the Cimbrians and Teutons, the Franks, the Alemanni (Alamanni), and the Ostrogoths were signs of the growing threat to the empire. On the other hand, the peaceful infiltration and coexistence of the barbarians with the Romans along the interior border zones brought about the "barbarization" of the empire on every level, even in its high command. In 401 the Visigoths, under Alaric, crossed the Alps, descended on Italy, and in 410 sacked Rome. Other tribes, made up of Suebi, Alans, and Vandals, had crossed the Pyrenees the previous year (409). After two years of turmoil and anarchy, they at

last ceased their movement, signing a treaty of alliance (*foedus*) with Rome that conceded them military settlements within the empire. The Suebi and the Asdingi Vandals settled in Galicia (Gallaecia), the Silingi Vandals in Baetica (Vandalusia = Andalusia), and the Alans in Lusitania and Carthaginiensis. Tarraconensis was occupied only after Ataulphus had been made king of the Visigoths and had invaded the Provincia, where he died the same year. The occupation of the peninsula by Germanic tribes was then complete. Wallia (415–419), through an agreement with Honorius, founded the kingdom of Tolosa in federation with Rome. With this agreement made, Wallia immediately attacked the Vandals and Alans of Andalusia and crushed them. The Vandals who survived joined with their Galician brothers and conquered Cartagena and Seville and, after sailing from Tarifa, crossed the Strait of Gibraltar under the command of Gaiseric (Genseric) and landed in North Africa (429). The Suebi, now in exclusive control of Galicia, established themselves there as an independent kingdom. They later divided the entire western region of the peninsula among their elite troops and extended their effective rule as far as the Tagus River. The kingdom of Galicia, as such, disappeared in 585.

Wallia

With Euric (466–484) the kingdom of Tolosa (Toulouse) became unified and independent from Rome. This kingdom's powers extended through the Galliae (except for the northern zone) and through the Hispaniae (except for Galicia, which continued, along with its Lusitanian possessions, to be a kingdom dependent on the Suebi). Euric was at that time the most powerful ruler in the West. He issued the first laws, which, because of the continuing predominance of Latin culture, were written in Latin, drawn up by Roman jurists, and were influenced heavily by Roman law (475). With the Alans and the Silingi Vandals having been defeated by Wallia and with the Asdingi Vandals in Africa, the only powers in the peninsula were the Suebi in the northwest and the Visigoths. Demographically and ethnically, they stood for very little, numbering only about 250,000, as compared with the 7,000,000 Hispano-Romans of the peninsula. But power was almost exclusively in their hands. They lived in relative harmony with the Hispanic people, a situation made possible largely by two great leaders, Orosius and Idatius, who strongly favoured the idea of a unified Spain under Visigothic protection.

When the Suebi and the Visigoths had arrived in Spain, they were to a great extent already Romanized; thus, there was no drastic cultural difference between the invaders and the invaded. Furthermore, for many years the Visigoths were considered allies and collaborators with the empire and, to a certain extent, its protectors. With the pact of 418, the Visigoths agreed to recognize the supreme authority of the Emperor and to submit to Roman political and administrative organization. The Visigoths arrived in Spain already bearing a certain political and social organization, although it is true that many aspects of it underwent important changes. They brought with them an elective monarchy that was assisted by a popular assembly made up of all free men who were eligible to go to war. The population of Visigothic Spain, however, was composed of two completely distinct societies. On one hand, there was the Hispano-Roman populace with its own laws, its Latin culture, and Catholic Christianity; and on the other hand, there were the Arian Visigothic invaders, who, even though they were a minority, held military power. They were ruled by a sovereign who, in theory, had authority only over his own people and not over the Hispano-Romans, since there was a code of Roman laws for these native people and a tradition of Visigothic common law for the invading people. Two-thirds of the land was redistributed into Visigothic hands, although vast territory remained unoccupied because of the scant numbers of the new colonists. A ban against marriage between Visigoths and Hispano-Romans lasted until its abolition by Leovigild (568–586). The Visigothic capital, which was originally in Tolosa, was transferred later to Toledo (Toletum) with the reign of Leovigild.

(A.G.y B.)

II. Christian Spain, c. 500–1479

THE VISIGOTHIC KINGDOM

Given their distinctive laws and customs and their profession of Arian Christianity, the Visigothic absorption by the Hispano-Roman population of perhaps 6,000,000 was not easy. Visigothic control of the peninsula remained incomplete so long as the Suebi (Suevi) maintained an independent kingdom in Galicia (Gallaecia) and the Basques refused submission to any external authority. Moreover, the Hispano-Romans in the southeast welcomed there the restoration of Byzantine rule early in the 6th century. Leovigild (568–586), the most capable of the Visigothic kings, who borrowed the Roman symbols and trappings of monarchy and ruled from a central position at Toledo (Toletum), made significant strides toward peninsular unification by conquering the Suebi and subjugating the Basques. Leovigild's attempt to convert the Hispano-Romans to Arianism, however, provoked a revolt in the southern province of Baetica under the leadership of his son, Hermenegild, who had accepted the Catholic religion. Rebel expectations of Byzantine help were unfulfilled, and so the uprising was finally crushed.

The evident futility of Leovigild's religious policy prompted his son, Recared (586–601), to convert to the Catholic faith professed by the majority of the people. His conversion, together with that of the Gothic nobles and bishops, removed one of the chief barriers to the assimilation of the two peoples. Thereafter, the Hispano-Romans, with less reason to be beguiled by Byzantium, developed a new and strong loyalty to the Visigothic monarchy. This facilitated the conquest of the last Byzantine strongholds in the peninsula during the reign of Swinthila (621–631). From then on, the Visigoths ruled over all of Spain.

Conversion of Recared to orthodoxy

The conversion also symbolized the triumph of Hispano-Roman civilization over the barbarians, and it helped to forge an alliance between the crown and the church. Realizing that the bishops were the leaders and spokesmen for the majority, the kings assumed the right to appoint them and to summon them to the Councils of Toledo. Those assemblies, though primarily ecclesiastical, had an extraordinary influence on the political life of the kingdom. After listening to the royal message describing problems of greatest concern, the prelates enacted appropriate canonical legislation as well as decrees relating to secular business, such as the election of the king or cases of treason. The bishops, in their councils, became a principal support of the monarchy, but in their quest for an ordered political regime, they lost some of their independence.

The monarchy was elective, partly because of aristocratic opposition to hereditary succession and partly because of a lack of natural heirs. Knowing the chronic Gothic tendency to assassinate the king, the bishops introduced the ceremony of anointing to give him a sacred character and to make known to all that he was under God's protection. In order to curtail the disorders attendant upon a royal election, they also elaborated an electoral procedure. The king ruled with the aid of a household (*officium palatinum*) based upon Roman models, and when important business required it, he took counsel with assemblies of magnates and other prominent persons (*aula regia*). Territorial administration was entrusted to dukes, counts, or judges. They were also responsible for the towns where self-government had long since decayed. Provinces and other administrative districts were essentially survivals from the Roman era. The economy was mainly agricultural, and there is scarcely any evidence of commercial or industrial activity.

The influence of Roman civilization can also be noted in the steady subordination of Germanic law to the law of the Hispano-Roman majority. The *Liber Judiciorum*, a code of law promulgated around 654 by Recceswinth (649–672), in both its structure and its substance gives testimony to Roman influence. Intended to have territorial rather than personal application (as was the case with barbarian customary law), it reflected a persistent accep-

The *Liber Judiciorum*

tance of Roman legal principles. Yet there are Germanic elements in the code, such as the ordeal of cold water. The *Liber Judiciorum* was one of the chief legacies of the Visigothic era and a tribute to the Hispano-Romans who were responsible for creating it.

The persistence of Roman tradition is also evident in the remarkable outburst of cultural activity in the 7th century. St. Isidore, bishop of Seville (Hispalis) from *c.* 600 to 636, friend and counsellor of kings, was a voluminous author who left his mark upon the entire medieval era. His works include a history of the Visigoths, a monastic rule, theological treatises, and, above all, the *Etymologies*, an encyclopaedic work that attempted to summarize the wisdom of the ancient world.

At the close of the 7th century, the Visigothic monarchy entered upon a critical time. The deposition, through trickery, of King Wamba (672–680), a competent ruler who tried to reform the military organization, indicated the trend of events. Numerous plots threatened his successors, who sought to blame the unrest upon the Jews, whom they tried to convert to Christianity or to reduce to slavery. Factionalism reached a climax when Witiza (700–710) died, hoping to leave the throne to his son; but Roderick, duke of Baetica (710–711), was acclaimed instead. Witiza's family apparently appealed to the Muslims in North Africa to help them overthrow Roderick. In response, Tāriq ibn Ziyād, the Muslim governor of Tangier, landed at Gibraltar (Calpe) in 711 and defeated King Roderick and the Visigoths near Río Guadalete on July 19. The conquest of the rest of Spain, left leaderless by the death of the king, followed quickly. The Visigothic kingdom perished, but its memory survived to inspire the kings of Asturias-Leon to attempt the reconquest of Spain.

CHRISTIAN SPAIN FROM THE MUSLIM INVASION TO C. 1260

The question of peninsular unity is central to the history of Christian Spain from the Muslim invasion in the 8th century to the advent of Ferdinand and Isabella in the late 15th. The Muslims destroyed what unity the Visigoths had been able to impose and introduced religious, cultural, legal, linguistic, and racial differences that impeded their assimilation by the natives. In the ensuing centuries, several Christian states emerged in the northern mountains; partly as a matter of self-preservation, partly for religious hostility, and partly to vindicate the Visigoths, they embarked upon the Reconquista (Reconquest). Their progress depended upon the strength of Muslim Spain at any given moment. In times of Muslim weakness or disintegration, the Christians usually made important gains. The kings of Asturias-Leon-Castile, eventually the largest state, aspired to dominance over the whole peninsula, but their ambitions were challenged and often thwarted by the rulers of Portugal, Navarre, and Aragon-Catalonia, which began to assume definitive form in the 11th and 12th centuries. The Reconquista was largely completed by the middle of the 13th century, with only the small Muslim kingdom of Granada (Gharnāṭah) surviving as tributary to Castile until 1492.

In the later Middle Ages, the problems of peninsular unity took on a new aspect as the kings of Castile tried to gain control of the other Christian states by marriage, diplomacy, and war. The Trastámara dynasty, which came to power in Castile in the late 14th century, consistently pursued this aim. It was achieved in part by Ferdinand and Isabella, whose marriage united Aragon and Castile. They also conquered Granada, thereby terminating the Reconquista, but their attempts to bring Portugal into a family union by marriage failed. In any case, the diversity of languages, laws, and traditions that had developed in the medieval centuries was not erased simply by the political union of kingdoms.

The Christian states, 711–1035. Christian resistance to the Muslim conquest began in the mountains of Asturias, where Visigothic nobles took refuge and joined the native mountaineers in opposing the invaders. Their leader was Pelayo (718–737), a Visigothic lord whose victory over the Muslims at Covadonga, around 722, later was hailed as the beginning of the Reconquista and the "'salvation of Spain." Under King Alfonso I (739–757), the Asturians occupied Galicia, abandoned by rebelling Berber troops who had been garrisoned there. Farther south, he laid waste the Duero River Valley, which remained for many years a no-man's-land separating Christian and Muslim Spain. In the western Pyrenees, the Basques seem to have regained their independence; and the Franks, after expelling the Muslims from Septimania, began to penetrate into northeastern Spain. Charlemagne's attempt to take Saragossa (Saraqusṭah) in 778 failed, but his forces eventually moved into Catalonia, seizing Barcelona in 801. The Franks organized this region into several counties, later known collectively as the Spanish March. For several centuries, the March was linked by political and cultural ties to the Carolingian Empire and then to the Kingdom of France.

"Salvation of Spain"

While the Catalans tended to be oriented to the north, the Asturians looked southward and under Alfonso II (791–842) cautiously moved their capital to Oviedo, where they attempted to restore the institutions of the Visigothic monarchy.

A breakdown of effective government in Muslim Spain in the late 9th century enabled Alfonso III (866–911) to cross enemy frontiers at will in search of plunder and to capture important places, notably Porto. Alfonso also began the process of colonizing the lands stretching southward to the Duero that had been depopulated for so long. On his eastern border, he erected many castles as protection against Muslim attacks, thus giving the area its name, Castile, and its distinctive character. The first extant Christian chronicles of the Reconquista were compiled during this period and sought to establish an essential continuity between the Visigothic and Asturian kingdoms. The Asturians saw themselves as the legitimate heirs of Visigothic power and tradition and consciously assumed responsibility for the reconquest of Muslim Spain. Future challenges to Asturian leadership were beginning to take shape, however, in the northeast, where a strong Basque kingdom, centred at Pamplona in Navarre, appeared under King Sancho I (905–926). In Catalonia, Count Wifred of Barcelona (873–898) took advantage of the disintegration of the Carolingian Empire to assert his independence and to secure control of several other counties. The dynasty he founded ruled in Catalonia until the 15th century.

Alfonso III

Encouraged by the expansion of the Asturian kingdom and evident Muslim weakness, García I (911–914) moved his capital southward from Oviedo to the city of León. The optimism revealed by this action was shattered during the course of the 10th century as order and unity were restored in Muslim Spain under the rule of the caliphs of Córdoba (Qurṭabah). Muslim incursions into the Christian north were resumed and caused great destruction, though from time to time the Christians achieved some successes. Ramiro II (931–951) defeated the caliph ʿAbd ar-Raḥmān III at Simancas in 939, a truly remarkable feat. Yet within his own kingdom, Ramiro was faced with the growing defiance of the Castilians, who, as frontiersmen exposed to the hazards of daily conflict with the enemy, seem to have had little use for Leonese tradition and law. Their count, Fernán González (*c.* 930–970), successfully resisted the kings of Leon and laid the basis for Castilian independence.

As the century progressed, the Christian position steadily deteriorated. Muslim ascendancy was demonstrated clearly by the procession of envoys representing Ramiro III of Leon (966–984), Sancho II of Navarre (970–994), Count Borrell of Barcelona (*c.* 940–992), and Count García Fernandez of Castile (970–995), who came to Córdoba to offer homage and tribute to the caliph. Even in these depressed circumstances, the Leonese seem to have maintained the traditional Asturian claim to the Visigothic inheritance. As a more explicit assertion of peninsular hegemony, they elaborated a concept of a Hispanic empire centred at León. The theory no doubt served to bolster morale at the close of the century when all the Christian states suffered from the scourge of al-Manṣūr, who assumed dictatorial power in the caliphate. In order to distract the Muslims from his usurpation, he

The scourge of al-Manṣūr

led armies on annual invasions of the north. In 985 he defeated Count Borrell and burned Barcelona; three years later he plundered León and many other towns.

The breakup of the caliphate following al-Manṣūr's death freed the Christian states from the threat of destruction. Count Ramón Borrell of Barcelona (992–1018) even gained a measure of retribution by intervening in the Muslim civil wars and sacking Córdoba in 1010. Alfonso V of Leon (999–1028) used the respite from war to repair his kingdom and in 1017, at a council held at León, promulgated the first general laws for the government of the realm. The cessation of Muslim attacks also encouraged a renewal of rivalry among the Christian states. For a few years, Sancho III the Great, king of Navarre (1000–35), succeeded in lording it over the rest of Christian Spain. During his reign, contacts with northern Christendom became more frequent. The pilgrimage route to Compostela, a principal avenue of French influence, was developed; the Cluniac reform of the monastic life was introduced, and feudal ideas began to affect the structure of aristocratic society. In pursuit of his own aggrandizement, Sancho III, who already held the counties of Aragon, Sobrarbe, and Ribagorza and numbered Count Berenguer Ramón I of Barcelona (1018–35) among his vassals, seized the county of Castile and threatened Vermudo III of Leon (1028–37). In 1034, Sancho occupied the city of León and assumed the title of emperor, but his triumph was cut short by his death in the following year.

The medieval empire, 1035–1157. Although Sancho III brought under his sway all the Christian states with the exception of Catalonia, he rejected the Leonese tradition of a unified, indivisible kingdom and preferred to regard his dominions as a family patrimony to be shared among his sons. Thus, he gave the kingdom of Navarre to García III (1035–54); Ferdinand I (1035–65) received Castile; and Ramiro I (1035–63) obtained Aragon, to which he added Sobrarbe and Ribagorza in 1045 after the murder of a fourth brother, Gonzalo. As each of the sons bore the royal title, Castile and Aragon were elevated to the status of kingdoms. Although Vermudo III had regained possession of Leon, he was defeated and killed in 1037 by Ferdinand I, who annexed his kingdom and took the imperial title. For the next 30 years, he aspired to domination throughout Spain, overcoming his brothers on the battlefield, conquering Coimbra, and compelling the petty Muslim rulers (*reyes de taifas*) of Toledo (Ṭulayṭulah), Seville (Ishbīliyah), and Badajoz (Baṭalyaws) to pay him tribute.

Ferdinand's contemporary Count Ramón Berenguer I of Barcelona (1035–76) was actively developing Catalan interests and relationships among the lords of Languedoc, in southern France. He also promulgated the earliest legal texts included in the compilation of Catalan law known as the *Usatges de Barcelona* ("Usages of Barcelona").

At his death, Ferdinand I followed his father's custom and divided his kingdoms between his sons, giving Castile to Sancho II (1065–72) and Leon to Alfonso VI (1065–1109). Discord promptly arose and ended only with Sancho's murder in 1072. Alfonso VI thereupon reunited the two kingdoms, but he had to swear to the Castilian nobility that he had had no hand in his brother's death. Chief among the Castilians, who now became his vassals, was Rodrigo Díaz de Vivar, known to history as El Cid Campeador. Court jealousies eventually forced him into exile, and for a time he served the Muslim king of Saragossa before becoming the protector of the king of Valencia.

Alfonso VI

The division of Muslim Spain into several small kingdoms enabled Alfonso VI to levy tribute upon them, but he decided to conquer them and in 1085 received the surrender of Toledo, the fall of which had great symbolic importance because it had been the ancient seat of the Visigothic monarchy, and its possession strengthened the king's claims to hegemony in Spain. For this reason, he called himself "Emperor of Toledo," but he also used the title Emperor of the Two Religions to emphasize his ascendancy over both Christians and Muslims. The conquest brought his frontiers to the Tagus and resulted in the incorporation into his kingdom of large numbers of Muslims and Jews, who in times past usually had withdrawn to the south in the face of advancing Christian armies. Large numbers of Mozarabs, or Arabic-speaking Christians, also lived in Toledo and its environs. The interaction among these differing religious and cultural traditions became especially intense in the ensuing years.

The fall of Toledo alarmed the other petty kings in Muslim Spain, who invited the Almoravids, a fanatical Muslim sect in North Africa, to come to their aid. The Almoravids crushed the Christian army at Zalacca (az-Zallāqah) in 1086, but they also swallowed up the petty kingdoms. The reunification of Muslim Spain kept Alfonso VI on the defensive for the rest of his reign and made further progress in the Reconquista temporarily impossible. The Cid successfully defended Valencia against the Almoravids, but after his death in 1099 his followers had to abandon the city. The Almoravids were then able to occupy the whole of eastern Spain as far north as Saragossa.

The deeper Christian involvement in the world of Islām brought about by these events was counterbalanced by the steady growth of northern European influences, especially that of the papacy. Pope Gregory VII (1073–85), who spearheaded the general reform of the church, imposed uniformity of liturgical practice by insisting upon the abandonment of the ancient Mozarabic liturgy native to Spain and its replacement by the Roman. His assertion of papal sovereignty over Spain was not admitted by the Spanish rulers, and he did not press it. In addition, French monks and clerics discovered ecclesiastical fortunes in Spain, and French knights came to participate in greater numbers in the wars of the Reconquista. Two of them, the cousins Raymond and Henry of Burgundy, married the daughters of Alfonso VI, Urraca and Teresa, and so became the progenitors of dynasties that ruled in Leon and Portugal until the late 14th century.

After Alfonso VI's death, Urraca (1109–26), now widowed, succeeded him. She married Alfonso I of Aragon, "the Battler" (1104–34), but disharmony between them kept the kingdom in constant turmoil. Alfonso I eventually retired in disgust to his own kingdom. Urraca's son by Raymond of Burgundy, Alfonso VII (1126–57), restored the prestige of the Leonese monarchy and affirmed Leonese aspirations to predominance in Spain by being solemnly crowned as emperor in the cathedral of León in 1135. This was apparently the first and last imperial coronation in Spain. Within a few years, however, two formidable challenges to Leonese ascendancy appeared in the newly formed federation of Aragon and Catalonia and in the newly independent kingdom of Portugal.

Alfonso I of Aragon, after ending his marriage to Urraca, conquered Saragossa in 1118, thus advancing his frontiers to the Ebro. Shortly after, he made an extraordinary incursion into the heart of Muslim Spain to liberate the Mozarabs of Granada (Gharnāṭah), whom he settled in Aragon. Thereafter, the Mozarabic population in Muslim Spain seems to have been negligible. When he died, he bequeathed his realms to the military Orders of the Hospital, the Temple, and the Holy Sepulchre, but his people refused to accept this. The Navarrese, who had been subject to Aragonese rule since 1076, elected their own king, García IV (1134–50), and the Aragonese summoned the dead king's brother, Ramiro II (1134–37), from a monastery to become their king. He assumed the responsibility only long enough to marry and to father a child, Petronila, who could inherit the kingdom. She was betrothed in 1137 to Count Ramón Berenguer IV of Barcelona (1131–62), who was entrusted with the administration of the kingdom. The child by this marriage, Alfonso II (1162–96), united in his person the kingdom of Aragon and the county of Barcelona, a union of peoples with different traditions, interests, and languages, which survived innumerable vicissitudes throughout the medieval era. Catalonia developed as a maritime power, interested in expansion in the Mediterranean, while Aragon, a landlocked region, remained largely agricultural, dominated by a landed aristocracy. Each region retained

its distinctive laws and institutions, and there was strong opposition on both sides to any attempt at assimilation. The term crown of Aragon appropriately describes the federation of the two territories.

The crown of Aragon

In the west, the county of Portugal, a part of the Kingdom of Leon, which Alfonso VI had given to Teresa and Henry of Burgundy, began to emerge as an independent state. Their son, Afonso I Henriques (1128–85), rejected Leonese overlordship and assumed the title of king, around 1139. In order to protect himself, he became a vassal of the papacy, pledging to pay an annual tribute; but the Pope did not explicitly recognize him as a king until 1179.

The Almoravid Empire, meanwhile, began to break up as a result of internal discord and the rise of the Almohads, another North African Muslim sect. Taking advantage of the Muslim civil wars, the Christian rulers penetrated deeply into Muslim Spain and made some significant conquests. In 1147, Afonso I, with the help of a crusading fleet from northern Europe, conquered Lisbon, while Alfonso VII and Ramón Berenguer IV, aided by Pisan ships, captured the great seaport of Almería on the southeastern coast. In the following year, the count of Barcelona seized Tortosa (Ṭurṭūshah) and Lérida (Lāridah), thereby completing the expansion of Catalonia and extending its frontiers to the mouth of the Ebro. The Almohads, however, after destroying Almoravid power, invaded Spain and reduced all the Muslim areas to their rule. In 1157 they reconquered Almería (al-Marīyah) and thereafter effectively checked any further Christian progress.

The rise of Castile and Aragon. Alfonso VII died in the same year. By dividing his realms between his sons, Sancho III (1157–58), to whom he gave Castile, and Ferdinand II (1157–88), who received Leon, he shattered the unity he had tried to impose upon the peninsula and ruined the concept of a Leonese empire. Although the Almohads kept the Christians on the defensive, Alfonso VIII of Castile (1158–1214) and Alfonso II of Aragon were confident enough of ultimate success to conclude a treaty in 1179 partitioning Muslim Spain. The territory of Valencia (Balansiyah) was reserved for Aragonese reconquest, while the rest of Andalusia fell to Castile. Rivalries among the Christian princes quickly arose, however, when Alfonso VIII tried to dominate his colleagues. Common action against the Almohads was impeded, and in 1195 they inflicted a humiliating defeat upon the king of Castile at Alarcos (al-Arak), south of Toledo. Realizing the seriousness of the Muslim threat, the other Christian kings made peace with Castile. In 1212, Alfonso VIII, supported by Sancho VII of Navarre (1194–1234) and Peter II of Aragon (1196–1213) and by contingents from Portugal and Leon, achieved a remarkable victory over the Almohads at Las Navas de Tolosa (al-'Iqāb). His triumph opened the road to Christian penetration into Andalusia, and it sounded the death knell for the Almohad Empire.

Active participation in the Reconquista was only one aspect of Aragonese policy. The king of Aragon, as ruler of Catalonia, also had significant interests in southern France, where several lords were his vassals. The Albigensian heresy was spreading through the region, however, and Peter II, while no friend of heretics, concluded that his rights there were jeopardized by the crusade launched by Pope Innocent III against them. When Peter tried to rescue his brother-in-law, the count of Toulouse, he was defeated and killed by the crusaders at Muret in 1213. His death marked the beginning of the end of Catalan ambition and power in southern France.

Advance against the Almohads

The steady collapse of the Almohad Empire in the second quarter of the century enabled the Christian rulers to carry the Reconquista almost to completion. James I of Aragon (1213–76), after a turbulent minority, conquered Majorca (Mayūrqah) in 1229. This was the first important step in the overseas expansion of Catalonia and the first great manifestation of Catalan naval power. He then began the more arduous task of conquering the kingdom of Valencia, but he was distracted briefly by the possibility of annexing Navarre. When Sancho VII died without children, the Navarrese proclaimed his nephew, Count Theobald of Champagne (1234–53), as their king, and thereafter Navarre was drawn closer to French politics. Putting aside his ambitions in Navarre, James I received the surrender of the city of Valencia in 1238 and promulgated a code of law for his new realm, which formed the third constituent element in the crown of Aragon. For the first time, thousands of Muslims were incorporated into his dominions.

In the meantime, Alfonso IX of Leon (1188–1230) was advancing southward to the Guadiana (Wādī Ānā), taking Mérida (Māridah) and Badajoz in 1230 and opening the road to Seville. Upon his death, his son Ferdinand III, already king of Castile (1217–52), by virtue of inheritance from his mother, Berenguela, daughter of Alfonso VIII, was received as king of Leon, thus reuniting the two realms for good. With their combined resources, he conquered Córdoba in 1236, Murcia (Mursiyah) in 1243, Jaén (Jayyān) in 1246, and Seville in 1248. Only the Kingdom of Granada remained in Muslim hands, but it was subject to the payment of annual tribute to Castile. For this reason and because it did not constitute any evident threat, serious attempts to conquer it were seldom undertaken in the next two centuries.

Society, economy, and culture. In the first 300 years following the Muslim invasion, Christian society and institutions developed slowly, but substantial change came rapidly in the 12th and 13th centuries. Among the reasons for this were a general rise in population, increasing contacts with northern Europe, participation in the common European growth of trade and town life, and progress in the reconquest.

By the middle of the 13th century, the kingdoms of Leon-Castile, Aragon-Catalonia, Navarre, and Portugal had attained the frontiers they were to have, with only slight modification, until the close of the medieval era. The crown of Aragon was unique among them because it was a confederation of several distinct entities: the kingdoms of Aragon, Valencia, and Majorca, and the principality of Catalonia.

The principle of hereditary succession was established early, but the idea of election survived in the acclamation of the king when he was anointed and crowned in accordance with Visigothic tradition. Peter II of Aragon was crowned in Rome by the Pope, whose vassal he became, thereby subordinating his kingdom to papal suzerainty. The king ruled with the collaboration of his household, whose chief officers were the chancellor, ordinarily a prelate or cleric responsible for the dispatch of royal letters and the keeping of records, and the *mayordomo* and *alférez* (*majordom* and *senyaler* in Catalonia), magnates whose tasks were, respectively, the management of the household and royal estates and the command and organization of the army. Magnates also served as administrators of provinces and in Castile were known as *merinos* or, later, as *adelantados*. In Catalonia, the several counts originally governed their counties in the name of the Carolingian emperors, but they became in fact independent, and by the 11th century, the counts of Barcelona established an effective sovereignty over the principality. Vicars (*vegueros*), responsible for justice, and bailiffs (*batlles*), for taxes, administered Catalan territorial subdivisions. Bishops, magnates, monasteries, and military orders often enjoyed the privilege of immunity, whereby royal officials were forbidden to enter their estates to administer justice or to collect taxes, except in case of negligence. Among the most important immunities were those of the archbishop of Compostela in Galicia and those of the military orders in the region south of Toledo.

As the old Roman concept of the state had been obscured, private and personal relationships of a feudal character had considerable influence upon the governmental and military structure. The development of feudalism was greatest in Catalonia due to its contacts with France, but in the western states it was retarded and incomplete. The Catalan nobles were vassals of the count of Barcelona, to whom they owed military and court service in return for fiefs, and they often had vassals of their own. In the west, royal vassals had the same obliga-

tions, but ordinarily they did not receive fiefs nor did they give fiefs to their vassals. The magnates, known as *ricos-hombres* (*i.e.*, rich or powerful men) in the west and as *barones* in Catalonia, usually were vassals of the king or count and served as his principal counsellors and as the backbone of the military organization. Nobles of secondary rank, called variously *infanzones, fijosdalgo, caballeros* or *cavallers*, usually were vassals of the magnates.

Inasmuch as the economy of the Christian states was mainly agricultural, the king and nobles of both ranks were landlords who derived the greater part of their income from the exploitation of landed estates. Peasants living on the estates cultivated the soil and paid a variety of rents and services to their lords. The burden fell heaviest upon the serfs (*solariegos* in Castile, *payeses de remensa* in Catalonia), who were effectively bound to the soil and denied freedom of movement. Catalan serfs, especially, were subject to abuse by their lords by means of the so-called evil usages. Leonese peasants occupying lands known as *behetrías* enjoyed the right of freely choosing their lord and of changing their allegiance as they wished, but this right was being limited in the 13th century. Peasants who were willing to accept the risks and the adventure of living on the frontier were the freest of all, and like pioneers in all ages, they developed a strong sense of personal worth and independence.

The serfs

As the Reconquista progressed, the crown encouraged the colonization of the Duero Valley by erecting fortified urban centres (*concejos*) with extensive rural areas dependent upon them. The people who settled there received charters (*fueros*) defining their liberties and giving them rights of self-government under their own elected magistrates (alcaldes). The economy of these towns was based not so much on industry and commerce as upon the raising of sheep and cattle and the booty taken by the urban militia in the wars of the Reconquista. Aragonese and Catalan towns enjoyed a lesser degree of autonomy, but the latter began to achieve a real importance as mercantile centres. The incorporation of such cities as Toledo, Saragossa, Lisbon, Córdoba, Valencia, and Seville greatly increased the urban population and contributed substantially to the development of trade and industry. Artisans appeared in greater numbers and began to organize guilds; a merchant class engaged in buying and selling on a large scale and commonly using money as a medium of exchange also emerged. The development of shipbuilding at such places as Santander and Barcelona made possible a native overseas-carrying trade.

The advance of the Reconquista also added large numbers of Muslims and Jews to the Christian kingdoms. The former, known as Mudéjares, were settled chiefly in rural areas, though there were always important Muslim quarters in the towns. The Jews lived mainly in the cities, where they were involved in trade and moneylending and were often employed as royal tax collectors. Both Muslims and Jews were permitted to practice their religions and to govern themselves according to Islāmic or Judaic law, but they were subject to the payment of regular tribute, and from time to time the Jews were assaulted by their Christian neighbours.

In view of the growing administrative, military, and economic importance of the towns, the crown summoned their representatives to meet in the royal council with prelates and nobles. In 1188 at León, Alfonso IX held the first such assembly (*curia plena*), but similar gatherings began to be held regularly in the other states in the early 13th century. Meetings of the Cortes, as these assemblies were called, served a number of purposes, one of which was to enable the king to obtain consent for the levy of extraordinary taxation to meet ever increasing financial needs caused by the expansion of royal activities and responsibilities. Parliamentary growth in the peninsular kingdoms was a manifestation of a common European phenomenon, but it is noteworthy that it occurred earlier in Spain than elsewhere.

Government of the towns

The Reconquista was also accompanied by the restoration of bishoprics or the expansion of older ones. Spain was divided ecclesiastically into five metropolitan sees: Toledo, which claimed the primacy, Tarragona, Braga, Compostela, and Seville. Papal intervention in their affairs became much more frequent in the 12th century. Monastic reforms emanating from Cluny and Citeaux in France profoundly affected religious life in the 11th and 12th centuries. Early in the 13th century, the mendicant orders, the Franciscans and Dominicans (whose founder was the Spaniard Domingo de Guzmán), spread through the peninsula. The military orders of the Temple and the Hospital, originating in the Holy Land, appeared in Spain in the 12th century, but in the second half of the century, several indigenous orders were founded—Calatrava, Alcántara, Santiago, and Avis. These knights tried to observe the monastic life, but they also provided much of the armed might responsible for the successes of the Reconquista.

The clergy were the guardians of the cultural tradition and the authors of the scant literary remains of the first centuries of the Reconquista. In the 8th century, the adoptionist controversy over the nature of the Trinity engendered many polemical works; the writings of the Mozarabs of Córdoba, Eulogius (died 859) and Alvarus (died 861), illustrate the difficulties faced by Christians living under Muslim rule. Under the influence of northern Europe, theology, philosophy, and Roman and canon law became formal subjects of study in the universities of Palencia and Salamanca, founded early in the 13th century by Alfonso VIII of Castile and Alfonso IX of Leon, respectively. The existence of the universities was problematical for some time, but eventually Salamanca attained international renown. In the meantime, a vernacular literature was developing, and around 1140 there appeared the first great epic in the Castilian tongue, *The Poem of the Cid*. The output of Spanish writers was still slim, however, though the historical works of Rodrigo Jiménez de Rada, archbishop of Toledo (died 1247), set the pattern for future generations of Spanish historians.

CHRISTIAN SPAIN, C. 1260–1479

Castile and Leon, 1252–1479. In the late medieval centuries, a power struggle developed between the kings of Castile, who sought to concentrate all authority in their hands, and the nobility, who tried to use the monarchy for their own aggrandizement. The struggle began during the reign of Alfonso X the Wise (1252–84), who is perhaps best known for the literary and scientific labours accomplished by the scholars who enjoyed his patronage. From a political standpoint, his reign ended in failure. Although he showed some interest in continuing the Reconquista and had to suppress a revolt by the Mudéjares of Andalusia and Murcia, he preferred to play a larger role upon the European scene. In 1257 he was elected Holy Roman emperor and spent most of his energy and resources in the next 17 years trying to overcome a rival candidate and to win papal recognition. His extraordinary need of money for this purpose eventually brought to a head a domestic crisis that had been developing since his first years in power.

Like his father, Ferdinand III, Alfonso perceived the need for a uniform royal law that would supersede all local and regional laws, such as the *fueros* of the towns and the Liber Judiciorum used in Leon (and translated into Castilian as the Fuero Juzgo on the orders of Ferdinand III). A code of law, known as the Fuero de las Leyes, intended to be common to the whole realm, was prepared; it was the earliest redaction of the Siete Partidas, the great code that received definitive form early in the 14th century. The new code drew heavily upon Roman law, but the King's attempt to impose it provoked outcries from the nobles and from the towns. In 1272 they forced him to confirm their traditional laws and customs and to abandon the attempt to create a single body of law for the kingdom. This disappointment was followed in 1274 by the Pope's final rejection of Alfonso X's claims to the empire. In 1275 Alfonso's eldest son and heir, Fernando de la Cerda, died, and the King recognized Fernando's son, Alfonso de la Cerda, as his successor. The King's second son, Sancho, refused to accept this decision, and, posing as the defender of the nobles against

Legal reform

the crown, took up arms against his father. An assembly of nobles, prelates, and townsmen deprived Alfonso X of his right to rule and entrusted the government to Sancho, but civil war raged until the King's death.

Sancho IV (1284–95) had to contend with domestic and foreign opposition in support of his nephew, and he was also threatened by invasion from Morocco. The latter danger initiated a long struggle to control the Strait of Gibraltar and thus to close that invasion route. The minority of Ferdinand IV (1295–1312) saw a renewal of efforts on behalf of Alfonso de la Cerda, but he was finally persuaded to renounce his claims to any share in the royal inheritance. In 1309 the King captured Gibraltar (Jabal Ṭāriq), but the Muslims recovered it a quarter century later. Disorders recurred in the minority of Alfonso XI (1312–50), but when he came of age he crushed his opponents among the nobility with great brutality. For 25 years he waged war against Granada. In 1340, with help from the other Christian states, he won a great victory over the combined forces of Granada and Morocco at Río Salado, thereby putting an end to any further Moroccan attempts to gain a foothold in Spain. Four years later he seized Algeciras (al-Jazīraha al-Khaḍrā'), but his effort to take the adjacent fortress of Gibraltar ended with his death as a victim, like so many other thousands, of the Black Death.

Conflict between the crown and the nobility flared again in the reign of Peter I the Cruel (1350–69), whose half brother, Henry of Trastámara, an illegitimate son of Alfonso XI, challenged his right to rule. To further his aims, Henry secured French support, and in 1366, with a mercenary army commanded by Bertrand du Guesclin, he was able to expel Peter from the kingdom. The King, however, persuaded Edward, the Black Prince, to assist him in regaining his throne. The combined Anglo-Castilian forces routed Henry of Trastámara at Nájera in 1367. But once the Black Prince withdrew, Henry and du Guesclin returned; they defeated and killed Peter at Montiel in 1369.

Henry II (1369–79), the first of the Trastámara dynasty, had to defend his title against his neighbours in the peninsula and to overcome domestic opposition. But he succeeded in consolidating his position and was even able to supply ships to support his French allies against the English. One consequence of the French alliance was Castilian recognition of the Avignonese pope during the Great Western Schism. The ambitions of the Trastámara family to unify the peninsula were revealed when Henry's son John I (1379–90) claimed Portugal by right of marriage. His intervention evoked the spirit of nationalism among the Portuguese, who defeated him at Aljubarrota in 1385. In the next year, John of Gaunt, duke of Lancaster, who had married Peter I's daughter, claimed the Castilian throne and with Portuguese help invaded Galicia. The campaign failed, however, and he made peace in 1388, agreeing to the marriage of his daughter Catharine to Henry III, oldest son of John I. Thus the enmity between the two branches of the Castilian royal house was laid to rest.

Avignon papacy

During the minority of Henry III (1390–1406), the nobility again tried to further their own fortunes at the expense of the crown, but when the king came of age he effectively repressed them. The long reign of his son, John II (1406–54), proved to be disastrous to the prestige and authority of the crown. Stability was maintained during his minority by his uncle, Fernando de Antequera, who acted as regent until his election as king of Aragon in 1412. Fernando's sons, Henry and John of Navarre, continued to participate actively in Castilian affairs and tried to dominate the King, but the royal favourite Alvaro de Luna had already gained ascendancy over the King and successfully resisted these intrigues. For most of the reign, he was the effective instrument of royal power, and only in 1453 was the King persuaded to be rid of him and to order his execution.

Rivalry among the nobles for influence and power was intensified during the reign of Henry IV (1454–74). Juan Pacheco, marquess of Villena, long dominated the King, but others succeeded in winning royal favour. Henry's alleged impotence caused the magnates to cast doubt upon the legitimacy of the infanta Joan, who, it was charged, was the child of the queen and of Beltrán de la Cueva, one of the King's most recent favourites. In derision, the little girl was called La Beltraneja. Henry IV disowned her at Los Toros de Guisando in 1468 and recognized his sister Isabella as heir to the throne. Villena and his followers evidently hoped to be able to control her, but they were mistaken. Although she had promised not to marry without her brother's consent, she did precisely that in 1469. Her husband was Ferdinand, son and heir of the king of Aragon. Angered by this, Henry IV tried to deprive Isabella of her rights, but when he died, she was proclaimed as Queen Isabella I (1474–1504). Afonso V of Portugal, who was betrothed to Joan (La Beltraneja), invaded Castile on her behalf, but in 1479 she renounced her claims. In the same year, Ferdinand ascended the Aragonese throne, thus effecting the personal union of Aragon and Castile.

The infanta Joan

Castilian institutions, society, and culture. The reemergence in the 13th century of the idea of the state as a public institution, partly as a result of the revival of Roman law and the rediscovery of Aristotle's *Politics*, had important consequences for the development of the Castilian monarchy. As the embodiment of the state, the king (whom the Siete Partidas described as God's vicar on earth) tended to concentrate power in his own hands. The jurists who staffed and dominated the royal bureaucracy tried to push royal claims to authority as far as they would go. The influence of Roman law, though originally opposed, grew steadily, and when Alfonso XI promulgated the Ordenamiento de Alcalá in 1348, the place of Roman law and legal procedure in the courts was assured. The royal tribunal (*audiencia*) was reorganized in the reign of Henry II, but it did not receive definitive form until the end of the Middle Ages. The tendency toward centralization also prompted the crown to intervene more frequently in the government of the towns, where rivalry among factions was becoming endemic. The urban aristocracy, originating as a class of non-noble mounted warriors, had gained control of the government and tended to exclude the lower orders from participation, while seeking the status of nobility—with its attendant exemption from taxation—for themselves. In order to repress urban conflicts, the crown began to send corregidores (governors) to supersede local governmental officials; at first this was a temporary measure, but by the 15th century it had become a permanent arrangement.

The trend toward absolutism encountered opposition both from the nobility and from the Cortes. Under pressure from the Cortes, John I in 1386 organized a royal council (*consejo real*) composed of representatives from the three estates, who would have a permanent influence on royal policy. The experiment failed, chiefly because the nobility succeeded in gaining control of the council and in the 15th century used it for their own advantage. The magnates wielded great power in the territorial administration of the realm, for they held the posts of *adelantado mayor* (governor) in Castile, Murcia, and Andalusia. Moreover, the Trastámara kings gave the magnates extensive territorial lordships as well as lordships over some of the principal towns and cities. In so doing, the crown divested itself of control of some of the major bulwarks of royal power in the past. In general, a proaristocratic mentality pervaded Castilian society in the 15th century, encouraging the usurpation of royal authority by the magnates.

The Cortes played a prominent role in the political life of the kingdom, especially from about 1250 to 1350, and again during the reigns of the first two Trastámara kings, who needed its support to consolidate the dynasty. The Cortes met frequently to give consent to taxation (*servicio*) for fixed periods and for specific purposes. Ferdinand IV was compelled to declare formally that he could not levy an extraordinary tax without consent, and John I had to allow the Cortes to audit his accounts. Petitions presented by the Cortes and accepted by the king became the law of the land. The role of the Cortes was circumscribed in the reigns of John II and Henry IV, who

The Cortes

abused the rights that the assembly had gained in the 14th century. The crown often appointed the procurators to represent the towns, and the number of towns summoned declined steadily. Many towns were not summoned because they were held in lordship by magnates who claimed to speak for them.

The Castilian economy, for the most part, remained agricultural and pastoral. Sheep and cattle grazing became extremely important as immense pasturelands were opened up in Extremadura and Andalusia. The Mesta, a national sheep-raising organization chartered by Alfonso X, wielded great influence. The manufacture of woolen cloth became an important industry in the towns, but on the whole the urban middle class was weak, and a powerful guild organization never really developed. The urban aristocracy, which had come to power in the wars of the Reconquista, was not inclined to engage in trade and crafts. A substantial maritime trade with England, Flanders, France, and Portugal was carried on by the ports of the Bay of Biscay. The Genoese maintained a large colony at Sevillé, the chief southern port, and had a great share in the carrying trade originating there.

The Black Death

The outbreak of the Black Death in the middle of the 14th century contributed to a radical decline in population and a steady succession of social and economic upheavals. In order to compel workers to accept the same wages as before the plague, Peter I enacted the Ordenamiento de Menestrales in 1351. Forty years later, social unrest vented itself in a great pogrom against the Jews, which quickly spread to many parts of the peninsula. Accepting Christianity rather than death, many Jewish converts (*conversos*) attained prominence in various aspects of public life, but they were always held in suspicion and accused of practicing Judaism in private. Deliberate attempts were made to exclude them from important positions by requiring them to prove *limpieza de sangre*; *i.e.*, that their ancestry was untainted by Jewish or Muslim blood.

Culturally, Castile was now fully integrated into western European life. Alfonso X gave great impetus to learning and to the development of Castilian as a literary language. Under his aegis, translations from the Arabic were undertaken, a "General Chronicle" of Spain was written, the Siete Partidas was compiled, and he personally gained renown as the author of the *Cantigas de Santa María*, a collection of more than 400 poems. His nephew Juan Manuel (died 1349) won lasting fame as the author of *El Conde Lucanor* (a collection of Oriental fables), the *Libro de los estados* (a treatise on social classes), and numerous other works. His contemporary Juan Ruíz, the archpriest of Hita (died *c.* 1350), who wrote the *Libro de buen amor*, ranks as one of the greatest medieval poets. Interspersed with beautiful hymns in praise of the Virgin Mary are ribald and comic poems of love. A series of royal chronicles relates the history of this period. Perhaps the best among them are the chronicles of the reigns of Peter I, Henry II, John I, and Henry III by Pedro López de Ayala (died 1407), a brilliant writer who searched for motives and who realized the importance of social and institutional developments. In the 15th century, the influence of the Italian Renaissance was being felt at the court of John II, a patron of many poets and scholars. Among them were the marquess of Santillana (died 1458), whose lyric poems have great beauty; Jorge Manrique (died 1479), whose *Coplas*, written on the death of his father, are a reflection on the vanities of the world; and Fernán Pérez de Guzmán (died *c.* 1460), whose *Generaciones y semblanzas* is a series of sketches of the characters and personalities in the court of Henry III.

Aragon, Catalonia, and Valencia, 1276–1479. The aristocratic challenge to royal power that characterized Castilian history of the later Middle Ages was also evident in the kingdom of Aragon. The growing influence of Roman law and lawyers evoked protest in both Aragon and Catalonia. In 1265, at the Cortes of Egea, James I had to confirm traditional Aragonese laws and agree that the nobles should be tried by the *justicia*, a judge appointed by the king from the knightly class rather than from the ranks of professional jurists. The controversy reached an acute stage during the reign of Peter III the Great (1276–85), who received Aragon, Catalonia, and Valencia as his share in the inheritance while his younger brother James received the kingdom of Majorca.

Annexation of Sicily

Peter III conquered Sicily in 1282 in the wake of the rebellion against French rule known as the Sicilian Vespers. Peter's wife represented Hohenstaufen claims to both Naples and Sicily, in opposition to Charles of Anjou, to whom the papacy had awarded the kingdom. The annexation of Sicily was the second major step in the expansion of Catalonia into the Mediterranean, but it was also the cause of a long international conflict with serious domestic complications. The Pope excommunicated and deposed Peter III, offering his realm to a French prince. In the midst of these troubles, the Aragonese nobles formed a union to defend their liberties and in 1283 forced the King to grant the General Privilege. Confirming the right of the nobles to be judged by the *justicia*, the King also promised to convene the Cortes annually and to take its counsel in matters affecting the realm. He made similar promises in Catalonia and Valencia, but there he could count more fully on the support of his subjects. The crusade launched against him by the Pope and led by King Philip III of France ended in failure, but the problem of Sicily remained the primary concern of Peter's three sons.

Alfonso III (1285–91) inherited the peninsular dominions and also annexed Majorca because of his uncle's treachery during the crusade. His younger brother, James, inherited Sicily. Together, they struggled valiantly to overcome the formidable opposition of the papacy, the French, and the Angevins. The Aragonese nobles again threatened the King and compelled him to issue two privileges in 1287, confirming the rights already ceded by his father and allowing them to appoint and to remove certain members of the royal council. James II (1291–1327) succeeded his brother as king of Aragon, but in order to obtain an undisputed title, he yielded his rights to Sicily in 1295 and also restored Majorca to his uncle James. In compensation, Pope Boniface VIII gave him title to Sardinia. The third brother, Frederick, was proclaimed king of Sicily, and in 1302 the Pope finally agreed to accept him. The Catalan Company, a mercenary force that had fought in the Sicilian wars, found employment thereafter in the Byzantine Empire and eventually established lordship over the duchy of Athens in 1311. Sicily and Athens, though not directly under the rule of the king of Aragon, remained outposts of Catalan influence and power.

James II, who achieved a favourable modification of his frontier with Murcia, occupied Sardinia only in 1325. Genoa contested Aragonese rights, and a series of wars on this account took place during the reigns of Alfonso IV (1327–36) and Peter IV the Ceremonious (1336–87). In 1343, Peter IV dispossessed his cousin, the king of Majorca, for disloyalty and reunited that kingdom to the crown of Aragon. A constitutional crisis arose in 1347 when the King proclaimed his daughter as heir to the throne, contrary to the ambitions of his brother, the Count of Urgel, who argued that only males could wear the crown. The Aragonese union, which had been relatively quiet in the preceding reign, renewed its challenge to royal authority and forced Peter to confirm the privileges granted by his predecessors. The Valencian nobility also formed a union and extorted similar concessions, but the ravages of the Black Death and the triumph of the royalists over the Aragonese at Epila in 1348 enabled the King to dissolve the union and to tear up its privileges. The union never appeared again.

The constitutional crisis of 1347

After midcentury, Peter IV had to defend his realm against a vigorous invasion by Peter I of Castile. He supported Henry of Trastámara's bid to gain the Castilian throne, but his subsequent efforts to secure territorial concessions from Henry II were unsuccessful. In 1377, however, he was able to reincorporate Sicily into the dominions of the crown of Aragon. Although Peter IV was inclined to neutrality in the Great Western Schism, his son John I (1387–96) recognized the Avignonese pope. Continued unrest in Sardinia and Sicily occupied much of his attention and that of his younger brother and

successor, Martin I (1396–1410). The death of the latter without immediate heirs created the most critical situation ever to confront the crown of Aragon. Numerous claimants came forward, but none had wide support. In order to resolve the issue, the estates of Aragon, Valencia, and Catalonia appointed nine commissioners who met at Caspe. Their decision, the compromise of Caspe, published in 1412, declared that the throne belonged by right of inheritance to Fernando de Antequera, brother of King Henry III of Castile. Despite opposition from the Count of Urgel, Ferdinand I (1412–16) became the first of the Trastámara family to rule in Aragon and thereby prepared the way for the eventual union of Aragon and Castile. His principal accomplishment was the withdrawal of obedience from the Avignonese pope Benedict XIII, thus helping to bring the schism to an end.

His son, Alfonso V the Magnanimous (1416–58), generally neglected his peninsular realms to pursue ambitions in Italy. There he seized the Kingdom of Naples in 1443, aspiring to dominate the rest of Italy and even to extend his power into the eastern Mediterranean. His long absence from Aragon fostered a spirit of discontent that reached a crisis during the reign of his brother, John II (1458–79), who inherited the mainland kingdoms as well as Sicily. (Naples was given to Alfonso V's illegitimate son, Ferrante.) By his marriage in 1420, John had become king of Navarre, thus adding a third kingdom to the Trastámara holdings. Conflicts with his son, Prince Charles of Viana, alienated his subjects and aroused overt Catalan hostility. When Charles died suddenly in 1461, many supposed he had been poisoned. The Catalans, who had grown restive because of economic and social hardships, revolted and tried to find another king to rule over them. Meanwhile, Louis XI of France took advantage of the situation to occupy Roussillon and Cerdagne, thus laying the groundwork for future hostilities between France and Spain. By 1472, however, John II had succeeded in overcoming the Catalan revolt and was able to assist his daughter-in-law Isabella in securing the Castilian throne. When he died in 1479, his son Ferdinand succeeded as king of Aragon and Sardinia, and his daughter Eleanor inherited Navarre.

Aragonese institutions, society, and culture. During the late Middle Ages, the Aragonese confederation, amplified and at times dismembered or nearly so, consisted of the essential nucleus of Aragon, Catalonia, and Valencia. In 1319 James II declared the indissoluble union of these three states. From time to time, Majorca, Sicily, Sardinia, and Naples were added to, occasionally separated from, and finally reunited to the crown of Aragon. An attempt to sever parts of Valencia for the benefit of Peter IV's younger brother was opposed not only by the King but also by the people of Valencia. Only at the close of the Middle Ages was the Aragonese federation menaced by dissolution, when the Catalans rebelled against John II and tried to secede, without success. The crown of Aragon had been held as a fief of the papacy ever since Peter II in 1204; this allowed the papacy to intervene directly in Aragonese affairs and even to try to dispose of the throne.

As in Castile, the bureaucracy was much more complicated; it came to be dominated by jurists who encouraged royal absolutism. Peter IV, known as the Ceremonious, took special care to organize the administration around the four principal officials—the chancellor, the chamberlain, the majordomo, and the *mestre racional*, or chief financial officer. There was intense opposition in each of the several states to the employment of non-natives in administrative positions. Even after the suppression of the union of Aragonese nobles by Peter IV, the *justicia* continued to be the ordinary judge in litigation involving them, but he always served at the king's pleasure.

The king's oldest son usually was named as procurator general for all the realms and served as his father's lieutenant. In Valencia and Majorca, which the king visited only occasionally, a procurator or governor general became a permanent representative of the crown. The towns gained a much greater degree of self-government, though the crown was represented in them by a vicar or *justicia*. In Barcelona, a council of five acted as the executive authority, responsible to a larger *consell de cent*, or council of 100.

Revisions of the law

The Usages of Barcelona remained the basic law for Catalonia and Majorca, and in 1251 James I had to pledge that Roman law would not replace them. After the conquest of Valencia, he promulgated the Fori regni Valentiae (1240), a code of law—largely Roman—for use in that kingdom. The customary law of Aragon was compiled in the Code of Huesca, promulgated in 1247. As additions were made to it, a new compilation called the Fueros de Aragón was issued in the 15th century.

The states of the crown of Aragon developed a vigorous parliamentary structure. Each of the mainland states had a parliament, and from time to time the king convened a general assembly for all three. In 1283, at Barcelona, Peter III made two significant pledges of fundamental importance for Catalan constitutional development. He promised to convene an annual Corts "to treat of the good estate and reformation of the land," and he also promised not to make any general statute for Catalonia without the consent of the Corts. Although the assembly did not meet annually, it did gain great power in legislation, as the body of constitutions and *capitols de Corts* enacted in these gatherings makes clear. The parliaments in the three states also voted taxes. The Catalans, distrustful of the crown in this respect, created a commission of the Corts to control collection and expenditure of taxes authorized by the assembly. This commission became permanent in the latter part of the 14th century and was known as the Diputació del General de Catalunya, or simply the Generalitat. It represented the Catalan community when the Corts was not in session, and it assumed the responsibility for defending Catalan liberties against royal encroachments. Similar bodies were created in Aragon and Valencia early in the 15th century.

In Aragon proper, agriculture and pasturage were the chief economic activities. In lower Aragon and Valencia, farming was largely in the hands of the numerous class of Mudéjares, Muslims living under Christian rule and retaining many of their own institutions. The manufacture of woolen goods was the main industry in the towns of Catalonia, where a strong guild organization was developed. Barcelona and Valencia ranked among the major ports of the Mediterranean, trading with Tunis, Alexandria, Sicily, Sardinia, the Holy Land, and the Black Sea area. Settlements of Catalan merchants and consuls to represent their interests were found in many of the principal African ports. A compilation of Catalan maritime law, known as the *Llibre del consolat de mar* (*Book of the Consulate of the Sea*), was prepared about 1370 and quickly gained general acceptance in Europe.

Massacre of the Jews

After the Black Death worked its destruction, frequent social and economic crises disturbed the peace and resulted in such horrors as the massacre of the Jews in Valencia and other cities in 1391. The decline of trade, financial difficulties, and the failure of many private banks in the late 14th century added to the general misery. In Barcelona a popular party (the Busca) began to agitate against the ascendancy of the wealthy urban aristocracy (the Biga) and helped to bring on the Catalan revolt against John II. The *payeses de remensa* joined in the struggle in hopes of gaining their liberty.

The crown of Aragon enjoyed a flourishing cultural life in the late Middle Ages. James II, in 1300, founded the University of Lérida, the first in Aragonese territory, and others were later established at Huesca, Barcelona, and Saragossa. Outstanding representatives of the university tradition were the Dominicans St. Raymond of Peñafort (died 1275)—a great canonist—and St. Vincent Ferrer (died 1419), a preacher of exceptional eloquence. A series of historical works, generally regarded as gems of Catalan literature, helped to mold the vernacular as a literary language: *The Chronicle of James I*, purportedly written by the King himself; Bernat Desclot's *Chronicle of the Reign of King Pedro III*; Ramón Muntaner's *Chronicle*, recounting the adventures of the Catalan Company; and the *Chronicle of Peter IV*, of which the King claimed to be the author. The most prolific writer of the era, equally facile in Latin, Catalan, and Arabic, was

Ramón Llull (died 1315). A philosopher, theologian, and mystic, he was also an active propagandist concerned about the recovery of the Holy Land and the conversion of the Muslims. His writings include *The Book of the Lover and the Beloved*; *Blanquerna*, a romance; and *The Book of the Order of Chivalry*. Francesc Eiximenis (died 1409) was best known for his *Lo Crestià*, an encyclopaedic work dealing with moral and political theory; and the poet Ausiàs March (died 1459) explored the psychological dimensions of love. (J.F.O'C.)

III. Muslim Spain

POLITICAL HISTORY

The conquest. During the 7th century, Byzantine strongholds in North Africa gave way before the Arab advance. Carthage fell near the end of the century, and in the early 8th century the caliph al-Walīd I appointed a new governor in the west, Mūsā ibn Nuṣayr, who annexed all North Africa as far as Tangier (Ṭanjah), leavin his general Ṭāriq ibn Ziyād to administer and Islāmize the native Berbers. Only Ceuta (Sabtah), which was supplied from Spain by the Goth Witiza, remained in Christian hands (see also NORTH AFRICA, HISTORY OF; CALIPHATE, EMPIRE OF THE).

But the Muslim invasion of Spain was the result of Visigoth invitation rather than Muslim initiative. The death of Witiza in 709 precipitated civil war in the Visigothic kingdom, in which the dispossessed family of Witiza and their partisans appealed to Mūsā for support against the usurper Roderick (see above *The Visigothic kingdom*). Ṭāriq crossed the Strait of Gibraltar with a Berber army and in 711 decisively defeated Roderick at a battle the precise location of which is disputed.

Ṭāriq

Ṭāriq, having accomplished his purpose, should have returned to Africa; but, believing that the legendary treasure of Solomon was located at Toledo (Ṭulayṭulah), he marched north and took the city. Much to his surprise, many Spaniards submitted to him voluntarily. In the following year, Mūsā himself brought an Arab army to the peninsula. Mūsā reduced Mérida (Māridah), the last stronghold of Roderick's followers, entered Toledo (Ṭulayṭulah) and Saragossa (Saraguṣṭah) (modern Zaragoza), and may have crossed the northern Meseta Central, forcing the Visigoths to submit or flee. Both Mūsā and Ṭāriq were recalled to the East by the Caliph; by the time of their departure most of the Iberian Peninsula was under their control.

The swift surrender of the peninsula to an army numbering at most 20,000 men appears to be accounted for by the fact that the Hispano-Visigoth society had not yet succeeded in achieving a compact and homogeneous integration. The Goths and Hispano-Romans preserved a clear consciousness of their differences, while lords and vassals found themselves in opposite camps because of profound economic and social factors. The Jews, harassed by the legal ordinances of Toledo, were particularly hostile toward the Christian government. Moreover, the Muslim conquest brought advantages to many elements of society: the burden of taxes was on the whole less onerous than it had been in the last years of the Visigoth epoch; serfs who converted to Islām advanced into the category of freedmen (*mawlās*) and enrolled among the dependents of some conquering noble; Jews were no longer persecuted and were placed on an equal footing with the Hispano-Romans and Goths who still remained within the Christian fold (the Mozarabs). Thus, in the first half of the 8th century, there was born a new and quite different society in Muslim Spain. The nobility of blood was now comprised of the Arab invaders ranked on the basis of the date of their arrival in the peninsula. The *baladiyyūn* were those Arabs who entered Spain in 712 under Mūsā ibn Nuṣayr; the Syrians were those who arrived in 740 under Balj. After the Arab aristocracy were the Berbers, who constituted the majority of the invading troops and whose numbers and influence constantly increased over the course of the centuries because of their continual influx from Africa. The Muslims, or *mawlās*, and their descendants, the *muwallads*, comprised by far the majority of the population, since, as indicated above, economic motives induced a considerable number of landowners to abandon Christianity. Then came the Mozarabs, whose numbers diminished in the course of the centuries; the Jews; and, on a lesser scale, Turkish and eastern European slaves who entered the peninsula later.

Muslim society

This new society was still in a period of gestation when decisive clashes took place between the Muslim troops, led by ʿAbd-ar-Raḥmān al-Ghāfiqī, and the army of the Franks, under Charles Martel at Poitiers (732). Between 711 and 756 Spain was ruled as part of the amirate dependency of Damascus. The period is so called because the governors of the peninsula were appointed, in the final analysis, by the Umayyad caliph at Damascus. This was a turbulent era, marked by continuous civil wars and by armed clashes between Arabs and Berbers. A major Berber uprising against the Arabs in Africa produced powerful repercussions in the peninsula. The Arab army that the caliph dispatched from Syria to subdue the Berber dissidents was defeated. The Syrian army under Balj was forced to seek refuge in Ceuta, whence it moved on to Spain, where its leaders seized power in Córdoba (Qurṭubah). This situation was altered with the establishment of an independent amirate (756) by ʿAbd ar-Raḥmān I ad-Dākkil, an Umayyad prince who, having succeeded in escaping from the slaughter of his family by the ʿAbbāsids, arrived at Córdoba and became politically independent of them (not religiously; he did not adopt the title of caliph).

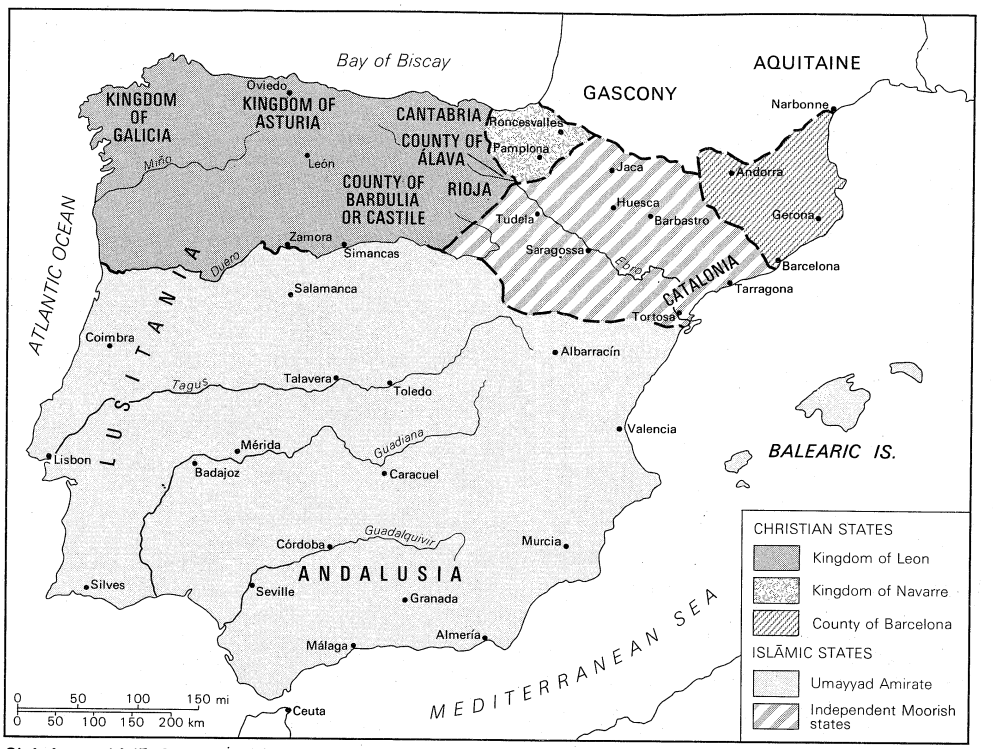
The independent amirate. The dynasty of the Andalusian Umayyads (756–1031) marked the zenith of the Arabic civilization in Spain. Its history may be divided into two major periods—that of the independent amirate (756–929) and that of the caliphate (929–1031)—and may be interpreted as revolving around three persons of like name—ʿAbd ar-Raḥmān I (756–788), II (822–852), III (912–961)—and one dictator, al-Manṣūr (976–1002).

ʿAbd ar-Raḥmān I was the organizer of the new Arab state. Cruel, energetic, and devoid of scruples, he vigorously checked all dissident elements in the peninsula. He endeavoured to base his power on the Eastern aristocracy affiliated with his house and heaped upon it property and riches, though he nonetheless treated it ruthlessly when it showed signs of rebellion; he protected the religious authorities who represented orthodoxy, and, through a series of punitive campaigns, he held in check the Christians of Asturias, who sought to reoccupy the desert to the north of the Duero. In the eastern part of the country he was troubled by intrigues of the ʿAbbāsids, and in the north he had to cope with the ambitions of Charlemagne, who menaced the valley of the Ebro (Ibruh). Charlemagne failed ignominiously in being forced to raise the siege of Saragossa and in the course of his retreat he suffered a defeat at Roncesvalles (778), which is celebrated in the great medieval epic *La Chanson de Roland*. The Franks had to be content with occupying the upper valleys of the Pyrenees. The Frankish advance ended with the Muslim seizure of Barcelona (Barshilūnah) in 801 and Old Catalonia.

ʿAbd ar-Raḥmān I's successors, Hishām I (788–796) and al-Ḥakam I (796–822), were confronted with severe internal dissidence among the Arab nobility. A rebellion in Toledo was put down savagely, and the internal warfare caused the amir to increase the numbers of Slav and Berber mercenaries and to impose new taxes to pay for them.

The reign of ʿAbd ar-Raḥmān II

ʿAbd ar-Raḥmān II inaugurated an era of great political and cultural splendour for Muslim Spain. The arrival in Spain of the Persian musician Ziryāb, for example, marked the beginning of a sharp "Orientalization," or, more precisely, of an "Iraquization" of customs that lent sophistication and refinement to the court. Certain technical advances were introduced, such as the technique of cutting rock crystal by ʿAbbās ibn Firnās and the importation of exotic seeds by the poet al-Ghazāl. ʿAbd ar-Raḥmān's greatest problems sprang from his restless vassals in the Ebro Valley, namely, the convert Banū Qasī family and, above all, before his death, from the Mozarabs,



Christian and Islāmic states of Spain in 910.
Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York

who, incited by extremist chiefs Alvaro and Eulogio (the latter, after his death, being canonized), sought to strengthen their ideas through the aura of martyrdom and began to revile publicly the Prophet Muḥammad—an action punishable by death (850). ʿAbd ar-Raḥmān II sought to persuade the "criminals" to retract, by alleging some pretext or by affirming that the insult had been proffered involuntarily. Failing in this effort, he was obliged to impose the death penalty. The "vogue" of seeking martyrdom voluntarily continued even after the death of the Amir. Both the latter and his successors, however, comported themselves with extraordinary tact and sought to soften the rough edges of friction, and this persecution, provoked by the Christians themselves, took a toll of 53 victims, the last of whom found themselves disavowed by their own ecclesiastical authorities.

In foreign policy ʿAbd ar-Raḥmān II conducted intensive diplomatic activity: he exchanged ambassadors with the Byzantine Empire and with the Frankish king Charles the Bald and maintained friendly relations with sovereigns of Tāhart, which supplied Spain with needed grain. He was able to confront the constantly growing incursions of the Northmen, whom he defeated in the vicinity of Seville. Furthermore, he established permanent defenses against the Viking invaders by the creation of two naval bases, one facing the Atlantic at Seville and another on the Mediterranean shore at Pechina near Almería.

His successors, Muḥammad I (852–886), al-Mundhir (886–888), and ʿAbd Allāh (888–912), were confronted with a new problem, which threatened to do away with the power of the Umayyads—and perhaps with Islām itself—in the peninsula: namely, the *muwallads*, who, having become more and more conscious of their power, rose in revolt in the north of the peninsula, led by the powerful Banū Qasī clan, and in the south (879), by ʿUmar ibn Ḥafṣūn. The struggle against both these enemies was long and tragic—especially against the latter, who was well protected by the walls of the fortress of Bobastro near Córdoba. Ibn Ḥafṣūn was on the verge of delivering the final blow to the Umayyad dynasty, but he lacked the power to win the decisive Battle of Poley (891), which, if it did not put an end to his excursions, at least confined them to eastern Andalusia. ʿAbd Allāh was able to reduce the rebels one after the other, starting with the weakest, and, when he died, he left in the hands of his grandson, ʿAbd ar-Raḥmān III, more than sufficient means to challenge Ibn Ḥafṣūn successfully. The latter, moreover, committed the political stupidity of reverting to the Christianity of his Spanish ancestors, and, consequently, he was deserted by numerous *muwallads*, who regarded themselves as good Muslims. When Ibn Ḥafṣūn died in 917 his sons were forced to capitulate, and in 928 ʿAbd ar-Raḥmān III was able to capture the hitherto impregnable fortress of Bobastro.

The Caliphate of Córdoba. One of the first political problems of an international nature that was posed for ʿAbd ar-Raḥmān III was that of his juridical status vis-à-vis the ʿAbbāsid caliphate at Baghdad. As long as religious unity existed in the Islām dominions, the Umayyads in Spain were content to acknowledge the religious leadership of Baghdad. When after 910 there grew up in Tunis the heterodox caliphate of the Fāṭimids (*q.v.*), ʿAbd ar-Raḥmān III did not hesitate to proclaim himself caliph or to adopt the caliphal title of an-Nāṣir (929). The new Muslim state, known as the Caliphate of Qurṭubah, was to rule in Spain for little over a century. His internal situation was already almost assured; the last bulwarks of resistance were not long in capitulating (Toledo, 933), and from that time on he was able to devote all his efforts to foreign affairs. As to Christian Spain, his successes were meagre, and, what was more serious, he suffered a severe defeat in 939 at Simancas (Shānt Mānkus). From then on, however, the internal debilitation of the Kingdom of Leon enabled him to restore his predominance in its peninsula by political means. He consolidated his position by a series of embassies to the Roman emperor Otto I, to the Christian sovereigns of the peninsula, to the Pope, and to Constantinople. He furthermore extended his protection to the Muslim pirates who established a base in southern France. In Africa the Fāṭimids fought the establishment of an empire that would reach as far as the Atlantic and encompass Andalusia (al-Andalus). And, in order to forestall a Fāṭimid attack on Spain, an-Nāṣir occupied Melilla (Malīlah) and Ceuta (931). Intense naval warfare between the two western caliphates coincided with clashes on land in Morocco and attempts at subversive wars in the enemy states. In the latter area, an-Nāṣir came near to overthrowing the Fāṭimid caliphate by his support of the rebel Abū Yazīd; the conflict between the Umayyads and the Fāṭimids dragged on and ended when the latter conquered Egypt (969) and lost interest in Morocco,

thus leaving a power vacuum that was rapidly filled by the Umayyads.

An-Nāṣir was succeeded by his son al-Ḥakam II (961–976), who adopted the caliphal title of al-Mustanṣir. His reign was peaceful, and he succeeded in resolving the problem of Morocco, thanks to the strategic ability of General Ghālib and the policy of the general of the African army, Abuʿ Āmir al-Maʿāfirī, later known as al-Manṣūr, the Victorious One (Almanzor).

Al-Manṣūr

On the death of al-Mustanṣir, his throne was occupied by his son, a minor, Hishām II al-Mu'ayyad, who grew up under the tutelage of his mother, Aurora, and of the prime minister al-Muṣḥafī, who before long was liquidated by al-Manṣūr (governed 976–1002). The latter succeeded in eliminating all temporal power of the caliphs, whom he dominated, and acquired dictatorial power for himself.

He won complete control over Morocco, which he transformed into the viceroyalty of Córdoba, and he arrested the expansion of the Christian kings from the north through a series of raids—usually every six months—in which he sacked nearly every Christian capital. With the support of a professional army consisting predominantly of Berbers, many of them recent arrivals from Africa who obeyed him blindly, he managed to dispense with the Arab aristocracy, which for the most part was pro-Umayyad, and to hold in check the influence of the slaves, whose numbers had been increasing since an-Nāṣir had placed them in posts of high responsibility. But this balancing of forces—Arabic aristocracy, Berbers, and slaves—could only be sustained by the strong hand of a ruler such as himself.

Al-Manṣūr played the role of a grand lord. A protector of poets and scholars, he concealed his rationalism under a cloak of piety and was the darling of the *faqīh*s (scholars versed in the traditions of Islām); he contrived to attract to himself the outstanding poets of that era. At the time of his death, after a victorious campaign in Aragon, he succeeded in leaving a robust and well-organized state for his son al-Muẓaffar.

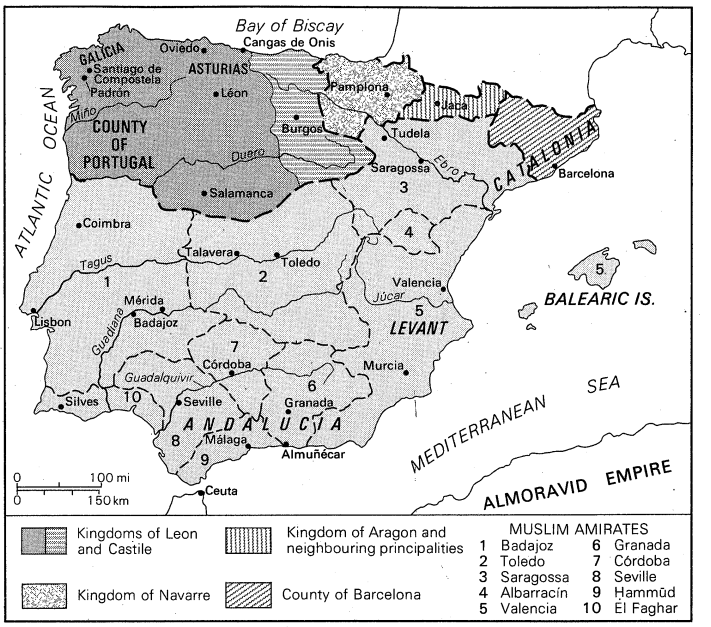
Al-Muẓaffar (1002–08) continued the policies of his father. He hemmed in Hishām II and fought against the Christians, but he died prematurely; his brother ʿAbd ar-Raḥmān Sanchuelo lacked the fortitude to maintain the structure built by his father. An uprising that sought to vindicate the political rights of Hishām II, whom Sanchuelo kept in seclusion, resulted in his death and brought about the beginning of the end of the Umayyad dynasty in Spain.

The ṭā'ifas. Upon the death of ʿAbd ar-Raḥmān Sanchuelo, 21 years of unrest followed (1009–31), during which al-Mu'ayyad was lost trace of and during which constant clashes occurred among Arabs, Berbers, and the slaves. The consequence of those years of anarchy was the formation of numerous independent kingdoms, or *ṭā'if*as, which may be classified into the following: (1) Arab factions in the valleys of the Guadalquivir (al-Wādī al-Kabīr) and Ebro rivers; (2) Berber factions in the Meseta Central, in the steppelike part of Aragon, in Granada (Gharnāṭah), and in nuclei dispersed throughout western Andalusia (Gharb al-Andalus); and (3) groups of slaves in the east.

The political history of the period constitutes an uninterrupted series of internecine wars. Pre-eminent is the confrontation between the Arab factions, under the leadership of Seville (Ishbīliyah), governed by the dynasty of the Banū ʿAbbād, and the Berbers, presided over by Granada. Little by little, Seville succeeded in uniting southern Spain under its aegis. This state was ruled over, first of all by al Muʿtaḍid, a sovereign devoid of scruples, who pretended at first to have found the vanished Hishām II al-Mu'ayyad (at the most, the pretender was a mat maker from Calatrava who bore some resemblance to the old caliph), and then by a poet-prince, his son al-Muʿtamid. In the east, except for a brief period when the petty state of Denia (Dāniyah) built a powerful fleet that enabled it to stage incursions throughout the western Mediterranean as far as Sardinia, the various *ṭā'if*as preserved a certain static equilibrium; farther to the north, the various *ṭā'if*as also wasted their time in interminable internal quarrels.

This fragmentation facilitated the expansion of the Christian states of the north, which, lacking the demographic potential to repopulate the lands they had succeeded in occupying, wisely annexed only those that they were capable of repopulating and garrisoning. The Christian states also imposed a heavy economic burden of tribute on the *ṭā'if*as. Christian armies forced the Andalusian petty kings to buy peace by paying annual tribute, the famous *parias*. This tribute, while it revitalized the economy of the Christian states, created sharp friction between the Muslim authorities and their subjects. The *ṭā'if*as had constantly to increase the yield from their imposts, and they constantly laid new and heavier tax burdens on their subjects; when cash was lacking, they devaluated the currency, minting low-standard coins that were not accepted by the Christians. This, in turn, gave rise to new tax increases and to popular discontent, which was considerably aggravated by the clerical party of the *faqīh*s. Furthermore, the petty local courts were characterized by extravagant luxury and lavish public outlays. This situation rendered Andalusia ripe for the foreign intervention that came when the Castilians occupied Toledo (1085), the key to the Meseta Central and to the peninsula. The factional chiefs, alarmed by the Christian advance, called in the help of the Almoravids.

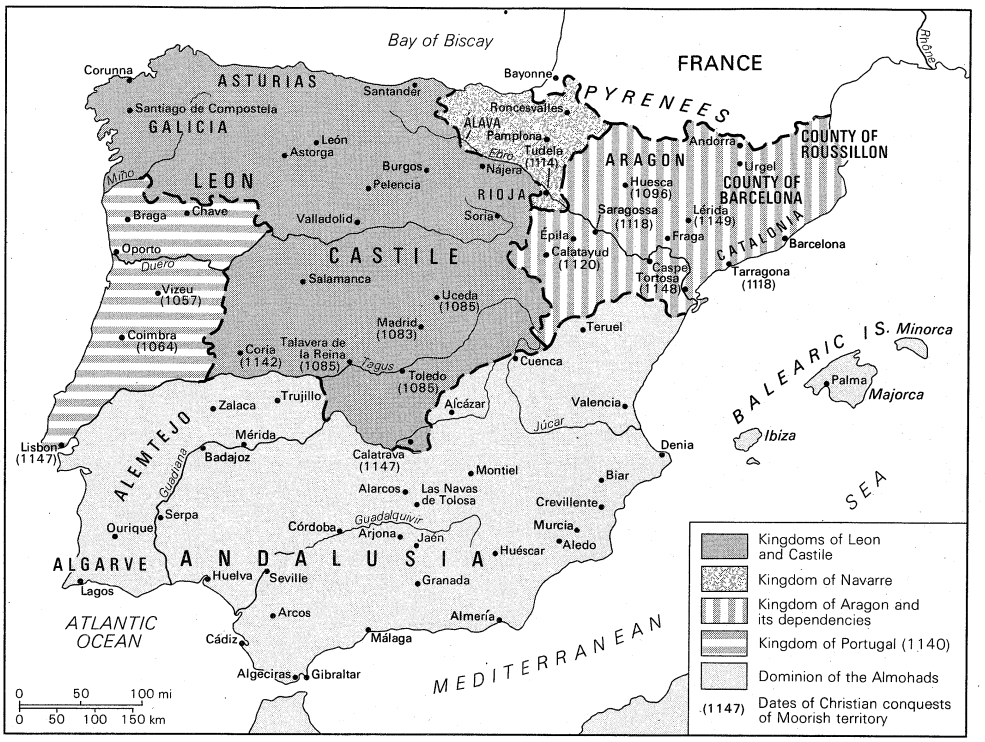
Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York



Spain in 1037.

The Almoravids. The Almoravid ruler Yūsuf ibn Tāshufīn entered the peninsula and by a slow advance reached the fields of az-Zallāqah, north of Badajoz (Baṭalyaws), where in 1086 he defeated a Castilian army under Alfonso VI; then he became an independent prince either unwilling or unable to exploit their victory. For four years Almoravid policy in Spain remained indecisive, but it appears that the siege of Aledo (1088) convinced Yūsuf ibn Tāshufīn of the urgent necessity of putting an end to the *ṭā'if*as if he were going to rescue Spanish Islām. In 1090–91, he deposed their rulers—except for Banū Hūd of Saragossa—and inflicted a series of defeats on the Christians. Only one of the chiefs of state, Rodrigo Díaz de Vivar (the Cid), exiled from his native Castile by King Alfonso VI, was able to resist the Africans, and he furthermore set himself up in an independent kingdom—a new *ṭā'if*a, that of Valencia. The figure of the Cid—the Lord, a title that the Arabs conferred upon him—is quite curious: first he served as a mercenary under Alfonso VI; then he became an independent prince in the east, ruling over states that were entirely inhabited by Muslims. He led a retinue of men who were hardened

The Cid



Spain in 1150.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

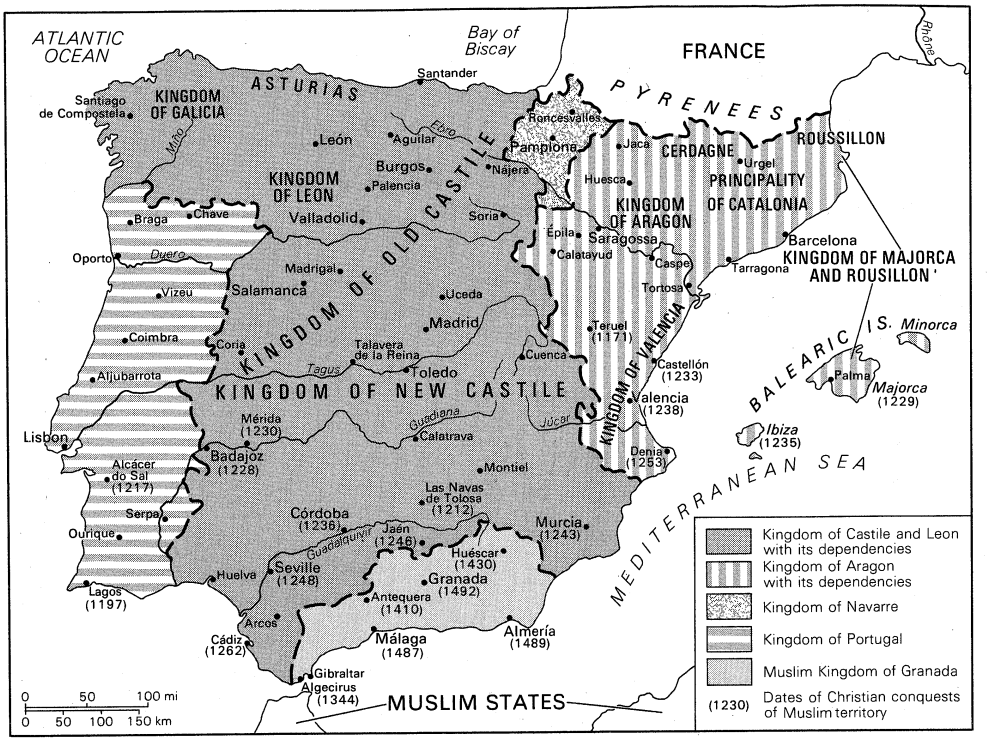
in combat and who understood nothing of the ways of peace. He had the good fortune, however, of finding efficient administrators from among the Mozarabs residing in his states; also, his superb grasp of Almoravid tactics enabled him to overcome his numerical inferiority. Upon his death, Valencia remained under the control of his men until 1102, when they had to evacuate it and seek refuge in Castile. After the fall of Valencia, the Almoravids were unopposed; and in 1110, under ʿAlī ibn Yūsuf (1106–43), they occupied Saragossa.

The conquest of Saragossa, however, marked the beginning of the Almoravid decline. The Aragonese king Alfonso el Batallador (the Battler) and his stepson, Alfonso VII of Castile, launched renewed Christian assaults against the entire frontier of Islām in Spain. The Almoravids, who after 1121 experienced serious difficulties in Africa as a result of the preachings of a Berber reformer, Ibn Tūmart, could not successfully parry the blow; indeed, they had to hire Christian mercenaries to help them. In 1118 Saragossa fell into the hands of the Battler, who furthermore reconquered a large part of the valleys of the Jalón and of the Jiloca. A resounding Almoravid victory over the Aragonese at Fraga (Ifrāgah) in 1134 bore no fruit, because the Almoravids lacked the resources to exploit their victory.

The Almohads. In Africa, the Almohads finally triumphed, and ʿAbd al-Muʾmin (1130–63), successor to Ibn Tūmart, was able to turn his attention to Spain and to undertake the integration of all the Muslim states—the second *ṭāʾifas*—formed under the shield of the latest internecine wars caused by the Almoravid decline. Of these states, there stood out especially that under Ibn Mardanīsh (1147–72), who, with a largely Christian army, succeeded in becoming the master of the entire Levant.

The Almohads assumed the title of caliph, introduced a series of severe religious measures, and sought to strengthen their states through religious unification, in practice, by compelling the Jews and Christians to emigrate. Two great sovereigns, Abū Yaʿqūb Yūsuf (1163–84) and Abū Yūsuf Yaʿqūb al-Manṣūr (1184–99), raised western Islām to the zenith of its power. Taking advantage of the quarrels that divided the Christian kingdoms of Castile and Leon, al-Manṣūr defeated the sovereign of the former, Alfonso VIII, in 1195 at the Battle of Alarcos (al-Arak); but, despite this victory, he proved unable to exploit his triumph—repeating the fate that befell the Almoravids after al-Zallāqah. Years later, during the reign of his successor an-Naṣīr (1199–1214), the Christians avenged this defeat in the Battle of Las Navas de Tolosa (al-ʿIqāb). This battle created a power vacuum into which stepped some of the *ṭāʾifas* or petty kingdom states, prominent among which were those of Banū Hūd of Murcia (Mursiyah) and of the Naṣrids of Arjona (Arjūnah) (later also of Granada). The policies of the two states were quite divergent: Muḥammad ibn Hūd (1228–38) emphasized resistance to the hilt on the part of the Muslims against the Christians who, led by Ferdinand III the Saint, were occupying the Guadalquivir Valley; Muḥammad I ibn al-Aḥmar (ruled in Granada 1238–73), on the other hand, acknowledged himself to be a vassal of the King of Castile and even helped him against his own co-religionists. This realistic policy enabled him to preserve in his possession the territory of what are now the provinces of Málaga, Granada, and Almería, together with portions of neighbouring provinces. Thus, after the middle of the 12th century, after the reconquest of Jaén (Jayyān), Córdoba, Seville, and Murcia by the Castilians and of Valencia and the Balearic Islands by the Aragonese under James I the Conqueror, there no longer remained any independent dominions of Islām in Spain, except Granada and the tiny area of Crevillente (Qarbalyān), which soon disappeared.

Granada. The Naṣrid dynasty founded by Muḥammad ibn al-Aḥmar in Granada, however, was destined to endure for two and a half centuries. The Muslims of Granada lacked sufficient forces to constitute a genuine danger to the Christians, who limited themselves to collecting tribute and launching, from time to time, an attack against them, snatching from them some city or other. The people of Granada, for their part, always bore in mind what had happened in the cases of the Almoravids and the Almohads, who, having arrived from Africa as auxiliary troops, ended up by becoming masters in Spain. Vis-à-vis the new Moroccan empires, particularly that of the Banū Marīns, they maintained a policy of balance of power. They permitted the influx of African volunteers to enroll in their army to fight against the Christians but never permitted the crossing of the Strait of Gibraltar by



Spain in 1300.

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision copyright © 1964 by Barnes & Noble, Inc.

massive organized contingents. The years between 1302 and 1340 are of an extraordinary complexity from a diplomatic as well as a military point of view. Both Morocco and Castile vied for the possession of the Granadine ports of Tarifa (Ṭarīf) and Algeciras (al-Jazīrah al-Khaḍra'), ports that dominated the Strait. Granada, therefore, allied alternately with the Africans and Christians, hoping thus to maintain the balance of power. A fourth state, Catalonia, torn by discord under James II, aspired to obtain a larger slice of the Reconquista and intervened with its fleet and laid siege to Almería (al-Marīyah), under the pretext of a crusade (1309).

When Ismā'īl I (1314–25) ascended the throne, another branch of the Naṣrid family gained the power. Ismā'īl was able to check the reconquest ambitions of Alfonso XI—who later, in 1340, with the aid of the Portuguese, won a decisive victory over the Berber army of Abu'l-Ḥasan at the Battle of the Salado.

The defeat of the Berbers and the lack of interest in reconquest on the part of the successors of Alfonso XI created a favourable climate for Granada, which found itself free from political pressures of both Moroccans and Castilians. During the reign of Muḥammad V (1354–59; 1362–91) Granada reached its greatest splendour; its ministers included some of the most learned men of the epoch, such as the historian Ibn al-Khaṭīb, the physician Ibn Khātima, and the poet Ibn Zamraq. It protected Ibn Khaldūn, the great philosopher of history.

The reign of Muḥammad V

During this era there grew up also the institution of the "judge of the frontier," *juez de la frontera y de los fieles del rastro*, a Moorish official who heard Christian complaints against the Granadinos; this did much to reduce frontier incidents between Moors and Christians.

Little is known about the decline of the Naṣrid dynasty, since with Ibn al-Khaṭīb died the last great Muslim historian of Andalusia. The extant records and reports from the 15th century are as a rule from Christian sources or from the tales of travellers. The narrative poems that are of the utmost interest as historic sources for other periods in Muslim history are completely lacking in this era. The conventional verses of the king-poet Yūsuf III (1408–17) or the anonymous Arab poet of the romance *Abenamar, Abenamar, moro de la morería* do little to illuminate the history of this period. More illustrative, however, are the neglected verses of ʿAbd al-Karīm al-Qaysī (*c.* 1485), an esteemed member of Granada's middle class, who eschewed classic themes and wrote of such mundane phenomena as the increase in the cost of living or the decline in the power of the sovereign.

Foreign relations enjoyed a long period of tranquillity as a result of the ghastly losses of life from the Black Plague (1348–51) and, afterward, from the internal wars that weakened Christian Castile. Furthermore, border incidents were quickly disposed of by the "judge of the frontier and the faithful of the trail." Only an occasional confrontation served to remind the Muslims and Christians that the Reconquista had not yet come to an end. In the 15th century, however, the Reconquista proceeded apace. Prince Ferdinand seized Antequera (Antaqīrah) in 1431; Jimena and Huéscar fell in 1435, Huelma in 1438, and Gibraltar in 1462. One result of these events was that the people of Granada became increasingly less tolerant of Christians, and the Granadine *faqīh*s professed the most extreme xenophobia. This policy, common to both point to the existence of a Granadine school of law, which before long exerted an influence on the other side of the Strait; for the Moroccans, subjected to the constant pressure of the Portuguese—who had gained possession of their coastal areas (Ceuta, first, in 1415)—realized like the Granadinos that the only way to escape Christian hegemony was through the profession of the most rigorous Islāmic ideals and in the practice of the most extreme xenophobia. This policy, common to both sides of the Strait of Gibraltar, did not achieve equal results. It saved Morocco from the hands of its enemies, but in Granada it became the *casus belli* for the campaign that was to inaugurate the end of the Reconquista.

The sultan Muley Hacén (Abū al-Ḥasan ʿAlī) refused to pay the annual tribute he owed to the Catholic kings and seized the fortified town of Zahara (1481), thus launching hostilities that were destined to liquidate the last bastion of Andalusian Islām. The campaign proved to be difficult for the Christian army despite the discord that split the royal Granada family, and that was exploited in Machiavellian fashion by Ferdinand the Catholic: Boabdil (Arabic: Muḥammad Abū ʿAbd Allāh), son of Muley Hacén, rebelled in Guadix against his father and was recognized in Granada with the aid of the Abencerrajes, a powerful Granada family. Muley Hacén, however, who had taken refuge in Málaga, succeeded in re-

The end of Andalusian Islām

capturing the capital with the assistance of the Zegries, or border people. But he was successfully deposed by his brother, el Zagal (Abū ʿAbd Allāh Muḥammad az-Zaghall—the Valiant One), who was supported by the Venegas family.

Boabdil was captured by the Catholic kings during his attack at Lucena and, in order to regain his freedom, signed the Pact of Córdoba, in which he pledged himself to deliver the portion of the kingdom that was in the hands of el Zagal, in exchange for help from the Castilians to recover Granada, part of which (Alhambra) was still in the hands of Muley Hacén. The latter and el Zagal allied themselves against Boabdil, who had to flee and seek asylum in the court of the Catholic king. The death of Muley Hacén in 1485 enabled Boabdil, with the help of the inhabitants of Albaicin, to occupy the Alhambra. El Zagal, who had been routed by the Christians before Vélez Málaga, in 1487 retreated to Guadix and, being incapable of further resistance, delivered his states to the Catholic kings and emigrated to Tlemcén (1491). Taking advantage of this civil war, the Christians had seized Ronda, Marbella, Loja, Málaga, and Razo and were in a position to lay siege to Granada. When the siege began, the population divided into factions: one consisted of pacifists and the other of belligerent persons who, despite their quarrels, fiercely defended the city.

By the end of 1491, the situation became desperate, and Boabdil capitulated, but, before making the news public, he brought a detachment of Castilian troops into the Alhambra on the night of January 1–2, for the purpose of avoiding a disturbance on the part of his vassals that might render it impossible for him to comply with the terms of the pact. The official surrender, and with it the end of Muslim occupation of the peninsula, took place the following day, January 2, 1492.

SOCIAL AND ECONOMIC LIFE

Society. In discussing the influx of the Muslims into Spain, the various categories of the population that settled there were earlier pointed out: Arabs (*baladiyyūn* and Syrians), Berbers, *muwallads*, Mozarabs, Jews, and slaves. The Muslim masses continued to increase during the early centuries of the occupation, because of the wave of conversions that markedly reduced the number of Christians. Precise figures cannot be given, but it is estimated that at the time of the conquest some 4,000,000 Spaniards inhabited the peninsula, while the total number of immigrant Arabs did not exceed 50,000. The population was primarily rural, and large cities were few in number. At the end of the 10th century one can estimate the following urban populations: Córdoba, 250,000; Toledo, 37,000; Almería, 27,000; Granada, 26,000; Saragossa, 17,000; Valencia, 15,000; and Málaga, 15,000.

At the peak of the administrative pyramid was the amir, caliph, sultan, or king, depending on the era, and all the functionaries exercised their power by delegation from the sovereign, who embodied within himself all executive, legislative, and judicial authority, even though at times he delegated power to a *ḥājib* (chamberlain) or, after the 11th century, to a prime minister (*dhū al-wizāratayn*). In the discharge of his functions he was assisted by various viziers. At times, there was, at the head of the various departments, a *kātib*, or official secretary. The provinces were governed by *wālī*s who enjoyed wide autonomy. There existed no uniform municipal organization, and the duties fulfilled by some officials cannot be considered as representative; such are, for example, the chief of police (*ṣāḥib al-shurta*) and the market inspector, known until the 10th century as *ṣāḥib al-sūq* (zabazoque) and later as *muḥtasib*. The Muslim cities of Spain, with their baths, parks, markets, mosques, and high cultural level, were quite different from and, some believe, superior to those of Christian Europe.

The military

The army was based on the voluntary recruitment of soldiers or on contracts with them from abroad. The units (*jund*), grouped according to the places of origin of their men, were strategically deployed along the borders and possessed extraordinary mobility at the time of the caliphate. Holding castles close to the enemy lands as their bases of operation, they were glad to welcome into their midst the Muslims, who were eager to die in combat, in order thus to open for themselves the gates of paradise. These volunteers, who became more and more numerous with the passage of time and about whom numerous details are known, were frequently second class soldiers, since they enrolled during years when they constituted a hindrance rather than a source of help. The *marina*, or navy, organized by ʿAbd ar-Raḥmān II, remained an effective force until the middle of the 14th century.

The entire state structure was supported, theoretically, on a foundation of the most rigid Islāmic orthodoxy and interpreted by the Mālikī school, which in Spain manifested special characteristics of hyperconservative nature. It is not known whether the school acquired these traits upon settling in the peninsula because intolerance was indigenous to the inhabitants there or whether it indoctrinated the Spanish Muslims in this manner, who transmitted it, in turn, to the Christian reconquerors.

The economy. The Moorish conquerors divided the lands seized from the Christians by force of arms and operated them, as a general rule, by means of "tenant farmer" leases. Possibly about the 10th century the woodlands achieved their widest expansion, and the cultivation of irrigated lands was encouraged by means of drastic regulations, which, however, were favourably received. Plants used in the manufacture of textile (flax, cotton, esparto grass, and silk) as well as those with medicinal properties were protected by the state.

In addition to agriculture, the raising of livestock (Christian sheep and Arabian horses) occupied a central position in the peninsular economy. As in the Roman period, lead, iron, gold, and mercury were mined. Domestic industry, which never went beyond the handicraft state, culminated in the production of luxury cloths such as silk (a state monopoly), in the tanning of hides (Cordoban leather), and in the export of ivory objects. Commerce was selective and carried on in products "of low weight and high value" that frequently reached the most remote regions of the known world. There are reports of Andalusian travellers as far as Sudan, central Europe, and even China.

The evolution of economic life was conditioned by political events; in proportion as the productive centres passed into Christian hands, the commercial vigour of the Muslims kept diminishing. No phenomenon is more illustrative of this than the confidence attached to the currency. In the 11th century, Barcelona was counterfeiting Muslim coins; in the 14th century, Granada was doing the same with Barcelona coins.

THE CULTURE OF MUSLIM SPAIN

Arab civilization in the peninsula reached its zenith when the political power of the Arabs began to decline. In the 8th century, in the years immediately following the conquest, there are no traces of a cultural level higher than that attained by the Mozarabs who lived among the Arab conquerors. All available evidence points to the fact that in this period popular works of medicine, agriculture, astrology, and geography were translated from Latin into Arabic. Many of these texts must have been derived from the *Etymologies* of Isidore of Seville and from other Christian writers. In the 9th century, the situation was changed abruptly: the Andalusians, who travelled east in order to comply with the injunction to conduct a pilgrimage to Mecca at least once in their lifetimes, took advantage of their stay in those regions to enhance their knowledge, which they then introduced into their native country.

Literature. In the 9th century there flourished such court poets as ʿAbbās ibn Nāṣih, ʿAbbās ibn Firnās, Yaḥyā al-Ghazāl, and the knight Saʿīd ibn Jūdī. Towering above all these, however, is Ibn Hāniʾ, nicknamed the "Mutanabbī of the West," who by virtue of his religious ideas was obliged to forsake his native land and enter in the service of the Fāṭimid caliph al-Muʿizz. In the 10th century, another bardic tradition grew up around al-Manṣūr. Bards performed the functions of modern jour-

Poetry

nalists; they accompanied their protector on war expeditions and celebrated his exploits in verse the singsong rhyme of which became engraved in the folk memory of the people of Andalusia. This occasional poetry did not always attain literary heights, but at times that was indeed the case. And al-Manṣūr himself chose as "poet-journalists" the foremost talents of his time—men such as Ibn Darrāj al-Qasṭallī, ar-Ramādī, Sa'īd of Baghdad, at-Ṭalīq, and numerous others. And it was in the 10th century that Ibn Faraj of Jaén deemed himself to possess sufficient background to compose the *Kitāb al-Hadā'iq* ("Book of Orchards")—the first anthology of Andalusian poets. This was soon followed by another, that of the physician Ibn al-Kittānī.

The highest peak in Islāmic literature in Spain was attained during the era of the *ṭā'ifa*s, when the poet-king al-Mu'tamid established an embryo of an academy of belles lettres, which included the foremost Spanish intellects, as well as Sicilians who emigrated from their native land before its conquest by the Normans. Other petty kings in the peninsula endeavoured to compete with al-Mu'tamid but did not succeed in assembling a constellation of writers of comparable stature.

In the 12th century, in the Andalusian Levant, outstanding poets included Ibn Khafajā of Alcira and his nephew Ibn az-Zaqqāq. In the era of greatest decadence in the 13th century were Abū al-Baqā' of Ronda (*c.* 1285), Ibn al-Khaṭīb, Ibn Sa'īd al-Maghribī, and Ibn Zamraq, whose verses are preserved in the stuccos of the Alhambra.

Prose

In Arab literature, including that of contemporary times, poetry possesses greater vitality than prose. But, even so, it is necessary to mention Ibn Shuhayd (*c.* 1035), author of a work that lent inspiration to al-Ma'arrī for his *Risālat al-ghufrān* ("Epistle of Pardon"). The prolific Ibn Ḥazm of Córdoba (died 1064) wrote the delightful *Tawq al-ḥamāmah* ("The Ring of the Dove"), dealing with love and lovers, which is still popular today. The enormous output of Ibn Ḥazm is not possible to enumerate here. Nevertheless, it suffices to cite his *Kitāb al-fiṣal,* a history of religions that was not surpassed by Western scholars until well into the 19th century. Another polymath was the vizier-historian Ibn al-Khaṭīb (died 1375). Two 12th-century histories of literature, by Ibn Bassām and Ibn Khāqān, are excellent sources of information concerning the apogee of Andalusian letters. Often the best grammars and dictionaries of a language are written by authors living in the peripheral zones who endeavour to prevent gross errors being committed by their countrymen in the region. This perhaps explains why Andalusia, located at the western fringe of the Muslim world, produced works that to this day are used as texts in certain traditional Islāmic universities. Among these grammarians should be mentioned al-Zubaydi, tutor of Hishām II and Ibn Maḍā of Córdoba, who proposed a drastic reform of grammatical methods. Ibn Mālik of Jaén's didactic poem *Alfiyya* ("The Thousand Verses") constitutes an excellent handbook of grammar; and Abū Ḥayyan of Granada (died 1344), who emigrated to the east, wrote an outstanding commentary on the Qu'ran as well as the first Turkish grammar. In the field of lexicology the blind Ibn Sīda of Denia (died 1066) is pre-eminent; he was also the author of a sort of "dictionary of ideas."

In the field of Qu'rānic science there stand out Abū 'Amr of Denia and Ibn Fierro of Játiva, whose handbooks made possible the correct cantilation, or intonation of the chanting, of the Qu'rān. Also produced there were various collections of *ḥadith*s (traditions referring to the Prophet), but none of these was of particular importance. In this domain, the Andalusians were imitators of the East, and personalities such as Ibn 'Abd al-Barr, a certain Ibn Rushd (the grandfather of Averroës), and an Ibn 'Āṣim are of merely marginal interest.

The first extant chronicles in Muslim Spain, such as the *Ta'rīkh iftitāḥ al-Andalus* ("History of the Conquest of Spain"), by Ibn al-Qūṭiya, date back to the 10th century. In the *ṭā'ifa* era the pre-eminent Spanish historian is Ibn Ḥayyān of Córdoba (died 1076), whose *Muqtabis,* preserved for the most part, is an anthology of historic texts based to a great extent on the work of his predecessors. From the point of view of human interest are the *Memoirs* of the king Ziri 'Abd Allāh, who was deposed by the Almoravids and who sought to justify in those memoirs his deeds as a statesman. In the Naṣrid era is found the frequently aforementioned Ibn al-Khaṭīb. The works of the North African historians Ibn Khaldūn and al-Maqqari (died 1631) supply much information concerning the Arab domination of Spain.

Science. In the mid-11th century, the *qāḍī* of Toledo, Ibn Ṣā'id, composed a noteworthy handbook of history of science that contained much information on technical subjects. Mathematical sciences received little attention, but mention should be made of Maslama of Madrid (died 1008), who probably took part in the translation of Ptolemy's *Planispherium* and made some contributions to pure mathematics. Also during the heyday of Granada, there was al-Qalaṣādī, commentator on Ibn al-Bannā', who did important work on fractions. Despite their lack of interest in the physical sciences, the Andalusians excelled in astronomy, both theoretical and practical. An array of these scholars sought to simplify the astrolabe, and finally al-Zarqālī (died 1087) achieved success by inventing the apparatus called the *azafea,* which was widely used until the 16th century. Al-Zarqālī also anticipated Kepler by suggesting that the orbits of the planets are not circular but ovoid. Jābir ibn Aflaḥ (12th century) criticized the Ptolemaic system.

Astrology was popular in Muslim Spain, and the Umayyad rulers, after 788, maintained an official astrologer in their courts. The most widely used astrological treatises were those of the Tunisian 'Alī ibn Abī Rijāl and another anonymous scientist, who made a translation from Vulgar Latin into Arabic in the 8th century. This book was translated from Arabic into Spanish during the era of Alfonso under the title of *Libro de las Cruces* ("Book of the Crosses").

Natural science

The treatises on esoteric or occult subjects attribute to Maslama two works on natural science that are not properly his but may be ascribed to one of his pupils. They are *Ghāyat al-ḥakīm* ("The Goals of the Scholar"; also known as *Picatrix*) and *Rutbat al-ḥakim* ("The Step of the Scholar"). Greater interest is merited by the *Materia Medica,* a revision of the Eastern Arabic text of the 1st-century Greek physician Pedanius Dioscorides, ordered by an-Naṣir, on which Jews, Arabs, and Christians collaborated. Gradually, the Andalusian Arabs kept adding new medicinal "simples" (which described the properties of various medicinal plants) to those described by Dioscorides, and the last eminent essayist on the subject, Ibn al-Bayṭār (died 1248), describes more than 1,400 of these. The Arabs of Andalusia were familiar with the texts of the great Latin classics relating to natural science, and Ibn Wāfid, Ibn Baṣṣāl, and Ibn al-'Awwām (11th and 12th centuries) quote Varro, Virgil, and others. The most notable geographers in Muslim Spain were al-Bakrī (died 1094), who wrote the *Kitāb al-masālik wa-l-mamālik* ("Book of Highways and of Kingdoms"), and al-Idrīsī (died 1166), who was in the service of Roger II of Sicily and is the author of the leading universal geography composed by the Arabs. Somewhat later (1323) there appeared the first Arabic nautical map, possibly of Granadine origin. In technology, Muslim Spain was noted for its windmills and for its manufacture of paper. (J.V.G.)

IV. Spain from 1479 to 1700

UNITED SPAIN UNDER THE CATHOLIC KINGS

The union of Aragon and Castile. When Ferdinand II (1479–1516, also Ferdinand V of Castile from 1474) succeeded to the crown of Aragon in 1479, the union of Aragon (roughly east Spain) and Castile (roughly west Spain) was finally achieved, and the Trastámara became, after the Valois of France, the second most powerful monarchs in Europe. The different royal houses of the Iberian Peninsula had long thought in terms of a union of their crowns and had practiced intermarriage for generations. Nevertheless, the union of the crowns of Castile and Aragon was far from inevitable in the last quarter of

the 15th century. A union between Castile and Portugal was equally feasible, and it has been argued that it would have made more sense, for it would have allowed the two western Hispanic kingdoms to concentrate on overseas exploration and expansion, and it would not have involved Castile in Aragon's traditional rivalry with France. The reasons that led John II of Aragon to arrange the marriage of his son and heir, Ferdinand, with Isabella of Castile, in 1469, were essentially tactical: he needed Castilian support against French aggression in the Pyrenees. In Castile an influential party of magnates, led by Archbishop Carillo of Toledo and opposed to King Henry IV, supported the succession claims of the princess Isabella, the King's sister, against those of his daughter, Joan. They were anxious for the help and leadership of the Aragonese prince and content with the alliance of a country in which the magnates had such far-reaching privileges as the Aragonese nobility. It needed a forged papal dispensation for the marriage, the blackmailing of Henry IV into (wrongly) denying the paternity of his daughter (Joan), and, finally, several years of bitter civil war before Ferdinand and Isabella defeated Joan's Castilian supporters and her husband, Alfonso V of Portugal.

Reasons for union

Ferdinand and Isabella ruled jointly in both kingdoms and were known as the Reyes Católicos (Catholic Kings). It was, however, a union of crowns and not of kingdoms. In size, institutions, traditions, and, partly, even in language, the two kingdoms differed greatly. Within the Kingdom of Aragon, Aragon and Valencia (easternmost Spain, halfway up the coast) each had about 270,000 inhabitants, of whom some 20 percent and more than 30 percent, respectively, were Moors and Moriscos (Moors officially converted to Christianity). Catalonia (northeast Spain) had about 300,000 inhabitants. In each of these kingdoms the powers of the crown were severely limited. The barons ruled their estates like kings, dispensing arbitrary justice over their peasants. In Catalonia they had the right to wage private war. In Aragon anyone arrested by order of the king could put himself under the jurisdiction of a *justicia* who held his office for life and was therefore independent of the king's pleasure. It was this highest judge who crowned the kneeling king and made him swear to observe the *fueros*, the laws and privileges, of the kingdom. It is now known that the formula "We who are as good as you swear to you who are no better than we, to accept you as our king and sovereign lord, provided you accept all our liberties and laws; but if not, not" is a forgery, most probably of the mid-16th century. It does, however, summarize very well the relations between the kings of Aragon and the Aragonese nobility.

Ferdinand made no attempt to change this position; nor did he do so in Catalonia, where the crown had just emerged successfully from a long and confused civil war. The nobility and the urban aristocracy of Barcelona (Spain's northeast coast) had been faced by violent social movements of the peasants and the lower classes of the cities and were themselves riven by family and factional strife. The crown intervened, mainly on the side of the lower classes but, inevitably, in alliance with some of the noble factions and against the French who had taken the opportunity to occupy Cerdagne and Roussillon. In 1486 Ferdinand settled the Catalan problem by a compromise, the Sentencia de Guadalupe, which effectively abolished serfdom and the more oppressive feudal obligations of the peasants, in return for monetary payments to the lords. Otherwise, the political and legal privileges of the rural nobility and the urban aristocracy were left intact. Effectively, therefore, Ferdinand made no attempt to strengthen the powers of the crown and to give the principality a more efficient system of government. One hundred and fifty years later his successors were to pay a high price for this failure. In the meantime, Ferdinand had given Catalonia peace, a considerable degree of social stability, and the opportunity to make good the ravages of the civil wars and the losses of commercial markets to Italian competitors. This opportunity was only partially taken by the Catalans. They failed completely to prevent the Genoese from establishing a dominant position in the economy of Castile and, more especially, in the vital and rapidly expanding Atlantic trade of Seville (south Spain, 90 miles [140 kilometres] north of Gibraltar). The union of the crowns of Aragon and Castile therefore led to neither a political and institutional union nor to an economic integration of the Iberian Peninsula.

Catalonia's situation

Castile. Castile, too, was a poor country. Much of its soil was arid, and its agriculture was backward. The armed shepherds of the powerful sheep-owner's guild, the Mesta, drove their flocks over hundreds of miles, from summer to winter pastures and back again, spoiling much cultivated land. Despite the violent hostility of the landowners, the government upheld the Mesta privileges, since the Mesta paid generously for them and was supported by the merchants who exported the raw wool to the cloth industry of Flanders (modern Belgium). The power of the Mesta and the harm it did increased in the 16th century, for the Catholic Kings supported its privileges, anxious, as always, to increase their revenues by following the line of least resistance. Thus, the impoverished Spanish peasant was unable to buy the manufactures of his urban industries, and hence the Castilian towns remained comparatively small and their industries underdeveloped, compared with those of Italy and the Netherlands. Only the cities of Andalusia (south Spain) were relatively large, but, except for Seville, they were not primarily trading or manufacturing centres but were dominated by an urban nobility living on revenues from the surrounding countryside.

Underdeveloped Castilian industries

In the northern provinces of Castile there lived a large class of minor nobles, the hidalgos. The inhabitants of Guipúzcoa (by the westernmost French border) even claimed that they were all of noble birth. But the south, New Castile (southeast of Madrid), Extremadura (southwest of Madrid), and especially Andalusia—that is to say, those provinces most recently reconquered from the Moors—were the domain of the great nobility. There the families of the Enríquez, the Mendoza, the Guzmán, and others owned vast estates, sometimes covering almost half a province. They had grown rich during the 15th-century boom in wool exports to Flanders, at a time when there were more than 2,500,000 sheep in Castile, and it was they, with their hordes of vassals and retainers, who had attempted to dominate a constitutionally almost absolute, but politically weak, monarchy.

It was in this kingdom that the Catholic Kings determined to restore the power of the crown. Once this was achieved, or so it seemed, the liberties of the smaller kingdoms would become relatively minor problems. Like their contemporary, Henry VII of England, they had the advantage of their subjects' yearning for strong and effective government after many years of civil war. Thus, they could count on the support of the cities in restoring law and order, without having to make any important concessions to the Castilian Cortes. The Cortes of Toledo of 1480 accepted the appointment of corregidores, or royal officials, to supervise the administration of the principal towns. At the same time, both municipal efficiency and civic pride were enhanced by the obligation imposed on all towns to build a town hall (*ayuntamiento*).

With the great nobles it was necessary to move more cautiously. The Catholic Kings revoked usurpations of land and revenues by the nobility if these had occurred since 1464, but most of the great noble estates had been built up before that date and were effectively left intact. From a contemporary chronicler, Hernán Pérez del Pulgar, historians know how they proceeded piecemeal but systematically against the magnates, here using a nobleman's defiance of the law, there a breach of the peace or of a pledge, to take over or destroy his castles and thus his independent military power. Even more effective in dealing with the nobility was the enormous increase in royal patronage. Isabella was stage manager to Ferdinand's election as grand master of one after another of the three great orders of knighthood: Santiago, Calatrava, and Alcántara. This position allowed the King to distribute several hundred commanderships with their attached income from the huge estates of the orders. Equally important

was royal control over all important ecclesiastical appointments, which the Catholic Kings insisted on with ruthless disregard of all papal claims to the contrary. In the Spanish dependencies in Italy, Ferdinand claimed the right of exequatur, according to which all papal bulls and breves (authorizing letters) could be published only with his permission. There exists a letter of Ferdinand to his viceroy in Naples, written in 1510, in which he upbraids the viceroy for permitting the Pope to publish a brief in Naples, threatens that he will renounce his own and his kingdoms' allegiance to the Holy See, and orders the viceroy to arrest the papal messenger, force him to declare he never published the brief, and then hang him. In Spain the Catholic Kings had no formal right of exequatur, but they and their Habsburg successors behaved very much as if they did. From now on the Spanish clergy had to look to the crown and not to Rome for advancement, and so did the great nobles who traditionally claimed the richest ecclesiastical benefices for their younger sons.

Perhaps most effective of all in reducing the political power of the high nobility was their virtual exclusion from the royal administration. The old royal council, a council of great nobles advising the king, was transformed into a bureaucratic body for the execution of royal policy, staffed by a prelate, three nobles, and eight or nine lawyers. These lawyers, mostly drawn from the poor hidalgo class, were entirely dependent on the royal will and became willing instruments of a more efficient and powerful central government. The Catholic Kings set up a Council of Finance (1480, but not fully developed until much later), the Council of the Hermandad (1476–98), the Council of the Inquisition (1483), and the Council of the Orders of Knighthood (for the administration of the property and patronage of the orders of Santiago, Calatrava, and Alcántara), and they reorganized the Council of Aragon. Charles I and Philip II were later to continue this work and to add further councils, notably those of the Indies (1524) and of Italy (1558).

When Isabella died, in 1504, the nobles appeared to be tamed and politically innocuous. In fact, their social position and its economic basis, their estates, had not been touched. Of the soil of Castile 97 percent was owned by 2 or 3 percent of the population, and some 50 percent of this was owned by a quite small number of families, mostly linked through intermarriage. In 1520 Charles I agreed to their demand for establishing a fixed hierarchy of rank, from the 25 grandees of Spain, through the rest of the titled nobility, down to some 60,000 hidalgos, or caballeros, and a similar number of urban nobility, all of them distinguished from the rest of the population, the *pecheros*, by far-reaching legal privileges and exemption from direct taxation. The Laws of Toro (1505), extending the right to entail family estates on the eldest child, further safeguarded the stability of noble property. The Catholic Kings had therefore shifted the balance of political power within the traditional ruling classes of Castile in favour of the crown, but they had not fundamentally altered the economic and social position of these classes.

Nobility's strong base

The Spanish Inquisition. With its large Moorish and Jewish populations, medieval Spain was the only multiracial and multireligious country in western Europe, and much of the development of Spanish civilization in religion, literature, art, and architecture during the later Middle Ages stemmed from this fact. The Jews had served Spain and its monarchs well, providing an active commercial class and an educated elite for many administrative posts. But, inevitably, their wealth created jealousy and their heterodoxy hatred in a population that traditionally saw itself as the defender of Christianity against the infidel. The Catholic Kings, ever good tacticians, profited from this feeling. In 1478 they first obtained a papal bull from Sixtus IV setting up the Inquisition to deal with the supposedly evil influence of the Jews and *conversos* (converted Jews). Since the Spanish Inquisition was constituted as a royal court, all appointments were made by the crown. Too late, Sixtus IV realized the enormous ecclesiastical powers that he had given away and the moral dangers inherent in an institution the proceedings of which were secret and that did not allow appeals to Rome.

With its army of lay familiars, who were exempt from normal jurisdiction and who acted both as bodyguards and as informers of the inquisitors, and with its combination of civil and ecclesiastical powers, the Spanish Inquisition became a formidable weapon in the armoury of royal absolutism. The Supreme Council of the Inquisition (or Suprema) was the only formal institution set up by the Catholic Kings for all their kingdoms together. Nevertheless, they thought of it primarily in religious and not in political terms. The Inquisition's secret procedures, its eagerness to accept denunciations, its use of torture, the absence of counsel for the accused, the lack of any right to confront hostile witnesses, and the practice of confiscating the property of those who were condemned and sharing it between the Inquisition, the crown, and the accusers—all this inspired great terror, as indeed it was meant to do. The number of those condemned for heresy was never very large and has often been exaggerated by Protestant writers. But during the reign of the Catholic Kings, several thousand *conversos* were condemned and burned for Judaizing practices. The whole family of the philosopher and Humanist Luis Vives was wiped out in this way. Many more thousands of *conversos* escaped similar fates only by fleeing the country. Many Catholics in Spain opposed the introduction of the Inquisition, and the Neapolitans and Milanese (who prided themselves on their Catholicism and who were supported by the popes) later successfully resisted the attempts by their Spanish rulers to impose the Spanish Inquisition on them. Even in Spain itself, it was the sumptuous autos-da-fé, or the ceremonial sentencings and executions of heretics, rather than the institution and its members that seem to have been popular. But most Spaniards never seem to have understood the horror and revulsion that this institution aroused in the rest of Europe.

The fate of the *conversos*

The first inquisitor general, Tomás de Torquemada, himself from a *converso* family, at once started a propaganda campaign against the Jews. In 1492 he persuaded the Catholic Kings to expel all Jews who refused to be baptized. Isabella and most of her contemporaries looked upon this expulsion of about 170,000 of her subjects as a pious duty. At the moment when Spain needed all its economic resources to sustain its new European position and its overseas empire (see below), it was deprived of its economically most active citizens and laid open to exploitation by German and Italian financiers.

The conversos. Nevertheless, the expulsion of 1492 did not signify the end of Jewish influence on Spanish history, as used to be thought until quite recently. It is not, however, easy to establish a clear-cut direction or pattern of this influence. At the end of the 15th century, there may have been up to 300,000 *conversos* in Spain, and the majority of these remained. They had constituted the educated urban bourgeoisie of Spain, and the richer families had frequently intermarried with the Spanish aristocracy and even transmitted their blood to the royal family itself. After 1492 their position remained precarious. Some reacted by stressing their Christian orthodoxy and denouncing other *conversos* to the Inquisition for Judaizing practices. Others embraced some form of less conventional, more spiritualized Christianity. Thus the followers of Sister Isabel de la Cruz, a Franciscan, organized the centres of the Illuminists (Alumbrados), mystics who believed that through inner purification their souls should submit to God's will and thus enter into direct communication with him. While they counted some of the high aristocracy among their numbers, most of the Illuminists seem to have been *conversos*. Again, it was among the *conversos* that Erasmianism (after Erasmus, a famous Humanist of the time), a more intellectual form of spiritualized Christianity, had its greatest successes in Spain. The Erasmians had powerful supporters at court in the early years of Charles I, when, as emperor, his policy was directed toward the healing of the religious schism by a general reform of the church. But in the 1530s and 1540s, the enemies of the Erasmians, especially the Dominican Order, launched a systematic

Converso reactions

campaign against them. The Inquisition annihilated them or forced them to flee the country, just as it had done in the case of the Illuminists as early as the 1520s. Nevertheless, the influence of Erasmus did not completely disappear from Spanish intellectual life, and it has been traced into the latter part of the 16th century.

But the majority of the *conversos* and their descendants probably became and remained orthodox Catholics, playing a prominent part in every aspect of Spanish religious and intellectual life. It has now become clear that without them the "Golden Century" of Spain would be inconceivable. They ranged from such saints as Teresa of Avila and St. John of God, one a mystical writer and founder of convents, the other an organizer of care for the sick, to Diego Laínez, a friend of St. Ignatius of Loyola and second general of the Jesuit Order. They included Fernando de Rojas, author of *La Celestina*, the first great literary work of the Spanish Renaissance, and, two generations later, Mateo Alemán, who wrote a picaresque novel, the *Guzmán de Alfarache;* and they could boast Luis de León, a Humanist and poet; a Dominican, Francisco de Vitoria, perhaps the greatest jurist of any country in the 16th century; and another famous Dominican, the defender of the American Indians and historian of the Indies, Bartolomé de Las Casas.

These, with Luis Vives, mentioned earlier, are only the most famous among the many distinguished *converso* names who played such a central and varied role in creating the cultural splendours of Spain's "Golden Century." It is an extraordinary phenomenon that had no parallel anywhere else in Europe before the 19th or even 20th century. Any attempt at an explanation is bound to be speculative, but the following may be suggested. The Spanish Jews and *conversos* formed a comparatively large section of the relatively small educated elite of Spain who were primarily responsible for the cultural achievements of the period. The *conversos*, moreover, having deliberately broken with the Jewish tradition of Talmudic (from the Talmud, the body of Jewish civil and canonical law) scholarship, found the glittering Renaissance world of Christian Spain ambivalently attractive and repellent but always stimulating. Their response to this stimulus was probably sharpened, their need to excel given a double urgency, by the hostility that they continued to meet from the "old" Christians. For these latter were very much aware of the ubiquity of the *conversos*, however much these were assimilated, and many resented it bitterly. Religious, racial, and even anti-aristocratic class prejudices combined to create the obsession with "purity of blood" (*limpieza de sangre*), which became characteristic of the Spaniards in the 16th and 17th centuries.

Converso contribution

It first crystallized with a statute of *limpieza*, imposed in 1547 on the cathedral chapter of Toledo, by which purity of ancestry from the taint both of *converso* blood and from any accusations of heresy by the Inquisition was made a condition of all future ecclesiastical appointments. The author of this statute was Archbishop Siliceo, a man of humble and hence, by definition, untainted origins, who had found himself despised by the aristocratic canons, many of whom, however, had *converso* blood in their veins. In 1556 Philip II gave his royal approval to the statute on the grounds that "all the heresies in Germany, France and Spain have been sown by descendants of Jews." As far as Germany and France were concerned, this remark was sheer fantasy, and it is especially ironic that, just at this time, Pope Paul IV, at war with Spain, described Philip II himself quite correctly as a Marrano, or descendant of Jews.

But statutes of *limpieza* now spread rapidly over Spain. They helped to perpetuate a set of values, the equation of pure ancestry, orthodoxy, and personal honour, which certainly helped to prevent the spread of heresies in Spain but which in the long run had a blighting effect on Spanish society, the more so as they were linked so closely with the basically corrupt institution of the Inquisition. For not only did they encourage the practice of denunciation, with inevitably corrupting and divisive effects, but they also educated generations of Spaniards to see their personal honour as a value depending on a set of circumstances substantially outside their own control or achievement.

Deadening effect of *limpieza* statutes

The conquest of Granada. The impact of the Moors on Spanish life and traditions had been rather different from that of the Jews. It was most evident, perhaps, in the position of women in southern Spain, who for long remained semi-veiled and in much greater seclusion than elsewhere in Christian Europe. It was evident also where Jewish influence was practically nonexistent, in the visual arts and especially in architecture. Not only did houses in southern Spain for a long time continue to be built facing inward, on to a patio, but a whole style of architecture, the Plateresque, derived from an imaginative fusion of the Moorish and the Christian: classical Renaissance structures were decorated with Gothic or Renaissance motifs but executed in the Moorish manner, as if a carpet had been hung over the outside wall of the building. This charming style, which was invented during the reign of the Catholic Kings, spread far and wide over Spain and eventually even to the New World.

To Ferdinand and Isabella, the Moorish problem presented itself in the first place in a political and military form, for the Moors still ruled their independent Kingdom of Granada (Spain's southeast corner). The Catholic Kings had to concentrate all their military resources and call on the enthusiastic support of their Castilian subjects to conquer the kingdom in a long and arduous campaign, which ended with the capture of Granada (some 130 miles east of Seville), the capital, in 1492. In this campaign Gonzalo Fernández de Córdoba, the "Great Captain," developed the tactics, training, and organization that made Spanish infantry almost unbeatable for 150 years. The Moors were granted generous terms and religious freedom. But, against the advice of the saintly Hernando de Talavera, the *converso* archbishop of Granada, who was trying to convert the Moors by precept and education, the Queen's confessor, Francisco (later Cardinal) Jiménez de Cisneros, introduced forced mass conversions. The Moors rebelled (1499–1500) and, after another defeat, were given the choice of conversion or expulsion, Jiménez and Isabella did not regard this new policy as a punishment of the Moors for rebellion, for Christian Baptism could never be that. It was rather that the rebellion released them from their previous undertaking, which they had entered into only with misgivings. Though many Moors chose conversion, the problem now became virtually insoluble. There were never enough Arab-speaking priests or money for education to make outward conversion a religious reality. The Moriscos remained an alien community, suspicious of and suspect to the "old" Christians. There was with the Moriscos very little of the intermarriage with Christians and of the deliberate acceptance of Spanish Christianity that, in spite of all the statutes of *limpieza*, allowed the *conversos* to become such an integral part of Spanish society.

Hopelessness of the Moorish position

Acquisition of Naples. Ferdinand's foreign policy was dominated by the traditional Aragonese rivalry with France along the Pyrenees and in Italy. Aragon still held Sicily and Sardinia from the much more extensive medieval Aragonese empire. French intervention in Italy, from 1494, gave Ferdinand his chance. Charles VIII of France, to secure his southern flank while he led his army into Italy, agreed to return to Ferdinand the counties of Cerdagne and Roussillon (Treaty of Barcelona, 1493), which Louis XI had seized during the Catalan civil wars, in 1463. But it was through Ferdinand's own diplomacy and through the generalship of Gonzalo de Córdoba that he acquired the Kingdom of Naples (1503). For the first time the union of Aragon and Castile had shown its strength, and Spain now rivalled France as the most powerful state in Europe. Ferdinand had carefully arranged the marriage of his children to strengthen his diplomatic position against France by alliances with Portugal, England, and Burgundy (which ruled the Netherlands). The unexpected deaths of the two eldest and their children, however, left the succession of Castile after Isabella's death (1504) to the third, Joan the Mad, and her husband, Philip I (the Handsome)

of Castile, ruler of the Burgundian Netherlands. The Netherlands nobility were delighted to see this enormous accretion of power to their ruler and looked forward to the advantages that they might reap from it. They accompanied him to Castile, where a large section of the high nobility, in their turn, were anxious to acclaim him rather than the redoubtable Ferdinand. Ferdinand was therefore forced to recognize Philip's claims; but this latter died in 1506, and Ferdinand was left as uncontested ruler. His last great success was the annexation of the Spanish part of the Kingdom of Navarre in 1512.

Spain and the New World. While the exploration of the Atlantic coast of Africa had been mainly a Portuguese concern in the 15th century, the Castilians had not been entirely disinterested in such activities and had occupied the Canary Islands (off northwest Africa). In the Treaty of Alcáçovas (1479), when Afonso V of Portugal renounced his claims to the crown of Castile, he also recognized Castilian possession of the Canaries in return for Spanish recognition of Portuguese possession of the Azores (in the Atlantic Ocean west of Portugal), the Cape Verde Islands (off West Africa), and Madeira (north of the Canaries). The conquest of Granada allowed Castile, for the first time, to concentrate major resources and effort on overseas exploration. The support that Columbus received from Isabella was indicative of this new policy. In 1492 Columbus discovered the West Indies, and, in the next half century, the Spaniards conquered huge empires in America and made their first settlements in the Far East. From the beginning there were disputes with the Portuguese, who were conquering their own colonial empire. The Catholic Kings obtained a series of papal bulls (1493) from the Spanish pope Alexander VI and as a result concluded the Treaty of Tordesillas with Portugal (1494) to settle their respective claims. Everything west of an imaginary line 370 leagues (here, the league was just over three nautical miles) to the west of the Cape Verde Islands in the Atlantic was assigned to Spain; everything east went to Portugal. The rest of Europe saw no reason to accept the Pope's decision, and the result was constant and brutal warfare in the overseas colonies, even when the European governments were officially at peace (see also LATIN AMERICA AND THE CARIBBEAN, COLONIAL).

Growing Castilian activity

Colonial policy. Unlike the other European colonists of that age, the Spaniards were vitally concerned with the moral problems of conquest, conversion, and government of heathen peoples. If the great majority of conquistadores ruthlessly pursued gold, power, and status, they took with them Dominican and Franciscan friars who set themselves to convert and educate the American Indians and, sometimes, to protect them from their Spanish masters. The Dominican Bartolomé de Las Casas fought long battles to modify at least the greatest evils of colonial exploitation. His debates with a theologian, Juan Ginés de Sepúlveda, and the writings of Francisco de Vitoria provide the first systematic discussions of the moral and legal problems of conquest and colonial rule. Their importance lay in their effects on Spanish colonial legislation. The Leyes Nuevas ("New Laws of the Indies") of 1542 were based largely on the arguments of Las Casas. While in the colonies these laws were honoured more in the breach than in the observance, yet they provided at least some protection for the Indians, and there was nothing like them in any of the other European overseas colonies of the period. Even Las Casas, however, supported the transatlantic slave trade of black Africans, until, late in his career, he came to recognize its evils.

The Atlantic trade. The crown insisted that all trade with the colonies should be carried on through Seville and should be reserved for Castilians on the argument that it was Castilian money and blood that had built the Spanish overseas empire. This trade was closely regulated by the Casa de Contratación (1503), or the House of Trade, in Seville. The city itself rapidly became one of the greatest trading centres in Europe, and its population rose from 25,000 in 1517 to 90,000 in 1594. Yet Castile was unable to supply all the manufactures that the colonists demanded and for which they paid in solid gold and silver. So far from seeing this trade as an opportunity for Castilian industry, the Cortes actually petitioned the crown in 1548 to prohibit exports to the Indies that, they claimed, were raising prices in Castile. The government did not accept this petition, but Castile had to import much of what its colonists needed from Italy and the Netherlands. The Castilian monopoly of trade with Spanish America in practice had only the effect of giving the rest of Europe the chance to compete on equal terms for the American trade with the monarchy's non-Castilian subjects. The organization and financing of the Spanish Atlantic trade were largely in the hands of Genoese and south German merchants. From the 1540s, when a new method of extracting silver from ore with the use of mercury was discovered, silver mining became a major industry in both Mexico and Peru, and silver shipments in rapidly increasing quantities soon surpassed the earlier gold shipments in value. Precious metal was exported from the Indies to Spain, partly as the crown's right of one-fifth (*quinto real*), but, more important, as payment for imports. The average annual quantities, as registered by the Casa de Contratación (not counting the unknown quantities that were smuggled), rose rapidly from about 1,000,000 pesos in the five-year period 1526–30, to 5,000,000 during 1541–54, and then to the peak of more than 35,000,000 during 1591–95. The growth of overall trade between Seville and the New World followed a very similar pattern, rising until 1550, then stagnating until the early 1560s, and rising again to a peak in the last decade of the 16th century. Prices, especially of agricultural produce, had started to rise in Spain, as in the rest of Europe, long before American silver was imported in considerable quantities. But, whatever the ultimate causes of the price revolution of the 16th century, there can be little doubt that American silver greatly aggravated the inflation in Spain in the second half of the 16th century. The theologians of the university of Salamanca (some 100 miles [160 kilometres] west–northwest to Madrid) in the 1550s were the first to see this connection and to formulate the earliest version of a quantity theory of money (in which money is worth more when scarce than when abundant). Very little of this American treasure seems to have been invested in economic production. Most of it was used for display by the court and ruling circles, to pay for Spanish imports, for the Spanish armies abroad, and to satisfy the government's German, Italian, and Netherlandish creditors. Thus Spain, with all the treasure of the New World at its command, remained a poor country.

Castile's demand for monopoly

The influx of silver from the New World

SPAIN UNDER THE HABSBURGS

Charles I. Ferdinand died on January 23, 1516, and the crowns of the Spanish kingdoms devolved on his grandson, Charles I (1516–56), the ruler of the Netherlands and heir to the Habsburg dominions in Austria and south Germany. This new union had not been planned in Spain, and, at first, it was deeply resented. Cardinal Jiménez, the regent until Charles's arrival in Spain, had to battle against the old antagonisms between nobles and towns that were flaring up again when the magnates took the opportunity of the regency to try to regain their old power. When Jiménez tried to raise a militia, nobles and cities both sabotaged the plan. The old hostilities between the different Spanish kingdoms were as bitter as ever, with the men of Navarre, for instance, claiming that they would rather accept a Turk than an Aragonese as governor of the fortress of Pamplona. Although the court at Brussels had been careful to hold its hand in the distribution of patronage, the Spaniards, nevertheless, accused the Netherlanders of greed and place hunting. It took Charles's Netherlandish ministers a year and a half to settle the Netherlandish government and to make agreements with France and England that would allow the boy-king to take possession of his new kingdom without outside interference. It was a considerable achievement, but for Spain the time was still too long. When Charles arrived in Spain, in September 1517, his supporters were already disillusioned, and the country was apprehensive of the rule of a foreigner. Charles himself, ugly, inex-

perienced, speaking no Spanish, and surrounded by Burgundian councillors and courtiers, did not initially make a good impression. The different Cortes of Castile, Aragon, and Catalonia granted his financial demands but attached to them much pointed advice and criticism.

The comunero movement. On June 28, 1519, Charles was elected Holy Roman emperor as Charles V and prepared to go to Germany. His chancellor, Mercurino Gattinara, summoned the Castilian Cortes to Santiago in northwest Spain (April 1520) to demand more money, even though the former grant had not yet expired. The towns immediately made difficulties. The Toledans refused to appear; the others demanded the discussion of grievances before supply. By a mixture of bribery and concessions, the government finally induced a majority of the delegates (who had transferred from Santiago to La Coruña on the northwest coast of Spain) to vote the new grant. Many of the delegates were immediately disowned in their home towns, and one from Segovia was murdered by an enraged mob. As Charles set sail (May 20, 1520) the Castilian revolution had already begun.

Insurrection against Charles V

The towns, led by Toledo, formed a league and set up a revolutionary government. They claimed—more boldly even than the Third Estate during the French Revolution in 1789—that they were the kingdom and that the Cortes had the right to assemble without a royal summons and to discuss all matters relating to the welfare of the realm. There was talk of dethroning Charles in favour of his mother, Joan the Mad. The *comunero* leader, Juan de Padilla, actually captured the castle of Tordesillas (100 miles northwest of Madrid), where Joan was kept as prisoner; but the Queen, whether out of madness or calculation of the interests of the monarchy, would not commit herself to Padilla's proposals. The *comunero* movement spread rapidly through Castile, and the nobles did nothing to check it. They had not forgiven Charles his quest for the imperial title (which they thought inferior to that of king of Castile) nor his foreign councillors and courtiers. They resented above all his bestowal of the archbishopric of Toledo on a young Burgundian, Guillaume de Croy, and the appointment of his former tutor, Adrian of Utrecht (later Pope Adrian VI), as regent of Castile. Even the appointment of the admiral Fadrique Enríquez and the constable of Castile, Iñigo de Velasco, as Adrian's co-regents did little to mollify the offended grandees. Only when the more radical and popular elements in the cities were gaining control of the *comunero* movement and beginning to spread it to the nobles' estates did the nobles combine to raise an army and defeat the *comunero* forces at Villalar (April 23, 1521).

The power of monarchy was thus restored in Castile, never to be seriously shaken again under the Habsburg kings. The towns kept much of their autonomy, but royal control in the town councils was safeguarded by the appointment of royal officials, the corregidores. The Cortes continued to function; but the deputies of the 18 "royal towns," which alone were summoned, were mostly members of the urban nobility, exempt from direct taxation and therefore not unduly concerned about voting the taxes the government asked for. As to their petitions, the King granted or ignored them as he saw fit. The nobles had won the civil war, but they could no longer break their alliance with the monarchy. The monarchy, now that it could manage the Cortes and afford a standing army, was the stronger partner. But when, in 1538, Charles I proposed a tax from which the nobles should not be exempt, there were immediate rumblings of revolt. Charles had to give way, but he never summoned the nobility again to the meetings of the Cortes. The monarchy had thus won its political victory in Castile only at the cost of letting the nobility contract out of the financial obligations of the state and the empire. The rising burden of taxes fell, therefore, on those least able to bear them and on the only classes whose activities and investments could have developed the Castilian economy.

The crown's expensive political triumph

The nobility. The traditions of the grandees and hidalgos, formed in the centuries of struggle against the Moors, made them even more averse to economic activities than the rest of the European nobility. They invested their money in land, without, however, improving agriculture, and preferred careers in the army, the church, and the civil service to the ignoble occupations of commerce. In the long run, the economic weakness of Spain, aggravated by its social traditions and its system of taxation, proved a serious handicap in Spain's struggle with its western European rivals. After Villalar, however, the Spanish nobility had come to accept Charles I. His championship of Catholic Christianity against the Muslim Turks and German heretics appealed to their own traditions of Christian warfare against the Moors. While Charles kept the grandees out of the central government of Spain itself, he had many prizes to offer in military commands, provincial governorships, and even viceroyalties in Italy and Spanish America. The hidalgos, trained as lawyers at Salamanca or as theologians at Alcalá de Henares (just east of Madrid), could look forward to dazzling careers in the King's councils and in the Spanish church. Even though Charles spent only 16 of the 40 years of his reign in Spain, the Spanish upper classes were beginning to accept and enjoy their monarch's position as the greatest ruler in Europe.

Charles I's foreign policy. Spain as a result, however, became involved in interminable wars. The necessity of defending southern Italy from the Turks brought Charles's empire into collision with the Ottoman Empire, with the central Mediterranean as the chief battle ground. Ferdinand's failure to complete the conquest of North Africa now brought a bitter revenge. The corsair leader Khayr ad-Dīn, known as Barbarossa, had made himself master of Algiers (1529) and acknowledged the suzerainty of the Sultan of Constantinople. Thus, the purely local problem of the Moorish raids on the Spanish south coast became merged into the much more formidable struggle with the Ottoman Empire. In 1535 Charles captured Tunis. It was, perhaps, his most satisfying triumph; for in this enterprise he manifestly appeared in his chosen role of "God's standard bearer," as he said himself. He now seriously considered carrying the war into the eastern Mediterranean, even conquering Constantinople itself. But in 1538 Barbarossa with a Turkish fleet defeated Charles's Genoese admiral, Andrea Doria, at Préveza (west Greece), and in 1541 the Emperor himself failed against Algiers. At the end of the reign, the balance of the two great naval powers in the Mediterranean, the Spanish and the Turkish, was still even.

Problems with the Turkish Empire

Rival Aragonese and Angevin (cousins of the ruling French house) claims to Naples also brought conflict with the French kings, against whom Charles fought four wars. His armies conquered Milan (north Italy) and reduced most of the still independent Italian states to Spanish satellites. An increasing part of the burden of these wars fell on Spain, and especially on Castile. The Spanish *tercios* (infantry regiments) were not only the Emperor's best troops, but it was in Castile that he could raise the largest part of his imperial revenues, and this, moreover, without having to account for the way he spent them, as he had to do with the taxes voted by the States General of the Netherlands. It is therefore not surprising that the empire in Europe with Charles V as head became gradually transformed into a Spanish, or, rather, Castilian, empire of Charles I. In the latter part of his reign, Spaniards and Hispanicized Italians monopolized all high positions in the empire south of the Alps and began to appear in Germany and the Netherlands. More and more they came to interpret the international and Catholic ideals of the Emperor in terms of the political predominance of Spain in Europe and overseas.

Philip II. When Charles abdicated his various lands (1555–56), Philip II (1556–98) succeeded to all his father's dominions except Germany. His empire in Europe, now without the imperial title, was still only a loose union of independent states recognizing the same head. Philip, a great traditionalist, was not the man to inspire his different subjects with a new unifying idea, though he improved the central administration of his empire by the creation of the Council of Italy (1558). But his own Castilian upbringing and preferences increased the tendency toward transforming the empire into a Castilian

empire. Six of the nine viceroys Philip appointed to govern Sicily were Spaniards, as were all those of Naples, with the single exception of one Cardinal Granvelle, and ten out of 13 governors of Milan. In the Spanish viceroyalties of Aragon, Catalonia, Valencia, and Navarre and in those of Mexico and Peru, none but Spaniards, and preferably Castilians, were ever thought of at all. These were the key figures in Philip II's empire, and they were backed by the commanders of the Spanish regiments. Fortresses were nearly always governed by Castilians.

Role of the Castilians in the armies

Only in the Low Countries was it necessary to appoint natives to military commands. When the viceroys and governors were appointed, they were given "secret" instructions—in short, ones not meant for purely propagandistic purposes. These instructions reflected the current commonplaces of Christian government that could be found in scores of "Mirrors of Princes" (handbooks of government popular at the time) published in the 16th century and that Philip had made his own. The governors were to represent the King—not the state or the Spanish Empire—as if he were present in person; it was stressed that they were not appointed for their own benefit but for that of the community they were sent to govern; they were to watch so that the King's subjects might sleep in peace and quiet and to dispense equal justice to rich and poor. Many of the Castilian grandees who were appointed to these high offices undoubtedly strove to live up to these precepts. In practice, however, their success depended largely on the strength of the local opposition they met: this was a great deal in Sicily, which had won the reputation of being "fatal to its viceroys," but very much less in Naples, about which at least one viceroy remarked that no one should wish to be viceroy there because of the pain he would have to suffer when he had to leave that post at the end of his term. A great part of the viceroys' difficulties, however, stemmed from the unreliability of the King himself. Philip was always anxious to maintain the dignity of their office, but he encouraged the local ministers and officials to report on their viceroys behind their backs, and he had no compunction about recalling a viceroy, governor, or minister when it suited him in this way to appease local opposition.

The King kept control over his viceroys and governors by weekly, sometimes daily, correspondence, carried by the excellent postal service that the House of Austria had organized in Europe. All important political decisions were thus taken in Madrid, and there the King relied almost entirely on Spaniards for advice. Only one non-Spaniard, Cardinal Granvelle of Franche-Comté, was ever summoned to Madrid to play a leading role in the King's inner councils (1579–86). It was Granvelle who had earlier, as the King's chief minister in the Netherlands, reminded his master of the international character of his empire. He advised a more international dispensation of royal patronage, as, for instance, the appointment of the Prince of Orange (William I the Silent) to the viceroyalty of Sicily so that Netherlanders and Italians would no longer think that the King regarded only the Spaniards as his "legitimate subjects." But Philip had refused to listen, and the bitter Castilian hostility at court to Granvelle ended by making the Cardinal's ministry in Madrid less and less effective in the last two or three years before his death in 1586.

Philip's refusal of non-Castilian appointments

Finance and imperial policy. Philip II on his accession inherited an unfinished war with France and a debt of some 20,000,000 ducats. While his ally England (to whose queen, Mary Tudor, Philip was married) lost Calais, Philip's own armies won considerable victories, and he was able to conclude the Peace of Cateau-Cambrésis (1559) with France, which confirmed Spanish possessions and hegemony in Italy and which left the frontiers of the Netherlands intact. But the financial position had become irretrievable, and Philip's governments, both in Madrid and in Brussels, had to declare a moratorium on their debts. It was the first of three such moratoriums in Philip II's reign—the other two were declared in 1575 and 1596—and it set the tone for the remainder of Habsburg rule in Spain, marked by growing disparity between the monarchy's imperial policies and the financial resources at its disposal to carry out these policies. For the rest of the 16th century this disparity was still largely masked by the fluctuating, but generally increasing, shipments of silver from the New World. These shipments inspired both the King and his German and Genoese creditors with the perennial hope of new treasure to pay off ever growing debts. But the armies and navies continued to swallow up more than the stream of American silver. In the end the financial burden of empire fell more and more on Castile, and it was this fact that came to determine the course of Spanish history for the next 100 years.

When Philip II returned to Spain, in 1559, he still had a naval war with the Turks on his hands, and, in the following year, his galleys suffered a humiliating and costly defeat at the island of Jarbah (off Tunisia's east coast). In 1566 the steadily deepening crisis of the Netherlands came to a head in an outbreak of image breaking. In that year Sultan Süleyman the Magnificent died, and, for a time, the Turkish danger faded into the background. Philip could therefore risk sending his commander, the Duke of Alba, with his best Spanish and Italian troops to the Netherlands (1567) in order to settle the problems of that dominion once and for all. Alba was to root out heresy, punish those responsible for the rebellion, and impose taxes sufficient to relieve Castile of the need to send any more financial help to the government in Brussels. It was the King's most terrible miscalculation, for rebellion now became revolt and involved Spain in an Eighty Years' War, 500 miles (800 kilometres) from its own borders (1568–1648). It was in the pursuit of this war that the Spanish Empire in Europe eventually foundered.

Repression in the Netherlands

The key to Philip II's and his successors' strategic thinking, however, was always France. This was reasonable; for France was potentially the strongest military power in Europe, and its hostility to Spanish greatness was absolute, despite occasional short periods of rapprochement. But, until 1595, France was paralyzed by a long succession of civil wars. Much as Philip II hated and feared a possible Huguenot (French Protestant) victory in France, he was content to see the civil wars continue, anxious, most often, to intervene on the side of the Catholics yet sometimes covertly offering help to the Huguenots. Until the late 1570s, the threat from the Turks rivalled in importance the problems of the Netherlands. Philip switched his limited resources from the Low Countries to the Mediterranean and back again, unable to achieve a decisive victory in either theatre. It was natural, therefore, that for 20 years after the Peace of Cateau-Cambrésis Spanish foreign policy remained on the defensive. There were, moreover, still formidable internal Iberian problems to be solved.

The Moriscos. The most immediate problem was that of the Moriscos of Granada. The attempt to Christianize and assimilate them had proceeded only very slowly. In the 1560s, the ineptitude and the mutual wrangling of the different public authorities in Andalusia brought government to a virtual standstill. The Captain General of Granada, in charge of defense and internal security, was quarrelling with the municipal council of Granada and with the *audiencia*, the supreme court for Andalusia, over precedence, rights of jurisdiction, and the ownership of some pastures. The *audiencia*, in its turn, quarrelled with the Inquisition over disputed rights of jurisdiction, as did the Captain General. He was supported by the Archbishop of Granada, who was, however, involved in a lawsuit with his cathedral chapter. Such disputes were typical of the Spanish system of government, and it was also characteristic that they became immediately involved in faction fights at Philip II's court. These were therefore rarely settled according to their merits but according to the prevailing political alignment at court. In this case the Governor General, who had usually acted as the protector of the Moriscos against the exploitation of the Christians, lost. The government in Madrid first sent a commission to inquire into titles of land, and this commission confiscated mainly Morisco land. In 1567 a decree was published forbidding the Moriscos the use of

their Moorish names, dresses, and even of their Arabic language. Internal security was transferred from the governor general to the *audiencia*. This was the last straw, for now there was no one to protect the peaceful Morisco farmers from the murderous Morisco bandits of the sierra ("mountains"). On Christmas Day, 1568, they rose against the hated Christians. It took two years of ferocious campaigning, with dreadful atrocities committed by both sides, before the rebellion was put down. The Moriscos of Granada were then deported in small groups to different parts of Castile and settled in a last attempt to achieve assimilation. In the absence of systematic education and in the face of the hostility of the Christian population, this attempt was also doomed to failure.

Portugal and Aragon. There remained the question of the complete unification of the Iberian Peninsula. In the case of Portugal, Philip's opportunity came when his nephew, King Sebastian of Portugal, lost his life and a great Portuguese army in an ill-prepared crusade in Morocco at the Battle of the Three Kings in north Morocco (1578). During the short reign of Sebastian's old uncle, King Henry (1578–80), Philip carefully prepared his ground in Portugal by intrigue and bribery. Nevertheless, when Henry died the opposition to Castile was still so strong in Portugal and the attitude of France and England so threatening that it was necessary for Philip to send the Duke of Alba with an army to conquer Portugal in 1580. Philip respected the laws and privileges of his new subjects and left them to administer their own colonial empire. But the union increased rather than diminished the old hostility between the Castilians and the Portuguese.

Philip II's last action in the peninsula was against Aragon. It was precipitated by a court intrigue that led to the flight (1590) of the King's secretary, Antonio Pérez, to Aragon. Since Pérez was unlikely to be convicted in the *justicia's* court there, the King demanded his transfer to the court of the Inquisition. The populace of Saragossa (some 160 miles [260 kilometres] west of Barcelona) rioted, freed Pérez, and killed the King's special representative (1591). To the Aragonese it meant the defense of their liberties; to Philip it meant open rebellion. A Castilian army marched into Aragon (1591), and Philip made a number of constitutional changes. The *justicia* was, from now on, removable at royal pleasure; the viceroy could be a Castilian, and majority voting was substituted for the principle of unanimity in the Castilian Cortes. These changes gave the crown the ultimate power of decision in Aragon but preserved the kingdom's autonomy.

Aragonese liberties defended

Lepanto. In the Mediterranean the Spanish fleet was inferior to that of the Turks, and Philip had to remain on the defensive, even when the Turks were besieging Malta (1565). Their failure to capture the island from the Knights of St. John, who had leased it from Charles V as emperor, marked the end of their great offensive. Six years later, the combined Spanish, Venetian, and papal fleets—in alliance the numerical equals of the Turks—virtually annihilated the Turkish fleet at Lepanto (1571). The strategic effects of this great victory were negligible, but its moral effects were immense. It confirmed the Spaniards in their chosen role of champions of Christendom and explains much of their continued willingness to support their King's religious and imperial policies, even in the face of ruinous costs and mounting disasters. After Lepanto, however, it became clear that the stalemate in the Mediterranean could not be broken. In 1580 Spain signed a truce with the Porte (Turkish government).

The Armada. From about 1580 the Spanish government became convinced that the rebellion (1568–1609) and heresy in the Netherlands could not be crushed as long as the rebels received help from England and France. These countries, moreover, gave active support to the Portuguese pretender, Antonio, prior of Crato (mid-Portugal), and their privateers committed continual acts of piracy against Spanish trade in the Americas. Philip began to give financial aid to the League, the ultra-Catholic party in France. From 1586 he prepared the invasion of England. The Armada, which set sail from Lisbon in May 1588, numbered 130 ships and nearly 30,000 men, bravely and not at all incompetently commanded by the Duke of Medina-Sidonia. But it had been set an impossible task: that of convoying the army under Alessandro Farnese (*q.v.*), Duque de Parma, from the Netherlands to England in the face of a better armed English fleet and without control of a single deepwater channel port. The defeat of the Armada was inevitable but not dishonourable.

Spanish intervention in France from 1590 was equally doomed to failure. The Duke of Parma, with his Spanish veterans, won great tactical victories, but Spain failed to prevent the succession of Henry of Navarre as Henry IV of France and the collapse of its ally, the League, when Henry became Catholic (1593).

Philip saw his role and that of Spain essentially as that of defender of the Catholic Church against the aggression of the heretics, an aggression that now seemed to have become mainly military and that therefore had to be met by military force. It was therefore essential that the King should safeguard and extend the power of Spain and the just claims of his house, such as those he claimed for his daughter for the throne of France. Every other consideration was subordinated to this obligation, even to the point at which the Spanish ambassador in Rome, the Count of Olivares, intervened in three successive conclaves in order to assure, by a mixture of promises and threats, the election of popes congenial to his master (conclaves of Urban VII, Gregory XIV, and Innocent IX, 1590–91). He just failed in the fourth, but crucial, election of Clement VIII, who was to receive Henry IV back into the Catholic Church (1595).

But, if Philip II could thus justify his aggressive policies to himself, both Spain's enemies and its allies were convinced that they were witnessing the quest for Spanish dominance over Europe. Many Spaniards themselves believed this; but, as the war dragged on and the costs mounted, even the faithful Castilian Cortes began to question the King's policy. After 1580, silver shipments from the New World to Seville reached new record levels, and this undoubtedly helped to persuade Philip II to embark on his grandiose schemes against England and France. Yet they only represented a quarter of his annual revenues. The rest was derived from taxation and from loans for which future revenues were pledged. The Armada campaign was said to have cost 10,000,000 ducats. The combined cost of the continuing naval war against England, the campaigns in the Netherlands, and the military intervention in France was even greater. In 1590 the Cortes accepted the royal demand for a new excise tax that was to raise 8,000,000 ducats in six years and that was appropriately nicknamed the *millones*. But by 1595, a deputy from Seville said bitterly that "the reason why taxes have been raised without noise is because they have not fallen on the rich who are those who have a voice . . . and the sweetness which they find, that is the blood of the poor." A year later, in 1596, Philip II's government declared its third bankruptcy (moratorium), but the *millones* were increased rather than lowered.

Spain had gambled its own prosperity and its American treasure on a decisive victory over the heretics in western Europe and with it its own hegemony over the Continent, and it had failed. Shortly before his death, Philip II concluded the Treaty of Vervins (1598) with France, which substantially re-established the position of 1559. Yet, if Spain failed in its highest ambitions, at the end of the 16th century it was still the greatest power in Europe. It had brought Christianity to millions overseas—and to most contemporaries, if they thought about it at all, it seemed worth the appalling price paid for it in terms of the lives and freedom of non-European peoples—and Protestantism, though not destroyed, had been contained. Spanish monks and mystics had given Catholicism a new content, and Spanish theologians and jurists had created the basis of international law. Spanish literature and art were only now entering their greatest period. Morally and economically there were dark sides to the picture, but to the Spaniards the 16th and early 17th centuries have always been their "Golden Century."

Philip's admission of defeat

Spain in 1600. It is not surprising that the enormous exertions of the last quarter of the 16th century, with its mixture of triumphs, disappointments, and miseries, should have been followed by a general mood of introspection and even disenchantment. This was particularly evident in economic and social thinking. The *arbitristas* (literally, "projectors") were writers who combined an economic analysis of the social ills of Spain with projects for economic recovery and social and moral regeneration. They saw clearly the central weakness of Spain, the attitude of mind that despised productive work and those who engaged in it. Far too many strove to live the life of a hidalgo. The treasures of Mexico and Peru, so far from stimulating investment and industrial production, had only encouraged men to look for shortcuts to riches and to live the life of *rentiers*, investing their money in the *censos*, the government annuities. These *censos* were the greatest plague and perdition of Spain, wrote González de Cellorigo, perhaps the most acute of the *arbitristas* of 1600. "It seems," he concluded, "as if we had wanted to turn these kingdoms into a republic of enchanted men, living outside the natural order."

The positive plans (among many fantasies) advocated by the *arbitristas* included the drastic cutting of government expenditure, the reform of the tax system, the encouragement of immigration into Castile, systematic and extensive irrigation, protection of industry, improvement of transport, and, finally, the sharing of the cost of empire among the constituent kingdoms of the monarchy. These were reasonable proposals, not unlike those put forward by mercantilist writers in the rest of Europe who treated economic activity as a means of increasing the power of the state. But it was not clear that the Castilian ruling classes would be capable and willing to act on them. Its attitudes were varied and often ambivalent, and this ambivalence is also reflected in the imaginative literature of the period.

Proposals for reform

Spain's Golden Age in literature. At one extreme there was the picaresque novel with its implicit satire of a society in which one could make one's way by cleverness and roguery rather than by honest work—that is, if one did not happen to be born a nobleman. Thus, the hidalgo in the *Lazarillo de Tormes* (published 1554), the first of the picaresque novels, is down and out but would rather starve than work, and he expects his servant, the boy Lazarillo, to scrounge for them both. In *Don Quixote* (published 1605 and 1615), the author, Cervantes, raised the picaresque novel to a completely new level of social and psychological insight. It is, among other things, a parable of Cellorigo's "republic of enchanted men," living in a world of illusions and tilting at windmills.

At the other extreme, there was the drama from exponents such as Lope de Vega to Tirso de Molina and Calderón. As with the picaresque novel, the comedy of the "Golden Century" was concerned with the contemporary social scene. The psychological problems faced by its characters arose nearly always directly out of social conflicts. But the social purpose of these plays was essentially conservative: the defense of the highly structured Spanish society of the time. This was achieved by insisting for all social ranks, from the king down to the peasants, on the special dignity and honour of their status. Thus Lope, for the first time, introduced the common people as fully rounded characters on the stage, allowing, for instance, to the daughter of a blacksmith the emotions of love formerly reserved on the stage to aristocratic ladies. Heredity and blood are the principles of a social order that in the comedies may be threatened but that are always reaffirmed in the end. Here may, perhaps, be seen a link with the visual arts of the age and with the Baroque style.

The "Golden Century" in architecture and painting. In the second half of the 16th century, in the more severe artistic climate of the Catholic Counter-Reformation, the playful Plateresque style of buildings fell into disfavour. Philip II preferred the unornamented and monumental classical architecture of Juan de Herrera, the greatest Spanish architect of the century. This could be very effective, as in Philip's monastic palace of El Escorial (20 miles [30 kilometres] northwest of Madrid), which embodied the gloomy and ascetic spirit of the King and also blended with the stark and forbidding landscape of the Sierra de Guadarrama (a range of mountains northwest of Madrid). But only too often the style produced an ugly and pompous monumentality, such as not infrequently afflicts the architecture of countries at the height of their imperial careers. But at the turn of the century this style gave way to the Baroque, which Spanish architects found no difficulty in acclimatizing in their country from Italy. This style could achieve monumentality without being pompous. It could display the grandeur of the church, or the monarchy, or anyone rich enough to build himself a palace. Perhaps most important of all, it was a style that, by its love of ornamentation and its essentially theatrical character, became immensely popular with the mass of the population.

Success of the Baroque style

The painting of the period does not lend itself as easily to a social interpretation; but certain patterns may still be observed. The greatest painter of the Spanish Counter-Reformation, the Cretan El Greco, made his home in Toledo, where the local aristocratic and ecclesiastical society (but not Philip II) seems to have fully appreciated his genius. El Greco's superb portraits, but above all his religious paintings, with their elongated figures rising like so many flames to heaven, seem like the embodiment of the most spiritual aspects of Spanish Catholicism.

El Greco left no school. The painters of the following generation, especially Murillo, had a different religious sensibility, more naturalistic than El Greco, more personal, and more romantic; and again, as did Baroque architecture, they appealed successfully to a popular religiosity. But the greatest of them, Velázquez, was hardly a religious painter at all. His subjects were the King with his family and court and, characteristic for the age and parallel with the drama of the period, the common people. All these he represented with a realism and psychological insight unmatched in the 17th century, except perhaps by Rembrandt.

The reign of Philip III. It was the tragedy of Spain that its ruling classes failed to respond to the social and political problems of the age as creatively as its writers and artists. For this failure there are at least some good reasons. In the first place, the system of royal government, as it was understood at the time, depended ultimately on the king's ability to lead and to make decisions. Philip II's very consciousness of his divinely imposed obligations in this respect, compounded by his almost pathological suspiciousness of the intentions and ambitions of other men, had led him to deprecate independent initiative by his ministers. Philip thus failed to educate an effective ruling class with a tradition of statesmanlike thinking and decision making.

Philip III (died 1621) was a devout, phlegmatic nonentity, totally incapable of carrying on his father's methods of personal government. He therefore had to have a minister who would do all his work for him, a *privado*. His choice, the Duke of Lerma, however, turned out to be a singularly unfortunate one. Amiable, incompetent, and, inevitably, under heavy attack from those who envied his position, Lerma strove to maintain himself by the lavish dispensation of royal patronage to the high nobility. He was not the man to translate the schemes of the *arbitristas* into effective reforms. During the reign of Philip III the government of Spain either became the victim of events that it did not attempt to control or allowed its hand to be forced by outsiders.

Lerma's incompetency

Not all events could have been controlled. In 1599–1600 an epidemic plague claimed some 500,000 victims in Castile. This sudden decimation of the labour force caused a sharp rise in wages that acted as another disincentive to capital investment by Spaniards. Yet the advantages that the labourers had reaped from the rise in wages were quickly offset by renewed inflation. This was the result of the government's decision to solve its perennial financial problems by the massive coinage of the *vellón*, a debased copper coinage. This did not prevent the need for another moratorium on government debts, in 1607, and in 1608

the King had to promise the Cortes of Castile not to issue any more *vellones* for 20 years. But by 1617 he had to ask the Cortes to allow another issue, and yet another one in 1621.

The expulsion of the Moriscos. The most serious social crisis of the reign was, however, entirely man-made. The great majority of the Moriscos lived in the Kingdom of Valencia. Like those of Andalusia, they had been forcibly but ineffectively converted to Christianity. Most of them were relatively poor farmers, agricultural labourers, or small tradesmen and hucksters. They were hated and despised by the poor Christian peasants but protected by the landowners for whom they provided industrious tenants and labourers.

For many years a controversy raged between those who wanted to "solve" the Morisco problem by expulsion and those who pleaded for time and money to achieve their genuine assimilation and Christianization. While the practical, economic aspects of these two views were not neglected, it was characteristic of the Spain of the period that the main emphasis of the debate was on the religious and moral problems involved. In 1609 Lerma's government ordered the expulsion of the Moriscos. By 1614 some 275,000 had been forced to leave Spain. The majority of Spaniards undoubtedly approved of the expulsion. About its economic effects there was at the time and there is still today considerable controversy. In Castile the effects were probably slight. In Aragon and Valencia, where the Moriscos had constituted between 20 and 30 percent of the population, they were certainly much greater. Some but by no means all Morisco land was resettled by "old" Christians. There was a shift from labour-intensive sugar and rice production to mulberry (silk) cultivation and viticulture (vines). The greatest difficulties were caused by the indebtedness of the Morisco peasants and the consequent losses suffered by their urban creditors. An ironic footnote to the expulsion was the plight of the Aragonese and Valencian inquisitions. Once hot for expulsion, they were now left without their major source of income, the composite fines for Moorish practices by which they had formerly milked the Morisco villages.

Effects of the expulsion

Spain and Europe. Neither Philip III nor Lerma were emotionally or intellectually capable of the fundamental reappraisal of foreign policy that Philip II's failures called for. Very few even of the *arbitristas* had seen this need sufficiently clearly. The court, the nobility, and, above all, the clergy and the King's confessors remained caught in the now hardening tradition of Spanish imperialism, simplistically interpreted as the cause of God. Once again, this attitude caused a serious misjudgment of the political forces in England: the absurd hope of placing the infanta Isabella on the English throne when Elizabeth I should die. In 1601 a small Spanish force was disembarked at Kinsale, in Ireland, to cooperate with the Irish rebels. The English army had no difficulty in forcing it to surrender.

Fortunately, the new government of James I was anxious for peace. On the Spanish side, the Treaty of London (1604), which ended 16 years of Anglo-Spanish war, was negotiated on the initiative of Philip II's son-in-law, the archduke Albert, to whom, in his last year, Philip II had handed over the nominal sovereignty of the Spanish Netherlands. It was also Albert and his Genoese general, Ambrogio Spinola, who virtually forced Madrid to negotiate with the Dutch rebels. Between 1604 and 1607, Spain sent unprecedentedly large sums to Flanders. Spinola captured Ostend (on modern Belgium's coast) and won victories in Friesland (north Holland). But it would take 300,000 ducats a month to continue the war successfully, he now wrote to Madrid. After the moratorium of 1607, Philip III was in no position to raise such sums; but still he could not bring himself to recognize the independence of his rebellious and heretical subjects. He compromised by concluding a truce for 12 years, to run from 1609.

In 1610 a new war with France threatened, but Henry IV was assassinated, and once more, for almost 20 years, Spain's most formidable opponent in Europe became preoccupied with its internal problems. The years from 1610 to 1630 were the last period in which Spain clearly dominated Europe. For the first of these two decades Europe enjoyed a kind of Pax Hispanica. Spanish armies controlled Italy, Flanders, and parts of the Rhineland. Spanish and Spanish-inclined Jesuits were confessors at the courts of the Austrian Habsburgs, of Poland, Bavaria, and of some of the minor German and Italian princes. Spanish subsidies, pensions, and bribes made clients even of Protestant politicians in England, Holland, and the Swiss cantons (although much less effectively so than Madrid hoped); and Spanish-paid spies fed the governments of Madrid and Brussels with valuable, if not always accurate, information about its potential enemies in the United Provinces (or Holland), England, and France. Yet, to a much greater degree than most contemporaries realized, this Spanish domination of Europe rested on default: the disunity and temporary weakness of Spain's political and religious opponents. The psychological effects on Spain of this position were wholly disastrous, for it confirmed the Castilian ruling classes in their imperialist attitudes.

While for Philip III and Lerma this attitude led, for reasons both of finance and temperament, to a largely defensive stance, it was quite otherwise for the Spanish representatives abroad. In the absence of an effective lead from Madrid, the Spanish grandees who were the King's viceroys and ambassadors in Europe took it upon themselves to advance Spanish interests as they saw them—that is, in terms of Spanish power. They fortified the route from Milan to the Tirol (now west Austria) through the Valtellina, the vital link with the Austrian Habsburgs; they annexed several small Italian lordships; they inveigled Dalmatian pirates (operating from what is now the Yugoslav coast), the Uskoks, to prey on the trade of Venice, and they even seem to have plotted the complete overthrow of that republic.

More fateful still were their activities in Prague and Brussels. At the courts of the emperors Rudolf II and Matthias, the ambassador Baltazar de Zúñiga organized an effective "Spanish" party. His successor, the Count of Oñate, negotiated the secret Treaty of Graz (1617) by which the Jesuit-educated archduke Ferdinand of Styria (later Emperor Ferdinand II) was designated as heir to Matthias. In return for giving up Philip III's claims to the Austrian succession, which Madrid had never seriously pursued in any case, Oñate obtained the promise of full Spanish sovereignty of the Tirol and Alsace (in modern east France), the two German pillars of the "Spanish Road" between Italy and the Netherlands. At the same time, the "Spanish" party in Prague managed the preelection of Ferdinand as king of Bohemia in case of Matthias' death. Zúñiga and Oñate had undoubtedly strengthened Spain's strategic position in central Europe; but they had also, for the first time since the abdication of Charles V, involved Spain again in the local politics of the Holy Roman Empire. For Philip IV, this involvement turned out to be even more disastrous than it had to Charles V. Spanish leadership, as practiced by the self-willed Castilian grandees abroad, had proved to be energetic and clever; but it was ultimately as devoid of true statesmanship as the slackness of the King and his *privado*.

Spain's position in central European politics

In 1617 Lerma's enemies at court finally managed to overthrow him. Zúñiga returned to Madrid and became the leading advocate of aggressive policies. The Marquis of Bedmar, former Spanish ambassador to Venice and the organizer of the anti-Venetian conspiracy, went as ambassador to Brussels and immediately began to press for the reopening of the war against the United Provinces. In 1621 Philip III died, and with him disappeared the last restraints on the neo-imperialists. Philip IV, a boy of 16, left the effective powers of kingship in the hands of his former gentleman of the chamber, the Count (later Count-Duke) of Olivares. Olivares shared his uncle Zúñiga's political views, and he soon dominated the Council of State.

Philip IV's reign. *Spain and the Thirty Years' War.* In 1620, following the defeat of Frederick V (the elector Palatine, or prince, from the Rhineland who had accept-

ed the crown of Bohemia, now western Czechoslovakia, when offered it in 1618) and the Bohemians, Spanish troops from the Netherlands entered the "Winter King's" hereditary dominions of the Rhine Palatinate. Militarily, Spain was now in a favourable position to restart the war with the United Provinces at the expiration of the truce in 1621. The decision to do so was, however, taken on more general grounds. The Dutch had used the truce only to capture the carrying trade of western Europe and the Baltic with Spain, Zúñiga argued. On the oceans they had never observed the truce but continued their piracies against Spanish and Portuguese shipping. If they were allowed to continue, first the Indies would be lost, then the rest of Flanders, Italy, and, finally, Spain itself, for it would have lost the dominions that had made it great. These were very different grounds for resuming the war from those habitually advanced by Philip II. Little was said about religion or even the King's authority, while the protection of the overseas empire had become the central consideration in Spanish relations with the Dutch "rebels." Olivares dismissed the counter-arguments of the Council of Finance. The young King, content to be told that he was not responsible for the debts of his predecessors, piously declared his intention not to burden his subjects any further. Neither he nor his ministers could foresee that a recent slump in silver shipments from America was not a temporary setback but heralded a rapid, long-term decline. The Dutch, it is true, were equally anxious for war, partly, at least, because of the vain hope that the Belgians would rebel against Spain and join the United Provinces.

Having decided on war, Olivares pursued a perfectly consistent strategy: communications between Spain and the Spanish Netherlands must be kept open at all costs, and the Dutch must be hit wherever they were most vulnerable. The first objective led Spain to cultivate the friendship of James I of England and even to contemplate the restoration of Frederick V to the Palatinate and the marriage of Philip IV's devoutly Catholic sister to the heretic Prince of Wales (later Charles I). It led to very close cooperation with the Austrian branch of the Habsburgs and the need to fight for the control of the Valtellina. The second objective, which followed the advance of the imperial armies under Wallenstein (an adventurer who made himself indispensable to the Habsburgs as a military organizer) to the Baltic, led to grandiose schemes of building an imperial Spanish fleet in the Baltic with Hanseatic (the Hanse towns on the Baltic were independent mercantile organizations) and Danish help in order to destroy the Dutch Baltic trade and with it the economic prosperity of the republic.

However rational and limited these aims and plans seemed in Spain, in the rest of Europe they appeared to show only too clearly the limitless ambitions of the House of Austria. The now habitual talk in Spanish court and military circles of restoring Spain's greatness did not help to persuade Europe otherwise. Spinola's and Wallenstein's victories in the mid-1620s convinced the Spanish Council of State that victory against the Dutch was possible and blinded them to the danger of raising up new and more powerful enemies. Thus, they let slip the last chances of a favourable peace. Yet, despite enormous sums sent annually from Castile to Flanders, the Spanish armies could not break Dutch resistance. They could not even supply their own provisions and ammunition without the covert help of Dutch merchants, who, in their turn, argued that this trade with the mortal enemy brought in the money needed to pay for the troops fighting this enemy. From 1630, when Sweden and France actively intervened in the war, Spain rapidly lost the initiative. Spanish armies could still win tactical victories in Italy and Germany, but the number and seriousness of Spanish reverses, especially at sea, were now steadily mounting.

Improved government under Olivares

The government of Olivares. Olivares was undoubtedly the most able politician directing the Spanish government since Cardinal Granvelle. All contemporaries remarked on the contrast between his energetic and bustling administration and the sleepy regime of Philip III and Lerma. The Catholic Kings, the Emperor, and Philip II had kept the high nobility, to a greater or lesser degree, out of the central government. Lerma had reversed this policy, and Olivares could not go back on this position, although he bitterly lamented the incompetence of his fellow aristocrats and sharply reduced the overgenerous flow of royal patronage to them. He could and did develop the system of committees (juntas) of experts within the councils, which took over a great deal of government business and made its administration more efficient.

In 1623 and 1624 Olivares presented to the King and Council of State a number of memorandums that were nothing less than plans for a far-reaching reform of government and society, on the lines advocated by the *arbitristas*. Like them, Olivares saw the need to change mental attitudes: restraints on the aristocratic love of splendour and display, the need to appreciate the dignity of work and productive economic activity, an end to the economically harmful and morally indefensible mania for *limpieza* (Olivares himself, through his grandmother, had *converso* blood flowing in his veins). On the more immediately practical level, Olivares' memorandums were concerned principally with finance; for, with an annual expenditure of 8,000,000 ducats, there was a deficit of 4,000,000. The Count-Duke proposed the abolition of some of the most harmful taxes, the *millones* and the *alcabala*, the sales tax, and their substitution by other, simpler, and more equitable taxes. Finally, Castile should not be expected to continue to bear almost alone the cost of the war. Like Granvelle, Olivares recognized that the King's non-Castilian dominions could only be expected to share in the burdens of empire if they could also enjoy its advantages, the honours, commands, and control over policy that had been all but completely reserved to the Castilians.

None of these plans was put into practice. The Spaniards were unwilling to change their mode of life and their ingrained beliefs at the behest of a royal favourite. Olivares did manage to arrange loans with a consortium of Portuguese Marrano businessmen, but he was bitterly attacked for this action. Philip IV consulted a committee of theologians about a proposed canalization of the Tagus River, flowing east–west across Spain and Portugal. They forbade it as an interference with God's design. The court itself gaily abandoned a short-lived austerity in the celebrations that followed the arrival of the Prince of Wales in his romantic but abortive quest for a Spanish bride (1623). The financial reforms foundered on the opposition of vested interests to taxation by the Cortes and on the opposition of the whole Castilian ruling class to the plan for the decentralization of the empire. Just as had happened to Granvelle's proposals, there was not even any serious discussion of Olivares' plan. In the 1560s the result of this failure had left Philip II with no alternative but Alba's policy of repression—which caused the revolt of the Netherlands; in the 1620s it left Olivares with no alternative but his Union of Arms—which caused the revolts of Catalonia and Portugal. The Union of Arms was a scheme for the creation of a reserve army of 140,000 men that was to be paid for by the dominions of the Spanish Empire in proportion to their estimated resources. But the non-Castilian dominions disliked this proposal because it infringed their liberties. They also distrusted Castilian intentions, and with some justice; for in 1625 Olivares had advised the King in a secret memorandum to "secretly plan and work to reduce these kingdoms of which Spain is composed to the style and laws of Castile."

Hopelessness of reform

The revolt of Catalonia. Apart from Portugal, Catalonia was the state with the greatest degree of autonomy. Its medieval form of government had not been changed since Ferdinand the Catholic had settled it in 1486. Its countryside, especially on the French border, was infested with smugglers and bandits and riven by local feuds. Its taxes were administered by the Diputació, a self-perpetuating and corrupt committee of the Catalan Corts that functioned during the long intervals between the meetings of that body. The viceroys, hemmed in on all sides by local privileges and without control over the finances of

the province, were virtually powerless. In 1626 Philip IV summoned the Cortes of the realms of the crown of Aragon. Aragon and Valencia reluctantly voted some money but refused conscription of troops. Catalonia refused everything. Nevertheless, Olivares published the royal decree for the Union of Arms. Relations between Madrid and Catalonia deteriorated rapidly.

Increasing financial troubles

As the costs of warfare continued to soar, the government resumed the inflationary minting of *vellones* and had yet to declare another moratorium, in 1627. In 1628 the *vellón* coins were withdrawn, causing a collapse of prices and a business recession. The 1630s added new taxes in Castile and outright confiscations from private individuals, both of income from government annuities and of American silver imported in commercial transactions. Not surprisingly, Madrid was becoming obsessed with what it considered to be the injustice of Catalonia's immunity from taxation. In 1639 Olivares opened a campaign against southern France from Catalonia. It had no rational strategic objective except to pitchfork Catalonia into the war. If the Catalans had to defend their country, Olivares argued, they would have to support the army.

This logic was lost on the Catalans. The peasants, urged on by their clergy, refused to support the troops. During the winter the soldiers were billeted in the countryside. Soon there were clashes with the population, then riots and open rebellion. Too late, Olivares attempted to draw back and conciliate the Catalans. On June 7 the mob murdered the Viceroy in Barcelona. The higher nobility and the urban aristocracies were still anxious for an accommodation; but the countryside was now completely out of control. The Diputació, the only remaining legal authority, was led by a strong-minded cleric, Pau Claris, canon of Urgel, west of Barcelona, who was unwilling to make concessions. In the autumn of 1640 Olivares scraped together the last available troops and sent them against the Catalan rebels. Claris countered by transferring Catalan allegiance to the king of France, "as in the time of Charlemagne" (January 1641). French troops now entered Catalonia. Only when the renewed French civil wars (the Fronde) induced the French to withdraw their army were the Castilians able to reconquer Catalonia (1652). The Catalan upper classes, at least, were relieved, for they had found the French even less congenial masters than the Castilians. For once, Madrid did not repeat its previous mistakes: the liberties and privileges of Catalonia were fully restored.

The revolt of Portugal. The revolt of Catalonia gave the Portuguese their opportunity. The lower classes and the clergy had always hated the Castilians. The Portuguese aristocracy and the commercial classes, previously content with the patronage and the economic opportunities that the union with Spain had provided, had become dissatisfied during the last 20 years. They resented the introduction of Castilians into their government (1634), the ineffectiveness of Spanish naval support in the defence of Brazil against the Dutch, and the growing reaction of the Spanish colonies against Portuguese economic penetration during this period of contracting economic activity. Rather than allow themselves to be sent to fight the Catalan rebels, the Portuguese nobility decided to seize power in Lisbon and proclaimed the Duke of Bragança as King John IV of Portugal (December 1640). Madrid, with an aristocratic conspiracy in Andalusia on its hands (1641), no longer had the means to react.

Portuguese grievances

The last years of Philip IV. The disasters on Spain's periphery were matched by continued mismanagement of Spanish finances at the centre. Once more the government tampered with the *vellón* coinage and then reversed course into a sudden and catastrophic deflation (1641–42). In January 1643 the Castilian grandees were finally able to force Philip IV to dismiss Olivares. The King now decided to run his own government. He dissolved the juntas, and the councils resumed their authority. Soon, control of the government slipped into the hands of Olivares' nephew, Don Luis Méndez de Haro, a clever but colourless politician with neither his uncle's imperial vision nor his panache.

The defeats continued. The Great Condé (the French King's cousin) broke the Spanish *tercios* and their reputation for invincibility in the Battle of Rocroi, northeast France (1643). In 1647, popular revolutions broke out in Naples and Palermo (Sicily), and soon these two cities were in the hands of revolutionary governments. In Naples, at least, it was again the excessive taxation, imposed for Spain's war effort, that had precipitated the rebellion. The Spanish monarchy, wrote the Venetian ambassador to Madrid at the time, "resembled that great colossus that during an earthquake had collapsed in a few moments while everyone hurried along to enrich himself with the fragments."

In fact, Spain survived and even managed to hold on to much of its empire. The revolts of Naples and Sicily, directed as much against the local nobility as against Spain, were put down in 1648. When the Emperor conceded French claims to Alsace and the Rhine bridgeheads, the "Spanish Road" to the Netherlands was irrevocably cut, and the close alliance between the Spanish and the Austrian branch of the House of Habsburg came to an end. With Portugal in revolt and Brazil no longer an issue between the Dutch and the Spaniards, Philip IV drew the only possible conclusion from this situation and rapidly came to terms with the United Provinces, recognizing their full independence and agreeing to the closing of the Scheldt, a river emptying into the North Sea west of Antwerp (Treaty of Münster, January 1648). But Philip IV had not changed his basic policy. He wanted to have his hands free for a final effort against France, even after Catalonia had surrendered. Once again the temporary weakness of France during the Fronde confirmed the Spanish court in its disastrous military policy. Haro passed up the chance of concluding a very favourable peace in 1656.

The war dragged on, with England joining France, capturing Jamaica, and contributing to the Spanish defeat in the Battle of the Dunes on the northern coast of France (1658). The Treaty of the Pyrenees (1659) cost Spain Artois (now northernmost France), Roussillon, and part of Cerdagne. More important than these relatively minor territorial losses was the realization throughout Europe that Spain's pretensions to hegemony had definitely and irremediably failed. The Spaniards themselves were slow to admit it. Philip IV had made concessions to France in order, once again, to have his hands free against the last unforgiven enemy, Portugal. There was no longer any rational basis for his hopes of success. All schemes for financial and tax reforms were still being blocked by vested interests, and the government had declared two more bankruptcies, in 1647 and 1653. Once more the Council of Finance issued a debased coinage to pay for the Portuguese campaign. But the Portuguese routed the last Spanish armies at Ameixial (1663) and at Villaviciosa on the north coast of Spain (1665). In 1668 Spain formally recognized the independence of Portugal.

The Treaty of the Pyrenees

Charles II. For ten years Philip IV's widow, Maria Anna of Austria, acted as regent for the child Charles II (1665–1700). She allowed her government to be dominated by her confessor, the Austrian Jesuit Nithard. He was overthrown by John of Austria, an illegitimate son of Philip IV, who, however, proved quite incapable of providing effective leadership for the country. From then on, the high nobility dominated the government as effectively as it had done in the days of Philip III, playing court intrigues, first with the regent and, later, with the two successive wives of the incapable Charles II. The majority of these aristocrats were self-seeking and incompetent; but some, notably the Count of Oropesa, were men of considerable ability. They finally restored the coinage, in 1680, although not before they had caused another catastrophic deflation. They set up a committee for commerce that pursued orthodox mercantilist policies, encouraging trade and industry. They even took the unprecedented step of investigating the Inquisition and recommending a right of appeal to the secular courts.

Decline of Castile

But these were palliatives. Until the mid-1680s the Castilian economy declined, and this at such a rate that the French ambassador, the Marquis of Villars, claimed to see an appalling deterioration between his two visits, in



The European dominions of the Spanish Habsburgs in 1648.
From R. Palmer and J. Colton, *A History of the Modern World*, Copyright © 1971; Alfred A. Knopf, Inc.

1668–69 and in 1671–73. In the 1690s the Venetian ambassador characterized the reign of Charles II as "an uninterrupted series of calamities." The population of Castile declined from about 6,500,000 at the end of the 16th century to a low point of under 5,000,000 about 1680. Figures for the whole of Spain followed a similar pattern, from 8,500,000 to about 6,600,000. The reasons for this decline were not so much emigration to the overseas colonies, which averaged 4,000–5,000 a year in the 17th century, as military casualties from all causes, which averaged the frightening figure of 10,000–12,000 a year. More devastating still were the recurrent plagues and, perhaps, the sheer misery of the rural population, who lived on estates that their noble and ecclesiastical owners could not be bothered to manage with even a minimum of efficiency. The shortage of labour, especially skilled labour, and the high wages attracted many foreign workers, perhaps as many as 70,000 Frenchmen. Nevertheless, Castilian industries continued to decline. The nadir was reached in the decade 1677–86, with crop failures, earthquakes, an epidemic that sharply reversed a slight upward trend in population since about 1650, and, on top of these natural disasters, the government's deflation of the coinage.

The French wars. In these circumstances it is not surprising that Spain now became the victim rather than the initiator of aggression. In three successive wars with France (1667–68, 1672–78, 1689–97), Spain lost Franche-Comté (Treaty of Nijmegen, 1678) and some Belgian frontier towns to France but still managed to hold on to the greater part of the southern Netherlands and its Italian dominions. The reason was less Spain's own military efforts, which were puny compared with those of the first half of the century, than the unwillingness of other European powers, especially the United Provinces, to see the Spanish dominions in Europe swallowed up by France. After the last and, for Spain, most disastrous of these wars, the War of the Grand Alliance (1689–97), Louis XIV himself restored Flanders and Catalonia, which his troops had occupied, for he now had his eye on the inheritance of the whole Spanish Empire.

The last years of the childless and evidently dying Charles II were occupied by the manoeuvres of the European powers for the Spanish succession or, alternatively, for the partition of the Spanish Empire. Amid cabals, intrigues, exorcisms of evil spirits, and blood feuds at court, while riots were going on in the streets of Madrid, the rule of the House of Austria came to an end with the death of Charles II, on November 1, 1700.

The decline of Spain. There can be no doubt about the economic and political decline of Spain in the 17th century and especially in its second half. Moreover, Calderón, Valázquez, and Murillo had no successors of remotely comparable stature. It need not be reasonable, however, to speak of decadence, as many non-Spaniards did at the time and later. The term decadence, except perhaps when applied to the person of Charles II, that last unhappy offspring of constant Habsburg intermarriage, does not explain the timing of the decline nor its duration. In the first place, the economic decline was mainly a Castilian phenomenon and did not affect Catalonia and Valencia to anything near the same degree. For Castile, it is perhaps best to see the problem of decline as the *arbitristas* saw it: in the depreciation by Castilians of economic activity, an attitude that was rooted deeply in Castile's past history but that was particularly baleful in a period of general European economic depression, such as the 17th century. To this must be added the aggressive militarism that was central to the Castilian aristocratic tradition and that led to the political hubris (overweening pride) of Spanish imperial policy, from Philip II to Philip IV. The Castilian ruling classes never produced, or perhaps gave no chance to, a leader who could break out of this tradition. Velázquez seems to have known it or felt it instinctively when he painted his "The Surrender of Breda" as the beginning of a hoped-for reconciliation and when, in his portraits of Philip IV, he showed the pathos of a man half aware of his personal inadequacy for the role he is called upon to play. It was the wars that devoured Castile, though they were fought beyond its borders. They do not directly explain the end of the "Golden Century." But it may be suggested that a society that has invested most of its

energies and all its pride in war, even though it be war for an ideal, will not provide a congenial ground for the exercise of creative genius when its ideal has failed and it is left with nothing but a now hollow pride. (H.G.K.)

V. Spain since 1700

THE EARLY BOURBONS, 1700–53

The wars of the 17th century, though they had weakened Spain's power in Europe, had left it still the greatest imperial power in the world. Spain's central problem in the 17th century had been to maintain what remained of its European possessions and control of its American empire. In 1700 both tasks appeared to be beyond the military and economic resources of the monarchy. In the 17th century the greatest threat had come from a land power, France, jealous of Habsburg power in Europe; in the 18th it was to come from a sea power, England, while the Austrian Habsburgs became the main continental enemy of Spain.

The War of the Spanish Succession. In 1700 (by the will of the childless Charles II) the Duc d'Anjou, grandson of Louis XIV of France, became Philip V of Spain. Austria refused to recognize Philip, a Bourbon, and concede the defeat of its hopes of placing an Austrian candidate on the throne of Spain. To England, a Bourbon king in Spain would disrupt the balance of power in Europe in favour of French hegemony.

Louis XIV conceived of Spain under a Bourbon king as a political and commercial appendage of France to be ruled by correspondence from Versailles; he wished to regenerate and strengthen his ally by a modern centralized administration. This task was both complicated and facilitated by the War of the Spanish Succession (1701–14) when the allied armies of Britain and Austria invaded Spain in order to drive out Philip V and establish the "Austrian" candidate, the archduke Charles, on the throne. An efficient administration had to be created in order to extract resources from Spain for the war effort, and thus relieve pressure on the French treasury; at the same time financial shortages imperilled administrative reform while war taxation and war levies drove Catalonia and Aragon to revolt against the demands of the Bourbon dynasty.

Administrative reforms

The instruments of centralizing reform were French civil servants (Jean-Jacques Amelot, Louiv XIV's ambassador, and Jean-Henri-Louis Orry, a financial expert) and a handful of Spanish lawyer-administrators such as Melchor de Macanaz. They were supported by the queen, María Luisa of Savoy, and her friend the 60-year-old Princess of Ursins. The opponents of reform were those who suffered by it: the grandees who had dominated the cumbersome inefficient councils; the councils themselves; the regions like Catalonia, Aragon, and Valencia, in which the establishment of effective royal rule was seen as a Castilian, centralizing, imposition in conflict with the local privileges, or *fueros;* the church, whose position was threatened by the ferocious and doctrinaire regalism of Macanaz, who wished to subject the independent jurisdictions of the church (especially that of the papal nuncios and the Inquisition) to the absolute monarch.

The disaffection of all these elements easily turned to opposition to Philip V as king. Opposition to the new dynasty accentuated the determination of Bourbon civil servants to end special privileges that covered treasonable sympathy with the Austrian and English invaders.

Castile was ferociously loyal to the new dynasty throughout the war, in spite of severe financial difficulties (consequent on the loss of revenues from the Indies). The support of Castile and of France (until 1711) enabled Philip V to survive severe defeats and two occupations of Madrid. In 1705 the archduke Charles landed in Catalonia and took Barcelona. When Philip V tried to attack Catalonia through Aragon, the Aragonese, in the name of their *fueros,* revolted against the passage of Castilian troops.

It was this revolt, backed by the local nobility, that turned the King's advisers resolutely against local privileges and aristocratic treason. After the victory over the archduke Charles at Almansa (April 1707) the *fueros* of the kingdoms of Valencia and Aragon were abolished and the property of rebels confiscated. When the Archbishop of Valencia resisted attempts to make priests of doubtful loyalty appear before civil courts, the regalism of Macanaz was given full course.

This was the last direct triumph of the reformers. With the death of Queen María Luisa in 1714 and the arrival of Philip's new wife, Isabella (Elizabeth Farnese), court support for radical reform disappeared. Macanaz was condemned by the Inquisition, and a less rigid administration, more inclined to compromise with the church and the higher nobility, controlled the country's policy.

The last stages of the war were a Spanish concern. The allies deserted the archduke Charles; the French gave little help to Philip V. In 1714 Philip recaptured the Archduke's capital at Barcelona. By the Decree of Nueva Planta (1716) the *fueros* were abolished and Catalonia was integrated into Spain. Integration, criticized as it was by later generations of Catalans as the destruction of Catalan "nationality," was nevertheless a precondition for industrial revival; it gave Catalonia a domestic market in Spain and later an overseas market in America.

The integration of Catalonia into Spain

Paradoxically, a disastrous war had for the first time created a unitary Spanish state: except for the Basque Provinces and Navarre, Spain was under direct royal administration.

It was in its possessions outside Spain that Spain paid the price of defeat: the treaties of Maastricht and Utrecht (1713) stripped the country of its European possessions (Belgium, Luxembourg, Milan, Sardinia, and Naples) and gave Britain Gibraltar and the right to send one ship a year to trade with Spanish America.

"American" and "Italian" policies. Two tendencies can be discerned in Spanish policy until 1748: a desire for revenge and recovery in Italy, and an "Atlantic" policy; *i.e.,* the protection of America from British incursions and the revivification of Spanish colonial rule. Both policies demanded a strong army and navy. The "Italian" tendency is connected with Philip V's second wife, Isabella, and her desire to get Italian thrones for her sons. The instruments of her ambitions were two foreigners: the Italian cardinal Julio Alberoni, the exiled son of an Italian gardener, and the Dutch-born adventurer Juan Guillermo Riperdá. The attempt to recover its possessions in Italy involved Spain in an unsuccessful war with Austria, now the great power in Italy. Spain suffered a serious naval defeat at Capo Passero in 1718. Nevertheless, Isabella's persistence was rewarded when her son, the future Charles III, became the duke of Parma in 1731 and king of Naples in 1733, relinquishing his claims to Parma.

The American–Atlantic tendency was the work of Spanish ministers with a particular interest in the navy and foreign trade—José Patiño, Marqués de la Ensenada, and José de Carvajal y Lancáster. The "Italian" and "Atlantic" tendencies existed side by side in the late years of Philip V's reign. Atlantic rivalries in the form of a dispute over the interpretation of British trading privileges in Spanish America granted at Utrecht brought on the War of Jenkins' Ear (1739–43), during which the British sacked Porto Bello in the Caribbean. The Spanish fleet nevertheless was surprisingly effective and worsted Adm. Edward Vernon at Cartagena. The Italian–Mediterranean policy led directly to Spanish involvement in the War of the Austrian Succession (1740–48). It made possible an alliance of France and Spain against Austria, now the dominant power in Italy, giving Isabella the opportunity to settle her second son, Philip, in an Italian duchy. Spanish troops entered Milan (1745). Ferdinand VI (reigned 1746–59) was concerned with the domestic recovery of Spain rather than the extension of its power in Europe. He hoped to recover Gibraltar at the general peace that ended the War of the Austrian Succession. But the Anglo-French rapprochement made recovery of Gibraltar an impossibility at the Treaty of Aix-la-Chapelle (1748); the treaty merely strengthened Spain's position in Italy when Philip became duke of Parma, Piacenza, and Gustalla.

With Ferdinand VI the Atlantic tendency became dom-

inant. Since Britain was Spain's great enemy in the Americas (as Austria had been in Italy) its "natural" ally was France as Ensenada and Carvajal had seen (hence a series of family pacts with France in 1733 and 1743). It was only in the last years of his reign that Ferdinand's minister, Ricardo Wall, attempted a policy of strict neutrality as the best means of saving Spain from the hostility of Britain, Austria, or France.

This American interest was reflected in increased trade (the old system of convoyed fleets was abandoned for individual sailings), the founding of privileged trading companies (1725), and the founding of new naval arsenals at Ferrod and Cartagena (1726). But the central weakness of Spain as an imperial power remained. The Spanish economy could not supply America with the consumer goods it needed in return for its increased exports. These were supplied either by British merchants through the "legitimate" trade from Cádiz or by smuggling. The Spanish navy, in spite of considerable efforts, could not suppress a contraband trade that, from the colonists' point of view, was a necessity.

The ministers of Ferdinand VI revived the reforming traditions weakened during the latter years of Philip V, whose lethargy deepened into chronic melancholia.

One of the most important imitations from French administrative practice was the use after 1749 of crown officers, or intendants, to rule in the provinces. The intendants strengthened royal control over local government, especially in its financial aspects; together with the captains general (set up by Philip V), they were responsible for a renewal of provincial public works. Carvajal reorganized the postal system; Ensenada, at one time the all-powerful minister of Ferdinand VI, was a great road builder; he opened the Seville tobacco factory, botanical gardens, and observatories; above all, he increased royal receipts by rationalization of the tax structure. The reforms of Ferdinand's successor, Charles III, though more dramatic, were a continuation of those of Ferdinand VI.

THE REIGN OF CHARLES III, 1759–88

Two features distinguished the Caroline reforms from those of the early Bourbons. First, Charles III was a "reformer's king" in that he consistently supported reforming ministers. This was surprising in a monarch who had no great intellectual gifts, was obsessed by hunting, and turned his court into the most boring society in Europe. Second, the civil servants were distinguished from their predecessors by their adherence to a philosophy of government derived from the commonplaces of the European Enlightenment.

Penetration of Enlightenment ideas

Nevertheless, there were sharp differences among the civil servants. The Conde de Aranda and Pablo de Olavide y Jáuregui were influenced by the French Philosophes; Gaspar Melchor de Jovellanos y Ramírez was a disciple of Adam Smith; Pedro Rodríguez Campomanes drew more directly on such Spanish reformers as Macanaz; José Moñino Floridablanca was a professional administrator. All would have taken as their slogan *Felicidad* ("Felicity")—a well-ordered monarchy based directly on the productivity of people who are made happy by the intelligent application of the principles of political economy. There were, however, impediments (*estorbos*), such as traditional privileges (*e.g.*, grazing rights held by a sheep breeders' corporation, the Mesta) or attitudes (*e.g.*, the prejudice of the nobility against the "mechanical trades"); they could not be allowed to stand in the way of greater prosperity and, therefore, of a higher tax income for the state.

Economic revival. The years after 1714 had seen a gradual economic recovery, quite marked in the last half of the 18th century. How much of this revival of the last half of the century was the work of conscious but often confused government policy must remain doubtful. Thus the famous decree (1783) "ennobling" the mechanical trades had little practical result; the attempt to establish government factories and replace religious charity by the productive employment of the poor was a relative failure. Government policy had little to do with the growth of population, which rose during the century from 8,000,000 to 12,000,000. The increased demand for food and the consequent sharp rise in prices encouraged agriculture, benefitting the large landowners of the south and the small farmers near growing towns such as Barcelona. The most remarkable feature of this economic revival was the emergence in Catalonia after 1745 of a modern textile industry, based on cotton. The industry benefitted from a protected market in Spain and the colonies. In the Basque Provinces the archaic iron industry began a slow process of modernization. In Galicia, Catalan immigrants established a flourishing fishing fleet. The brandy trade brought sudden prosperity to the ports of Catalonia and their hinterland.

Nevertheless, the financial reforms of the earlier years and the abolition of internal customs duties removed obstacles to the expansion of the internal market, and the opening of the American trade (see below) acted as a strong stimulant.

One important feature of these developments was that they took place not in the old Castilian core of the monarchy but in the peripheral regions; not in the countryside but in the towns. It was on the periphery that price rises were steepest, outstripping wages in Barcelona and thus creating the beginnings of a social problem. The role of Catalans in the economic revival lies at the origins of the Castilian stereotype of the Catalans as selfish, thrusting businessmen, indifferent to traditional values and exploiting their fellow Spaniards.

Imperial problems. To Charles III the key to prosperity lay in the Indies, for a prosperous Spain demanded an American market. He saw clearly that Spain alone could not preserve an overseas market closed to the outside world against Britain. Ricardo Wall's policy of strict neutrality had allowed Britain to make gains in Canada that must weaken France—the only other anti-British power in America. If Britain would respect Spain's possessions, or if Charles could mediate between France and Britain in such a way as to preserve the balance between them in America, then to commit Spain to the French alliance was a dynastic luxury. Once it was clear to Charles that British terms were non-negotiable, then the Family Compact with France of 1761 was a piece of realpolitik. It was signed by the "anglophile" Ricardo Wall.

The Seven Years' War

The consequence of such an alliance was involvement in the Seven Years' War (*q.v.*)—too late to save France. In 1762 the British occupied Havana—the greatest single blow sustained by Spain in the war. Spain theoretically allowed no foreigners to share directly in the colonial trade: the effect of this had been to starve the colonies of necessary imports and to encourage smuggling. In 1762, 15 ships entered the port of Havana; during the 11 months of British occupation, 700. This was a dramatic indication to the colonists of the drawbacks of the Spanish monopoly, especially when that monopoly was exercised by what was, in European terms, an underdeveloped nation.

The Treaty of Paris (1763) concluding the Seven Years' War destroyed France as an American power. Spain lost the territory between Florida and the Mississippi, in return gaining Louisiana from France. Spain also had to recognize Portuguese advances in the Río de la Plata (the fort of Sacramento) and the British right to cut mahogany in Central America. The Family Compact was therefore an immediate military failure, and it was only the revolt of the American colonies against Britain that enabled Spain to recover the ground it had lost; the successful alliance with France to aid the colonists resulted in the Treaty of Versailles (1783), which gave back Sacramento, the two Floridas, and Minorca. Not until later did it become apparent that an alliance of revenge with colonial insurgents is a short-sighted policy for an imperial power.

Reform of imperial administration

The problems of imperial defense were thus temporarily solved by British weakness after 1765. The positive side of Charles III's imperial policy was an attempt to create an efficiently administered colonial empire that would provide the crown with increased revenues and with a closed market for the exports of an expanding Spanish economy.

The rationalization of the administration of the Indies had started before Charles III with the creation of a new viceroyalty at Santa Fe de Bogotá (now Bogotá) in 1717. To protect the south against the British and against Portuguese incursions from Brazil, the Viceroyalty of La Plata was created in 1776. The centrepiece of Caroline reform was the introduction of the intendant system in the colonies to tighten up local administration. Most dramatic of all was the abolition of the monopoly of Cádiz, by which all trade to the colonies had to go through that port. Beginning in 1778 non-Castilian ports could trade directly with the colonies.

The new policies brought some immediate and striking results. Energetic viceroys and ministers and ruthless intendants doubled or tripled imperial revenues. The volume of Spanish goods in the American trade increased tenfold in ten years, so that Britain became seriously concerned at the Spanish revival. But imperial free trade would not satisfy the growing demand from Creole producers for free trade with all nations. Nor did the colonial oligarchs desire efficient government and higher taxation; they preferred bad government that let them control their own affairs. They showed their discontent in a series of revolts in the 1780s; only the fear of the native Indians drove them back to allegiance to the Spanish crown.

Domestic reforms. The domestic reforms of Charles III are more interesting for what they intended than for what they accomplished. They were not directed, as has often been maintained, at fostering a "bourgeois revolution." The middle classes were too weak, in a predominantly agrarian country, for the role of a modernizing elite; nor did Charles III contemplate a frontal attack on the traditional nobility. The purpose of reform was to remove what seemed to civil servants to be "traditional" constrictions on economic growth and administrative anachronisms that prevented the efficient exercise of royal power.

Reorganizations of the machinery of central government made for greater executive efficiency, but complete rationalization was never achieved; the old machinery of the councils persisted, with the Council of Castile as the ultimate decision-making body. An attempt to establish royal control of municipalities (without which, reforms could not get past the oligarchic councils) was likewise only a partial success. Most of the public works that characterized the late 18th century were the achievement of vigorous captains general. The extensive civil functions of these military officials was the first sign of a hybrid military–civilian government that, in another form, was to be developed in the 19th century.

Nor was the economic structure of agrarian Spain modified. All the chief reformers believed that the great and extensively cultivated estates, especially in Andalusia and Extremadura, constituted the greatest bar to agricultural prosperity. The landless underemployed proletariat who worked the large estates began to alarm reformers. "Why in our villages and towns are these men without land and in the countryside land without men? Bring them together and all will be served." Property should be more widely distributed, it was felt; there should be a free market in land. Yet none of the reformers was radical enough to push through a wholesale assault on private property or on the civil entail (the juridical instrument by which the *latifundios*, or large estates, were preserved intact). Acts such as the limitation of future entail, which preserved great estate intact over generations (1789), the limitation of the privileges of the Mesta (1779), and the right to enclose olive groves and irrigated land (1788) showed that the reformers believed primarily in the right of private individuals to do what they liked with their own; the unrestricted pursuit of private profit, they believed, would bring public prosperity. *Corporate* property was the enemy: hence the proposal that common lands owned by municipalities and the crown should be sold for individual cultivation and hence the proposals to end ecclesiastical entail (mortmain).

The attack on the privileges of the greatest corporation in Spain, the church, was less radical than has sometimes been maintained. Charles III himself was a devoted Catholic—he dedicated Spain to the Immaculate Conception. Some of his servants were fashionable anticlericals. Most were regalists; that is, they asserted the right of the crown to control over the church in civil matters. In the extreme regalists' view, the state should take care of charity and education; it should subject priests to civil jurisdiction for civil crimes and assert the traditional rights of the crown over church appointments.

The regalist attack on the church

The main attack of the regalists fell on the Jesuit order. In 1766 a serious riot in Madrid revealed some of the difficulties confronting the reformers. The abolition of fixed wheat prices during a bad harvest (a step that reflected the reformers' belief in the virtues of a free market) and an attempt to reform outlandish fashions in popular dress brought out the mob in Madrid. The Jesuits were alleged to have fostered the riot and were expelled from both Spain and America in 1767. The importance of this expulsion has been overestimated. Already expelled from France and Portugal, the Jesuits were bitterly criticized by rival orders and by the secular clergy: 42 of the 56 bishops approved of the expulsion. Again the expulsion was a *negative* achievement; more ambitious plans to set up a state university system and a state welfare organization failed.

The question arises of the extent to which the policies of Charles III resulted from the acceptance by his servant of the precepts of the Enlightenment. Certainly Aranda, the "Hammer of the Jesuits," and Olavide, whose flippant anti-clericalism was to be his undoing at the hands of the Inquisition, were what were called *esprits forts* (*i.e.*, French-influenced radicals); their views gave a sharp edge to traditional regalism. Jovellanos was a disciple of Adam Smith. Although his famous *Informe sobre la ley agraria* ("Report on the Agrarian Law") contains not a single original thought, the book is significant in that it attempts to apply dogmatic laissez-faire ideology to Spanish conditions. Here was one of the foundations of Spanish liberalism.

One of the aims of the Enlightenment was to produce a bourgeois society in which no traditional prejudices or traditional institutions should limit the pursuit of private profit—the foundation of national prosperity. But in Spain the social base was too narrow. In spite of economic changes over the century, the bourgeoisie was limited to the coastal towns, Barcelona and Cádiz in particular, with branches in Saragossa and the Basque country. Patriotic societies, organized with government encouragement from 1765 onward, were meant to provide the provincial basis for a progressive society and to familiarize Spaniards with European advances in technology and agriculture. They did not get much beyond the status of local reading rooms and debating societies.

Traditional Catholic society was still strong, if under attack from a minority of intellectuals and civil servants. As the reaction of the countryside after 1808 was to show, the church was still a great social power. The Duke of Wellington observed that "the real power in Spain is in the clergy. They kept the people right against France." To most of the clergy the new ideas of the Enlightenment were "foreign" and dangerous, and there could be no such thing as moderate progress encouraged by the king himself—the notion of a "revolution from above" that was to haunt subsequent Spanish history. Voltaire, Locke, and Rousseau were quite simply dangerous heretics, though the Inquisition proved powerless to prevent the clandestine circulation of their works. It was the clerical attacks on heretics as much as the subversive works themselves that familiarized a narrow stratum of society with new ideas. When the French Revolution exposed the dangers of progressive thought, the traditionalist cause was immensely strengthened and the Inquisition appeared to the crown itself to be a useful instrument to control the spread of dangerous ideas.

CHARLES IV AND THE FRENCH REVOLUTION

In 1788 Charles III, who had been the "nerve" of reform in the sense that he loyally supported able ministers, was succeeded by his son, Charles IV, a weak, amiable man dominated by a lascivious wife, María Luisa. Spain was

ruled after 1792 by her favourite, Manuel de Godoy, a handsome, plump officer from the lower nobility. It was not merely that Godoy was incompetent and self-seeking; the French Revolutionary and Napoleonic Wars (*q.v.*) put unbearable pressures on a weak power. Reform was now dangerous. Neutrality was impossible; alliance with either France or the anti-Revolutionary coalitions engineered by Britain proved equally disastrous.

The influence of Godoy

Floridablanca, the prime minister, disliked both the internal effects of the Revolution (the spread of radical as opposed to government guided reform) and its external consequences (the weakening of the anti-British alliance). His hostility to France was the cause of his dismissal. Aranda, a friend of France, was discredited by the excesses of the Revolution (1792). Godoy became prime minister at age 25 because the old ministers of Charles III had failed to devise a foreign policy. In domestic affairs, Godoy supported a mild version of the enlightened reformism of his predecessors; but, like them, he failed to find a satisfactory place for Spain in the Europe of the French Revolutionary and Napoleonic Wars (1800–15).

Spain had no alternative but to declare war on France after the execution of Louis XVI (1793). The war was popular but disastrous; in 1794 the French armies invaded Spain, taking Bilbao, San Sebastián, and Figueras. Godoy feared the spread of revolutionary propaganda in the wake of the French armies in Catalonia and the north (there was a republican conspiracy in 1795). Above all, he was convinced that Britain was the true enemy of Spain. Thus the Treaty of San Ildefonso (1796) represented a deliberate choice: the French alliance, irrespective of the nature of the French regime, as the only policy for a weak imperial power.

The consequences of this choice were disastrous. War with Britain became inevitable and cut off Spain from America, opening its colonial markets to Britain and the United States; Godoy invaded Portugal with success in the War of Oranges (1801) but the defeat of the Franco-Spanish fleet by the British at Trafalgar in 1805 and the "selfishness" of Napoleon caused him to seek a rapprochement with the allies. Had Napoleon lost the Battle of Jena (1806) against the Prussians, Godoy would have joined the Fourth Coalition against him.

Godoy's position was now extremely weak. He was unpopular at home. The court aristocracy disliked him as an upstart; severe inflation produced great hardship among the poorer classes; a campaign of gossip was mounted against him by those intellectuals whom he did not patronize. He now hoped, with French help, to dismember Portugal and to secure personal salvation in a principality. This curious hope was the basis of the Treaty of Fontainebleau (1807) by which Napoleon and the Spanish government agreed upon the conquest and partition of Portugal. When French troops on the way to Portugal occupied the fortresses of northern and central Spain and when Napoleon demanded territorial gains in Spain itself, Godoy's policy was bankrupt. It was obvious that Napoleon had lost all faith in Godoy and Spain as an ally: the "dirty intrigues" of the heir to the throne, Ferdinand, prince of Asturias, against his father and Godoy led Napoleon to think of drastic intervention in Spanish affairs.

Appointment of Joseph Bonaparte

The opportunity for direct intervention by Napoleon was given by the Revolt of Aranjuez (March 17, 1808), in which the partisans of Ferdinand (an alliance of discontented aristocrats and others opposed to Godoy) impelled the abdication of Charles IV and the dismissal of the favourite. Napoleon summoned both the old king and Ferdinand VII to Bayonne, where both were impelled to abdicate. The Spanish throne was offered to Napoleon's brother Joseph.

THE FRENCH INVASION AND THE WAR OF INDEPENDENCE, 1804–14

Joseph could count on the support of cautious, legalistic administrators and soldiers, of those who believed resistance to French power impossible, and of those who considered that Napoleon might "regenerate" Spain by modern reforms. These groups became convinced *afrancesados*, as members of the pro-French party were called. Relying on their support, Napoleon entirely underestimated the possibility of popular resistance to the occupation of Spain by French armies. Although the rising of May 2 in Madrid was suppressed, local risings against the French were successful wherever French military power was weak.

The War of Independence. After the deposition of King Ferdinand, "patriot" Spain outside the control of the French armies split into a number of autonomous provinces. Resistance centred in provincial committees (juntas) that organized armies. A Central Junta at Aranjuez sought to control this nascent federalism and the local levies, and Spanish regular troops defeated a French army of inferior, ill-supplied troops under Dupont at Bailén in July 1808. The French retired from Madrid. Napoleon himself invaded Spain and by 1809 was in control of most of the peninsula. The Spanish regular army, led by incompetent generals, suffered defeat after defeat.

The War of Independence, or, as the English call it, the Peninsular War, became for Napoleon the "Spanish ulcer," and he attributed his defeat in Europe to its requirements for men and money. He was defeated not by the inefficient Spanish regular army but by British troops under Wellington advancing from Portugal with the aid of Spanish guerrillas. The guerrillas represented popular resistance as against the supposed pro-French proclivities of the "possessing classes." As the main battles—Talavera (July 1809) and Vitoria (June 1813)—were fought by Wellington, the guerrillas pinned down French garrisons, intercepted dispatches, and isolated convoys.

The Constitution of Cádiz. The significance of the war years lies in the re-creation of a patriotic spirit to cover the bare bones of Bourbon administrative centralism and in the explicit formulation of a liberal ideology that was to be a dynamic factor in Spanish history.

The Central Junta and its successor, the Regency, were compelled to summon a Cortes (Parliament) in order to legitimize the situation created by the absence of Ferdinand VII, who was in France. Conservatives conceived of its task as the mere supply of the sinews of war on behalf of an absent king. The Cortes, when it met at Cádiz in 1810, was dominated by liberals who wished to go beyond the mere support of the war effort and establish a constitution that would make impossible the revival of rule by a favourite like Godoy. The Constitution of 1812 was to become the "sacred codex" of Latin liberalism.

The Constitution of Cádiz gave Spain a strictly limited monarchy (the king must work through his responsible ministers), a single-chamber Parliament with no special representation for the church or the nobility, and a modern centralized administrative system based on provinces and municipalities. All this had little basis in the medieval precedents quoted in the debates and was inspired by the constitutions of revolutionary France. Bourgeois individualism inspired legislation against entail, favouring instead the sale of common lands and the individual's right to dispose of his property as he might wish. The abolition of the Inquisition represented a mixture of historic regalism and modern anti-clericalism. This measure produced a conservative reaction, as did all liberal anti-clericalism through to the Second Republic in the 1930s. This reaction gave a popular underpinning to Ferdinand VII's destruction of liberalism and all its works in 1814 (see below).

Abolition of the Inquisition

In addition to initiating a liberal tradition, the War of Independence bequeathed two problems: first, generals chafed at control by civilian juntas and on occasion overthrew them, thus initiating the phenomenon of the pronunciamento, or military revolution; second, the *afrancesados*, often men of liberal inclinations but tarred with the accusation of collaborationism with the French, were left as an indigestible element within liberalism itself.

FERDINAND VII, 1814–33

With the help of an army corps and of conservative sentiment that had been outraged by the liberalism of 1812, Ferdinand returned from exile in France to rule Spain as

an absolute monarch. In 1820 he was forced by military sedition to return to constitutionalism (1820–23). For the last "ominous" decade of his reign he returned to a relatively enlightened form of ministerial despotism. The years 1814–20 were dominated by the attempt to re-establish Spanish rule in America and by the problem of maintaining an inflated wartime army with a permanent economic deficit.

The failure of liberalism. The solution of the Cádiz liberals to the imperial problem had been to make the colonies constitutionally part of metropolitan Spain by giving them representation in the Cortes. This did not stop the revolt of the colonies, where the Creoles wanted local self-government and free trade rather than liberal centralization. In 1814 it was not clear that the rebels under Simón Bolívar in the north and San Martín in the south would succeed; but all of Ferdinand's efforts to assemble a large army and a fleet to send to America failed. In 1820 the army that was to subdue the colonies revolted against the king in a pronunciamento organized by Major Riego Nuñez and supported by the local liberals organized in Masonic lodges.

The Revolution of 1820

The Revolution of 1820 brought into power the "jail birds"—liberals of the 1812 vintage who had been persecuted by Ferdinand VII. The Constitution of 1812 was re-established together with other liberal legislation, including the sale of monastic property.

The liberal system failed once more because it was a minority creed sustained by a section of the army—the military radicals such as Riego—against a mounting conservative reaction that had been fed once more by an attack on the church, especially the monasteries. The liberals themselves split. The more conservative wing (led by Martínez de la Rosa, a dramatist) wished for a more moderate constitution, based on the French Charter of 1814, which would give better representation to the upper classes and would not be totally unacceptable to the King, as was the "prison" of the Constitution of 1812. The King gave no support whatsoever to this movement and, in a cowardly fashion, disowned a rising of the guards' regiments that backed it. Thus the extreme radicals (*exaltados*) gained control by means of demonstrations in the streets, organized by clubs run on the lines of the Jacobins of the French Revolution.

The conservative reaction developed in the north around the regency set up at Seo de Urgel. Without French help, the movement would not have been successful, but when Louis XVIII sent French troops (the "Hundred Thousand Sons of St. Louis"), the liberal armies disintegrated and the liberal system fell.

Once more revolution at home favoured revolution in the colonies. Mexican conservatives, who had no desire to be ruled by Spanish anti-clericals, successfully established an independent Mexico under Agustín de Iturbide (1822). Spanish military power in South America finally foundered on the decisive Battle of Ayacucho (1824). Of Spain's far-flung empire, only the islands of Cuba, Puerto Rico, and the Philippines remained.

The "ominous decade," 1823–33. The "ominous decade", as it was called by the persecuted liberals, began with a severe purge of liberals; but in its later stages the regime became increasingly unacceptable to extreme conservatives, who looked to the King's reactionary brother, Don Carlos. Ferdinand had to rely either on inefficient traditionalists who could raise no money on the European money markets or on the more liberal ministers who were able financiers. Ministers such as Luis López Ballesteros, a friend of the *afrancesados*, set the tone with a serious attempt at a government-fostered economic revival.

The Carlists

This return to 18th-century "ministerial despotism" did not satisfy the liberal exiles, who mounted ineffective invasions in 1824 and 1830. More important, the conservatives of Don Carlos' court faction accepted Ferdinand VII's rule only on condition that Don Carlos would inherit the crown. In 1829, Ferdinand married his niece María Cristina of Naples, who realized that her influence depended on the elimination of that of Don Carlos. In March 1830, her faction at court persuaded the King to exclude Don Carlos from the succession even if María Cristina produced a female heir. This attack on the rights of Don Carlos was the origin of the Carlist party and of the Carlist Wars, which were for over a half century to be a disrupting influence in Spain. After the defeat of an attempt to force María Cristina to recognize Don Carlos' rights during Ferdinand's illness (September 1832), María Cristina's faction was dominant at court. She succeeded in securing all important military commands in the hands of supporters of the claims of her daughter, Isabella. When, on September 29, 1833, Ferdinand died, Isabella was proclaimed queen, with María Cristina as regent. The First Carlist War (1833–39) broke out almost immediately.

ISABELLA II, 1833–68

The Carlist Wars. The dynastic war between Isabelline liberalism and Carlism was a savage civil war between urban liberalism and rural traditionalism, between the poorly paid and equipped regular army of the liberal governments, supporting Isabella, and the semi-guerrilla forces of the Carlists. The Carlist strength lay in the north, especially in the Basque Provinces and Navarre, where there was strong support for the *fueros* against liberal centralism and for the traditional Catholic order represented by the religious bigotry of Don Carlos and his circle. But the Carlists could not break out of their bases in the north to capture an important city. The great Carlist leader Tomás Zumalacárregui y de Imaz was killed in an attempt to capture Bilbao, and Don Carlos' expedition to Madrid failed (1837). In 1839 the Carlist commander staged a mutiny against the clerical court of Don Carlos and came to terms with Baldomero Espartero, the most successful of Isabella's generals.

María Cristina allied with liberalism out of military necessity, not from conviction. She would have preferred to grant administrative reforms rather than consent that her daughter should become a constitutional monarch. But the liberals alone could save her daughter's throne from the Carlists, and the minimum demand of all liberals was a constitution. As regent from 1833 to 1840, she therefore consistently supported conservative liberals against the radicals. The Royal Statute (1834) represented this alliance between respectable upper-middle-class liberals and the crown.

This conservative constitution, with its property franchise and the great powers it gave the crown in the choice of ministers, could not stop the drift toward the left implicit in liberalism itself. The radicals, who were the heirs of the *exaltados* of 1820–23, were installed in power first by a series of provincial town risings in 1835 and later, after a short-lived conservative reaction, by an army mutiny that forced María Cristina to accept the Constitution of 1812. The radical politicians, however, accepted a more moderate compromise constitution—that of 1837. Of more enduring importance was the sale of church lands to finance the war. The great disentailment carried out by Prime Minister Juan Alvarez Mendizábal and his successors altered profoundly the social structure of Spain by putting on the market large quantities of land, most of which was bought by large landowners or prosperous peasants.

Moderates, progressives, and the generals. The Carlist Wars left civilian politicians discredited, and generals became the arbiters of politics not, as in 1814–20, as intruders but as part of the political machinery. They became the "swords" of the two main political groups. The *moderados*, who were upper-middle-class oligarchic liberals fearful of democratic violence and upholders of the prerogatives of the crown, represented the conservative stream in liberalism. Their rivals, the *progresistas*, were the heirs of the *exaltados* and represented a lower stratum of the middle class; the *progresistas* were prepared to use the discontent of the urban masses in order to bring pressure on the crown to give them office. Their instrument was the Urban Militia. Espartero used his military faction and his supporters among the younger *progresista* politicians and their artisan followers in the great cities to oust María Cristina and establish himself as regent

(1841–43). Espartero proved a disappointment to the radical *progresistas*, who now allied with his conservative opponents under his military and political rival, Ramón María Narváez. In 1843 Espartero was thrown out by a pronunciamento.

The pronunciamento of Narváez

Narváez jettisoned his *progresista* allies through a court intrigue, and between 1844 and 1854 he and his fellow generals dominated domestic politics as representatives of the *moderados*. Their administrative, educational, and financial measures and the formation of the Civil Guard were lasting achievements; but the generals could not stabilize their rule on the basis of their constitution of 1845, a conservative revision of 1837. The period was disturbed by a series of *progresista* military risings.

To the left of the *progresistas*, who were prepared to accept the monarchy if it gave them office, a Democratic Party developed; it was prepared to dethrone Isabella II, who was declared of age to rule in 1843. Never strong in numbers outside the towns, the Democrats radicalized politics. Orthodox *progresista* politicians were embarrassed by their extremist attitudes but could not neglect their potential role as urban revolutionaries.

It was not the Democrats but a group of discontented generals led by Leopoldo O'Donnell who successfully revolted in 1854. They were prepared to sacrifice the dynasty because the Queen and her mother favoured a rigid court conservatism that effectively excluded them from influence.

The rebellious military oligarchs were forced to call in civilian and radical support. This turned their pronunciamento into a mild revolution. Isabella survived only through Espartero's political timidity. Unprepared to accept the backing of the Democrats as the "George Washington of Spain," he accepted an alliance with O'Donnell, who was determined to arrest the drift to radicalism. In 1856 he broke with Espartero, defeated a demonstration in his favour, and dissolved the National Militia, the instrument of the left-wing Progressives. The radical thrust was thus defeated.

Economic expansion. The "revolutionary" period of 1854–56 saw no important constitutional change; but the further extension of disentailment and the new law for limited companies provided the legal structure for a rapid expansion of the economy. To promote this expansion, there were new injections of foreign credit—particularly French—and new banks. This capital made it possible to begin the railroad network that was to provide the transport infrastructure for a national market. The textile industry of Catalonia flourished, as a modern wool manufacture grew up; in the Basque Country, the second pole of an industrial economy developed slowly around iron. Agriculture, still a dominant sector of the economy, expanded as the church and common lands provided new fields for the easiest of cash crops—wheat; the growth of such towns as Barcelona created a market for vegetables, wines, and fruit.

Expansion resulted in the classic expression of 19th-century liberalism—the haute bourgeoisie of finance and industry. Members of this new class ranged from solid Catalan manufacturers demanding protective tariffs to safeguard their gains, to daring speculators such as the railway financier Marqués de Salamanca, who took an active part in political life and the urbanization of Madrid. It also included successful generals ready to forget their humble origins. Together with the large landowners, these groups formed the oligarchy of liberalism.

The government that presided over this prosperity was O'Donnell's Liberal Union—an attempt to fuse all dynastic parties in a broad-based coalition. It provided a long period of stable government (1856–63), and had the Progressive politicians been less afraid of losing their left wing to the Democrats and had the Queen been willing to grant Progressives power, the dynasty might have survived. Instead, she excluded the Progressives and forced them, first to withdraw formally from political life, then to contemplate revolution. Their "sword," General Prim y Prats (a chemist's son and a striking example of the social mobility of the liberal army), was a resolute conspirator and the ablest of the political generals of the 19th century. When Isabella's court ministers alienated O'Donnell's followers, a powerful coalition was formed, and Prim dropped his alliance with the Democrats. A slump caused by bad harvests, a series of mediocre ultra-conservative governments, and a growth in democratic agitation among university intellectuals (whose main concern was the hold of the church over education) gave wide support to the military rising against Isabella. Her armies would not defend her, and she was forced to leave for France in September 1868.

O'Donnell's Liberal Union

THE REVOLUTION OF 1868 AND THE REPUBLIC OF 1873

The revolution that led to the dethronement of Isabella was the work of army oligarchs led by Francisco Serrano y Domínguez and Progressive conspirators behind Prim. The Democrats became active in setting up juntas after the revolution: for the most part they rapidly became Federal Republicans under the influence of the theories of the French anarchist Pierre-Joseph Proudhon as presented by their leader, Pí y Margall. The Democratic intellectuals' main contribution was to add a radical democratic content to the demands of the military oligarchy.

The generals were determined to keep the leadership of the revolution in their own hands by channelling it into a constitutional monarchy. Although they had to concede universal suffrage in the Constitution of 1869, they ruthlessly suppressed republican risings in the summer of that year. Their problem was to find a constitutional monarch. Prim's attempt to persuade a Hohenzollern to accept the throne was opposed by France and set off the Franco-German War in 1870. In November 1870, Amadeus, duke of Aosta, was elected king, and Prim, the kingmaker, was assassinated the day Amadeus entered Madrid.

Amadeus, duke of Aosta

Amadeus attempted to rule as a constitutional monarch. Opposed both by Republicans and by Carlists, he could form no stable government from the "September coalition" of former conservative Liberal Unionists, the ex-Progressives, and the moderate Democrats—now called Radicals. Once Amadeus called the Radicals to power, the conservatives deserted the dynasty. Amadeus abdicated after an attack by the Radicals on the army in February 1873. The Cortes proclaimed Spain a republic.

The Republic of 1873 came into existence to fill the political vacuum created by Amadeus' abdication. The Republican Party was neither strong nor united. When the Republican leaders, on legal scruples, refused to declare for a *federal* republic, the provincial Federal extremists revolted.

This Cantonalist revolt was serious in Cartagena, Alcoy, and Málaga. The Republican leaders had allowed attacks on the army that had reduced it to impotence. To conservatives and men of order, the country seemed on the verge of total dissolution; the Carlists were immensely strengthened by the "excesses" of Cantonalists. Too late, Emilio Castelar y Ripoll, the last president of the republic, tried to recapture the loyalty of the army. In January 1874, Gen. Manuel Pavia y Rodríguez de Alburquerque drove the Republican deputies from the Cortes building in the hope of creating a government of order. Pavia turned power over to Gen. Francisco Serrano to form a coalition government.

General Serrano took over as president of a unitary republic. His main task was victory over the Carlists, now a strong force in northern Spain. In this he failed, and on December 29, 1874, a young brigadier, Arsenio Martínez Campos, "declared" for Alfonso XII, the son of Isabella.

There was no resistance. The extreme threat of anarchy in 1873 had resulted in a strong conservative reaction, strengthened by the religious policies pursued since 1868. The Constitution of 1869 for the first time had allowed complete freedom of religion. In spite of a total failure to create political stability, the Revolution of 1868 bequeathed to Spain the model of a modern secular state based on universal suffrage. On the extreme left these years saw the effective penetration of anarchism, especially in rural Andalusia and in industrial Catalonia, where the lower classes deserted a long tradition of political action via the Democratic and Republican parties and

moderate unionism. Though anarchists were persecuted after 1873, the movement was kept alive by small groups of enthusiasts.

The Cuban revolt

The worst problem of the years 1868–75 was posed by the independence movement in Cuba, which, with Puerto Rico, was the last possession of Spain in America. Cubans had long resented the failure to reform rule by captains general, to grant some autonomy, and to ease the economic sacrifices imposed by the Spanish tariff system. The revolt that began in October 1868 made great demands on Spain in men (100,000 by 1870) and money. The Ten Years' War (1868–78) made difficulties for all governments in power in Spain after 1868 and forced abandonment of the most popular of the pledges made by the rebels in 1868: the abolition of the arbitrary and socially selective recruitment system. Like the Carlist wars, the Cuban War favoured the monarchical reaction.

THE RESTORED MONARCHY, 1875–1923

Once the Carlists had been defeated and the Cubans had accepted the peace settlement of El Zanjón (1878), the restored monarchy provided the most stable government Spain had known since 1833. This stability was sustained by respectable economic growth.

Stability, 1875–98. The architect of the restoration itself and of the Constitution of 1876 was Cánovas del Castillo. A superb politician, Cánovas had hoped for a civilian restoration; he accepted Martínez Campos' coup but used the young Alfonso XII to keep the military out of politics.

The Canovite system

The Canovite system was artificial in that it required the contrived rotation in office (*turno pacífico*) of a Liberal and a Conservative party; this in turn demanded governmental control of elections, which were run by caciques, or local political bosses, who controlled votes in their districts. Only in this way could the government selected by the king and the politicians in Madrid obtain a parliamentary majority; extensive corruption and the use of administrative pressures on electors were perhaps the only ways to make the parliamentary system work in an underdeveloped society. This system survived the death of Alfonso XII (1885) and only began to falter in the 1890s, toward the end of his wife's regency. The Carlist threat weakened with defeat, and the majority of Republicans in Spain were domesticated and reconciled to the use of "legal" means.

"Without being a rich country," wrote an economist in the early 1880s, "Spain has become comfortably off." This prosperity, untroubled by the claims of organized labour, was the result of the demand for iron ore after the invention of the Bessemer process (England and France invested heavily in mineral production), of the demand for Spanish wine after the devastations by phylloxera in France, and of the resumption of railway construction. The third largest wool industry in Europe grew up beside the older cotton mills in Catalonia. The boom did not break until the late 1880s, when an agricultural depression set in. A wave of economic pessimism preceded the political and intellectual reaction of 1898.

The loss of Cuba, Puerto Rico, and the Philippines in 1898 exposed the political system to severe criticism. No fiscal and political reform sufficient to satisfy Cuban demands could be effected within the framework of the monarchy, partly because of the pressure of the Spanish loyalist party in Cuba. A revolt of 1895 set off another costly war against Cuban guerrillas. The intervention of the United States could not be staved off by a last-minute grant of autonomy. Their humiliating and total naval defeat in 1898 became known to Spaniards as "the Disaster." Spain now lost all of its American possessions and the Philippines at the very time when the great European powers were building their overseas empires.

War with the United States

Opposition movements, 1898–1923. Criticism of the restoration monarchy came from the Catalan and Basque regionalists, from a revived Republican Party, from the proletarian parties, from the army, from the more forward looking politicians, and from the intellectuals.

Basque regionalism, though more akin in its ideas to extreme nationalist movements, was less a challenge than Catalan regionalism, which dominated Spanish politics until 1923. With its own Catalan language and a revived cultural tradition, it moved from a demand for protection of Catalan industry against "Castilian" free trade to a demand for political autonomy. The *Lliga Regionalista* (founded 1901), dominated by the Catalan industrialist Francisco Cambó de Asís and the theoretician of Catalan nationalism, Prat de la Riba, demanded the end of the *turno* and a revival of regionalism within a genuine party system. Cambó wished to solve the Catalan question "within Spain"—*i.e.*, by legal means and in cooperation with monarchical politicians. The revival of Catalanism, however, set off a "Castilian" reaction in which moderate Catalans were accused of selfishness or of hiding separatist aims under "respectable" regionalism. In the 1907 election the united Catalan front, Solidaridad Catalana, defeated the parties of the system but then divided into a right (which accepted a solution within the monarchy) and a left (which was to drift to Republicanism). Cambó's cooperation with Madrid brought Catalonia no tangible concessions.

Republicanism, which had degenerated into local politicking and lived on its memories of 1873, was revived by the remarkable oratorical and organizational talents of Alejandro Lerroux, whose Radical Republican Party offered a program that appealed to the alienated working class voter of Barcelona.

Here he was competing with a slow revival of anarchism, which veered between terrorism and educational propaganda and union activity; the anarcho-syndicalist union, the Confederación Nacional del Trabajo (CNT), was founded in 1910. The Socialist movement with its union (the Unión General de Trabajadores, or UGT, founded 1888) was relatively weak except in the mining districts of the north and in Madrid, where it was dominated by its French-influenced founder, Pablo Iglesias. In 1909 the Socialists abandoned their boycott of "bourgeois" politics and allied with the Republicans. This alliance was to give the party a political leverage in excess of its voting strength.

The "Generation of 1898"

The intellectuals' protest embraced the writers of widely differing ideas collectively known as the "Generation of 1898." Joaquín Costa y Martínez, a voluminous writer, was an especially harsh critic of *caciquismo*; he wanted a revived, effectively democratic, modernized Spain. His attempt to create a coalition of the "productive classes" against the political drones in the Unión Nacional failed. Miguel de Unamuno y Jugo saw regeneration in terms of a return to "pure" Spanish values. All of the Generation of 1898 distrusted politics as managed by professional politicians who were out of touch with the "real" Spain.

Among the politicians themselves, the conservative leaders Francisco Silvela and Antonio Maura and the democratic liberal José Canalejas sought to regenerate the system by widening the degree of political participation through "sincere" elections. They were opposed by the cratic liberal José Canalejas sought to regenerate the system by widening the degree of political participaty. This exposed Alfonso XIII (who succeeded in 1902) to the charge of political interference in a constitutional monarchy in which "opinion" did not provide the king with a choice between two strong parties. The danger of "sincere" elections to the political establishment was revealed by the Republican victories in 1903.

Revolt in Morocco

The call up of troops for Morocco, where Spanish troops were engaged in operations protecting the Spanish coastal possessions, set off the Tragic Week of 1909 in Barcelona. Public order collapsed; anarchists and Radical Republicans burned churches and convents. Maura was driven from office because Alfonso XIII (who had succeeded in 1902) accepted the Liberals' estimate of the harm Maura's firm repression would inflict on the monarchy. Rather than resist a Liberal Party that had allied with the "sewer" (the Republican parties), the King held that the Liberals were a useful "lightning conductor," protecting the monarchy from the threat on the left. Ever since Maura's fall, the King's diagnosis had been challenged by conservatives.

World War I produced increased strains. Real wages fell, making the unions restive, and in 1917, junior officers formed juntas and struck for better conditions. The ensuing crisis was exploited by Catalan politicians, increasingly pressing for autonomy. The revolutionary coalition split up, and a general strike frightened Cambó and the Catalans, who threatened to call a national convention that would unite all critics of the monarchy. The formation of a national government under Maura ended the last serious attempt to regenerate the political system and to make it respond to reformist and Catalan demands. No subsequent government was strong enough to face the anarchist agitation in Catalonia and the Moroccan War.

Spain had been granted a protectorate in Morocco in 1912. In the effort to pacify it, economizing politicians were ready for compromise with the tribes, but the generals saw conquest as the only solution. A bid by Gen. Fernández Silvestre, reputedly backed by Alfonso XIII, for a crowning victory ended in the terrible massacre of Spanish troops at Annual in 1921. Opposition politicians were determined to expose the King's action and criticize the army.

Anarchist influences

World War I had given the anarcho-syndicalist movement great power. At the same time there grew a terrorist fringe, which the leaders of the CNT could not control. These leaders (Salvador Sequí and Angel Pestaña), though they shared the anarchist contempt for political action, wished to build unions powerful enough to challenge the employers by direct action. They mistrusted the libertarian tradition of spontaneous revolution as a means of toppling the bourgeois state. The great Barcelona strike of 1919 was the most impressive in Spanish history. When the employers' violent reaction discredited the moderates in the Confederación Nacional del Trabajo, it was followed by a wave of assassinations by gangsters employed by both the anarchists and the employers.

Thus, when Gen. Primo de Rivera staged his pronunciamento in September 1923, he could count on the support of the conservative classes fearful for social order. The Lliga, for instance, became more concerned about suppressing strikes than about Catalan autonomy, for which it had campaigned from 1915 to 1918. He could also count on the tacit alliance of Alfonso XIII, who was tired of politicians who could not provide him with effective governments.

PRIMO DE RIVERA (1923–30) AND THE SECOND REPUBLIC (1931–36)

Primo de Rivera. Primo de Rivera was a political improviser who believed his mission was to save Spain from the old politicians and to hand over government (after an interval of personal rule) to "clean" patriots. He failed to complete the process because his rule became increasingly unpopular, especially among the intellectuals and Catalans. The September 1923 coup by which he had gained power had been widely welcomed in Catalonia, where, as captain general, Primo had listened sympathetically to Catalan demands. Soon afterward, however, Primo became a Spanish patriot and permitted an "anti-Catalan crusade." His followers' attempts to build up a political party (the Patriotic Union) to run a regenerated Spain and to provide it with an ideology collapsed.

At first he ruled via the army. In spite of initial quarrels with the African commanders, whom he forced to retreat in Morocco, the Military Directory was responsible for final victory in the protectorate. The Spanish, collaborating for the first time with the French, landed at Alhucemas (al-Hoceima) in September 1925 and defeated the most successful tribal leader, Abd el-Krim. By 1927 the whole of the protectorate was successfully occupied.

The Civil Directory (1925–30) was responsible for a thorough overhaul of local government and for an ambitious public works program to increase irrigation, hydraulic power, and road building. Primo's economic nationalism entailed strict protectionist policies and an attack on foreign oil monopolies. The complicated bureaucratic control of industry did not endear him to capitalists after 1926; on the other hand he collaborated successfully with the UGT while suppressing the CNT. The Civil Directory failed in its main task: to win sufficient political support in the National Assembly summoned for 1928 to facilitate a return to quasi-constitutional government.

Throughout his early years Primo was favoured by an economic expansion based on favourable terms of trade for Spanish exports. By 1929 the peseta began to fall in value in spite of desperate measures to save it. Economic recession alone would not have ruined the dictator, but he lost the support of both the army and the King. The army turned against him as a result of his attempts to abolish the privileges of the artillery and engineer corps. The King believed that student protests, the growing discontent in Catalonia, and the increasing conspiracies of the "old" politicians imperilled the dynasty. On January 28, 1930, he forced Primo's resignation.

Alfonso acted too late. His earlier support of the dictatorship ruined him in the eyes of the politicians and public. The weak governments of Gen. Dámaso Berenguer and Adm. Juan Bautista Aznar could barely keep order. At San Sebastián (August 17, 1930) an alliance of former liberal monarchists, Catalan politicians, and Republicans agreed to overthrow the monarchy. The failure of a Republican military rising at Jaca (December 12, 1930) saved them from establishing a republic by force. The municipal elections of April 12, 1931, proved that the great cities were overwhelmingly Republican. Rather than face civil war and street demonstrations in Madrid, Alfonso XIII left Spain.

Abdication of Alfonso XIII

The Second Republic. The history of the Second Republic falls into four distinct phases: (1) the Provisional Government, which lasted until the religious issue forced its resignation in October 1931; (2) the governments of the Left Republicans and Socialists, which ruled from October 1931 and were defeated in the elections of November 1933; (3) the conservative government of the Radical Republicans and the Catholic right from November 1933 to February 1936, punctuated by the revolution of October 1934, and that ended with the electoral victory of the Popular Front in February 1936; and (4) the government of the Popular Front and "the descent into violence" that culminated in the military rising of July 1936.

The Provisional Government was a coalition government, presided over by Alcalá Zamora, whose Catholicism reassured moderate opinion. The coalition included all the groups represented at San Sebastian: Lerroux's Radicals, the Catalan left, the Socialists, and the Left Republicans dominated by Manuel Azaña y Diaz.

The elections to the Constitutent Cortes strengthened the Socialists and Left Republicans and thus upset the parliamentary balance between moderate Catholic Republicans and the left. It was the left that imprinted its views on the constitution, especially its religious clauses. The historically conditioned anti-clericalism had led the government to tolerate an outburst of church burning (May 1931). The Socialists and Left Republicans inserted in the constitution an attack on religious education and the regular orders, which forced the resignation of Alcalá Zamora and Miguel Maura, his minister of the interior.

This direct and, it would seem, ill-advised clash with Catholic sentiment was to provide a base for the construction of a party of the right devoted to the reversal of the church settlement. This party, the creation of the Catholic politician Gil Robles y Quiñones, became Acción Popular; it was to become the main component of the right-wing electoral grouping, the Confederación Española de Derechas Autónomas (CEDA). To the left its "accidentalism" (the doctrine that forms of government were irrelevant provided the church could fulfill its mission) was suspect.

The Acción Popular

From October 1931 the government, with Azaña as premier, was in the hands of Left Republicans and Socialists, with the Catholic right, the Basque Catholics, the Navarrese Carlists, and Lerroux's Radicals in opposition. Azaña aimed to create a modern democracy; labour legislation would be the work of the Socialists, with the UGT leader, Largo Caballero, as minister of labour.

In April 1931 there was a danger that Catalonia might

declare its independence within a federal state. Azaña's greatest achievement was the settlement of the Catalan question. He overcame conservative Republican opposition to limited home rule under the Generalidad, which was controlled by the Catalan left (Esquerra) under Luis Companys. Largo Caballero's legislation provided labour with a strong negotiating position but could not, in itself, mitigate the mounting unemployment, which was particularly serious in the *latifundios* of the southwest. Since new machinery for the settlement of labour disputes was dominated by the UGT, it was opposed by the CNT, now influenced by the extreme revolutionary apoliticism of an anarchist group, the Federación Anarquista Ibérica (FAI). Violent strikes were frequent.

Sedition from the right came to a head in General Sanjurjo's pronunciamento in Seville (August 10, 1932). A fiasco, it reflected the army's resentment at Azaña's military reforms, which were intended to create an efficient army by pensioning off superfluous officers; though the reforms were necessary, they were carried out with what Gen. Emilio Mola called a "vindictive spirit" that made them resented. Politically more dangerous than Sanjurjo's abortive coup were the steady growth of Gil Robles' Acción Popular and the Socialists' desertion of the Azaña coalition, as Largo Caballero wearied of cooperation with "bourgeois" parties. In the elections of November 1933, therefore, the left was divided and the right relatively united in the electoral union of the CEDA. Given an electoral law that favoured electoral coalitions, the CEDA and Lerroux's Radicals, now a respectable middle class party, triumphed.

Conservative reaction

In spite of substantial achievements, the first two years of the republic dissatisfied important groups. The Socialists believed they could extract no further gains for labour from the "bourgeois" republic. A limited agrarian reform did little to ease chronic underemployment in the southwest. At the same time half-hearted Socialism alienated capitalists. Perhaps the most serious failure was in the realm of public order. The sedition of the extreme right was parallelled on the left by the strikes of the CNT against what it called "the republic of hangmen." The Law of Public Order and the government's sharp repression failed to control outbreaks and laid it open to the accusation of dictatorship.

The newly elected Cortes was dominated by Lerroux's Radicals and the CEDA. Lerroux could not govern without the support of Gil Robles. With power within sight, Gil Robles accentuated his legalism, to the distaste of the militant monarchists among his supporters. Nevertheless, his party was regarded by the Left Republicans and Socialists as a Catholic brand of the Fascist trends now winning out in western Europe.

When CEDA entered Lerroux's government the Socialists staged a revolution, which the Left Republicans did not disown. Revolutionary councils were set up in the mining districts of Asturias, where there was considerable destruction of property. In Barcelona the revolution was the work of Catalan nationalists, who believed autonomy was imperilled by the actions of the Madrid government in overruling an agrarian law passed by the Generalidad. Unsupported by the CNT, it was quickly suppressed.

The October Revolution of 1934 is the dividing point in the Second Republic. The Socialists, fearing the fate of their Austrian and German brothers, had revolted against a legal government and thereby established in the minds of the right the fear of "Red" rebellion. In the subsequent repression by the army lay the emotional origins of the Popular Front against "Fascism"—a re-creation of the Azaña coalition of Left Republicans and Socialists in order to fight the elections of February 1936.

The Popular Front

The Popular Front won the election of February 1936 by a narrow majority. Spain had become polarized, and the division was to intensify.

The Popular Front government was exclusively Republican, though it was supported in the Cortes by the Socialists. Under Largo Caballero, the left Socialists put increasing pressure on the government, using revolutionary language if not intending revolution. Largo Caballero's revolutionary rhetoric concealed his hope that the orthodox Republicans could be eased from power, leaving the field open to a pure Socialist government, which his followers chose to call, in Leninist language, the "dictatorship of the proletariat."

Just as the fears of the "Fascism" of the right justified the defensive reaction of the Socialists, so the right argued that the Republican government was a prisoner of the revolutionary left (they could point to a rapid growth of Communist influence). The Falange, founded by José Antonio Primo de Rivera, the son of the dictator, was a nationalist anti-Marxist grouping appealing to youth ready to engage in street affrays with the Socialist Youth. Conservatives rallied behind the right-wing National Front, which openly appealed to the military to save Spain from Marxism.

The army was the key factor; by the early summer a young officers' conspiracy was backed by Mola, Manuel Goded, and, finally, Francisco Franco. The murder of Calvo Sotelo with the connivance of government security forces was the final outrage for the right and the army.

THE CIVIL WAR

The rising of the military started in Morocco on July 17, 1936, and spread to the garrisons of metropolitan Spain in the following days. The Civil War took place because the rising was successful only in Old Castile, in Navarre, where Carlist support was decisive, and, of the larger towns, in Saragossa, Seville, Córdoba, Valladolid, and Cádiz. Galicia soon went over to the Nationalists, as did most of Andalusia. Catalonia and the Basque Provinces were loyal to the government because the Republic guaranteed their autonomy. In Madrid and Barcelona, the security forces, helped by the workers who were armed belatedly by the government, defeated the officers. Thus, in broadest terms, the Republic held the centre, the Levant, Catalonia, and the Basque industrial zones; the Nationalists held the food-producing areas. One consequence was to be an increasingly acute food shortage in the Republican zone.

The role of the workers in defeating the rising made their organizations the power in the Republican zone. The legal government was by-passed or totally supplanted by local committees and trade unions; the workers' militia replaced the dissolved army. In many parts of Spain a social revolution took place in July 1936: factories and farms were collectivized. The English novelist George Orwell described Barcelona, where the CNT was all-powerful, as "a town where the working class was in the saddle." The success of working class control, in terms of increased production, is difficult to estimate.

This revolution was distasteful to the Left Republicans and to the Communist Party, which rapidly grew in number and in political influence because it controlled the supply of arms from the Soviet Union. In the name of an efficient war effort and the preservation of "bourgeois" elements in the Popular Front, the Communists pressed for a popular army and central government control. In September–November 1936, the CNT was brought into the government of Catalonia and into Largo Caballero's ministry in Madrid—an astonishing step for a movement that had consistently rejected "bourgeois" politics. The CNT militants did not approve the leaders' "surrender" and the dismantling of the militia-backed revolution.

A small Marxist revolutionary party, the Partido Obrero de Unificación Marxista (POUM), which rejected the Popular Front in favour of a workers' government, set off a rebellion in Barcelona in May 1937. The Communists, Republicans, and anti-Caballero Socialists used this as an excuse to oust Largo Caballero, who proved insufficiently pliable to Communist demands. The government led by the Socialist doctor Juan Negrín was a Republican–Socialist–Communist concern. The great unions, the UGT and CNT, were replaced by the political parties.

The Communists were correct in arguing that the committee–militia system was militarily ineffective. General Franco's army, ferried over from Morocco, cut through the militia and arrived before Madrid by November 1936. The successful resistance of the city, which was stiffened by the arrival of the International Brigades and

Soviet arms, meant that the Civil War would be prolonged for two years.

Victory would go to the side with the best army, with unified political control, and with adequate arms supply. The core of the Nationalist army was the African army commanded by General Franco. Abundantly supplied with generals, the troops and junior officers of the Nationalist armies were well trained. Given the confused political control in Republican Spain, the secure military and political command of Franco (from October 1936) was decisive. In April 1937 he ruthlessly incorporated the Falange into a unified movement under his leadership.

Both sides sought help from abroad. The Republic consistently hoped that France and Great Britain would allow them to acquire arms. Partly because of fear of a general war, partly because of domestic pressures, both powers backed non-intervention (*i.e.*, self-imposed restriction of arms supply by all powers). General Franco appealed immediately to Hitler in Germany and to Mussolini in Italy, which supplied aircraft early in the war. The Germans, in return for mineral concessions, supplied the Condor Legion (100 combat planes), and the Italians sent ground troops; both supplied tanks and artillery.

The Western democracies protested but did nothing. The Soviet Union alone responded to the breakdown of non-intervention by supplying arms to the Republican side. Soviet supplies were of great importance (tanks, aircraft, and a military mission) after October 1936. The Communist International also organized the International Brigades. In 1938, Soviet supplies dropped off, and the balance of arms supply was decisively in favour of the Nationalists.

Once the Popular Army replaced the militia (which was held to be responsible for the easy advance on Madrid and the fall of Málaga to Italian troops in February 1937) the Republic held Madrid and defeated two flanking attacks in the battles of Jarama (February 1937) and Guadalajara (March 1937), where the International Brigades decisively defeated a motorized Italian corps.

After his failure at Madrid, General Franco transferred his effort to the north, where the bombing of Guérnica, on April 26, 1937, by German planes outraged public opinion in the democracies: by October 1937, General Franco had captured the industrial zone, shortened his front, and won a decisive advantage. Diversionary battles at Brunete and in Aragon were costly and revealed the great weakness of the Republican Popular Army: its incapacity to sustain an initially successful offensive.

When General Franco concentrated again on Madrid the Republican army staged its most effective offensive in the winter Battle of Teruel (launched December 15, 1937). General Franco recovered Teruel, drove to the sea and committed his one strategic error in deciding on the difficult attack on Valencia rather than turning against Catalonia. To relieve Valencia the Republicans attacked across the Ebro (July 24, 1938); once more they failed to exploit the breakthrough, and the bloody battle exhausted the Popular Army.

The final Nationalist campaign in Catalonia was relatively easy. On the Republican side, the feasibility of continued resistance, which was supported by the Communists and Negrín, caused acute political divisions. On March 7, 1939, a civil war broke out in Madrid between Communists and anti-Communists. On March 28 the Nationalist forces entered a starving capital.

Capture of Madrid

FRANCO'S SPAIN, 1939 TO PRESENT

The first decade of General Franco's government was harsh. As a result of the Civil War roughly 1,000,000 Spaniards were dead or had emigrated. Fierce political repression was matched by economic hardship. General Franco's alliance with the Axis powers brought diplomatic isolation from the Western democracies. After 1950, growing foreign investment and change in economic policy brought an increase in material prosperity and a diminution of opposition activities.

Recovery after 1939 was made difficult by the destruction caused by the Civil War (especially of railway rolling stock), by a severe loss of skilled labour, by a series of bad droughts (especially 1949), by restrictions on the import of capital goods imposed by the European war, and by misguided attempts at state-controlled industrial development within a protected economy. The national income dropped back to the levels of 1900, and real wages fell to half of those of 1936. These were the years of near famine, of the black market, and of rural misery and flight to the towns.

General Franco's sympathies (and, even more so, those of his foreign minister, Serrano Súñer) lay openly with Germany and Italy, to whom he gave moral and material support. He nevertheless demanded a high price for his military cooperation against the Western Allies, on whom he was dependent for food and oil imports. He consistently maintained his mission against world Communism, however, sending the Blue Division against the Soviet Union (1941–43).

Franco's relations with the democracies

When it appeared likely that the Allies would win the war, in 1943, General Franco declared Spain's neutrality, but this late change earned no reward from them. Though Britain's Winston Churchill recognized the value of Spanish neutrality, the Allies were unwilling to desert the Soviet alliance; to the American and Soviet leaders, Roosevelt and Stalin, General Franco was a "Fascist." With Churchill's defeat by the Labour Party in Great Britain in 1945, the three great powers all declared their hostility to General Franco's government as undemocratic.

The declared hostility of the dominant great powers and the exclusion of Spain from the United Nations gave the opposition in Spain new life. Don Juan, the heir of Alfonso XIII, presented himself as a democratic alternative, and agreements were negotiated between the liberal monarchists and the opposition in exile, excluding the Communists, in 1947. Valiant but futile guerrilla activities, inspired largely by the CNT and the Communist Party (1944–48), were easily suppressed.

General Franco met these difficulties, the most severe of his regime, with confidence. The Fuero de los Españoles (July 1945), guaranteeing personal freedoms (provided no attack was made on the government), was an attempt to establish a more democratic image. More significant was his alliance with the church, which was given control of education, and his reduction of the Falange (which had supplied a national Fascist ideology) to a mere appendage of his government. The ostracism of Spain by the United Nations and the consequent withdrawal of ambassadors from its member nations was skillfully turned into a means of rallying support for General Franco in the name of national unity.

General Franco's confidence came from his divination that, with the Cold War between the Western democracies and the Communist states, the United States could not sustain its hostility to Spain and that the Allies, while declaring for the democratic opposition, would not directly intervene to overthrow his regime. Hence the hopes of the opposition came to nothing; once more the monarchists and Republicans split, and the "internal" and "external" oppositions drew apart. In 1953 an agreement with the United States gave General Franco considerable financial advantages in return for the establishment of four United States bases in Spain. "At last," he could declare as the ally of a great power, "I have won the Civil War."

Establishment of United States air and naval bases

Thus by 1953, Spain appeared to be a stable state. Internal political control remained in General Franco's hands, ensured by his mastery of the army and his ability to play off the groups that supported him against each other. In these ministerial changes the Falange steadily lost power to the technocrats of the Opus Dei (a lay Catholic organization), which broke with the economic nationalism of the Falange in favour of a "normal" capitalist system, with market mechanisms replacing state control.

The new policies seemed justified by their economic success, which was aided by a rapid growth of tourism and foreign investment. As in many countries that have seen an economic miracle, spectacular expansion (a growth rate of 7.4 percent between 1962 and 1966) brought problems of inflation and a wages–cost spiral.

Investment went mainly into industry and some spectacular irrigation projects; the Spanish countryside remained poor and its productivity low. The rural poor of the south emigrated to the factories of Europe or to the cities of Spain.

While the new prosperity brought a degree of social mobility that was novel to Spanish society and satisfied the middle classes, rising prices entailed a revival of labour troubles. The workers became disillusioned with the "official" syndicates run by the Falange. They set up workers' commissions outside the official framework, and there were serious strikes, in spite of legislation that forbade them.

Sections of the church were sensitive to these discontents. While the episcopate, on the whole, held to the view that the church should support the state, younger priests saw the long-term dangers of such an alliance. Thus they were sympathetic to the workers' commissions.

An even more intractable problem was that of Basque nationalism, and, again, some of the Basque lower clergy were sympathetic to the nationalists. The nationalist movement developed an extremist wing that supported violent action.

The most serious problem was the failure of the system to provide a safety valve for discontents. In 1966 a new Organic Law of the State reintroduced a measure of direct election of a proportion of members of the Cortes by heads of families and their wives. But clearly this was not intended to reintroduce political parties: they remained, in General Franco's words, a symptom of the "particular furies" that had so often destroyed Spain. The trade unions were freed from Falangist control. Press censorship was replaced by a press law, which, though it left publishers open to prosecution by the state, nevertheless allowed a freedom of expression unknown since 1939. It looked as if gradual liberalization was under way.

The Organic Law of 1966

This prospect was clouded by student troubles and labour unrest. Just as the workers distrusted official unionism, so the students disliked official students' syndicates. They had protested violently as far back as 1956, and in 1968 their protests forced a closing of the universities. The students' grievances against the educational system were largely justified; but their political aims were ill-considered and their alliance with the workers tenuous. Student violence in Madrid and Barcelona alarmed the government, and in early 1969 a state of emergency lasted two months. This action took place against a background of rumours about General Franco's health and brought the problem of the succession to the forefront.

Since Spain was legally a monarchy, the choice lay between Don Juan, the son of Alfonso XIII, and Don Juan's son Juan Carlos. A choice of Don Juan, a critic of Franco, would lead to a parliamentary monarchy and give political life to Franco's enemies. Juan Carlos, who was close to Franco, was declared legal heir in 1969. In 1972 Franco decreed that at his death he be succeeded by Juan Carlos as chief of state and by the vice president, Luis Cerrero Blanco, as chief of government.

In the early 1970s Franco's government seemed to gain in strength. Spain made attempts to improve relations with eastern Europe and Latin America; a trade pact was signed with the Soviet Union in September 1972.

(R.Ca.)

BIBLIOGRAPHY. The principal bibliographical guide to Spanish history is BENITO SANCHEZ ALONSO, *Fuentes de la historia española e hispanoamericana*, 3rd ed., 3 vol. (1952). LUIS G. DE VALDEAVELLANO, *Historia de España*, vol. 1 (1952), is the best general history, but it extends so far only to 1212. FERRAN SOLDEVILA, *Historia de España*, 3 vol. (1952–54), is beautifully illustrated, but it sees the history of Spain from the Catalan point of view. The most comprehensive work eventually will be the collaborative *Historia de España* under the editorial direction of RAMON MENENDEZ PIDAL, of which five volumes have thus far appeared. JAIME VICENS VIVES is the editor of another collaborative work, the *Historia social y económica de España y América*, 5 vol. (1957–59). Among the most important interpretive works is AMERICO CASTRO, *España en su historia* (1948; Eng. trans., *The Spaniards: An Introduction to Their History*, rev. ed., 1971), the most recent statement of his thesis emphasizing the medieval interaction of Christians, Muslims, and Jews. CLAUDIO SANCHEZ-ALBORNOZ has taken sharp issue with him in his *España, un enigma histórico*, 2 vol. (1956). RAMON MENENDEZ PIDAL, *Historia de España*, vol. 1 (1947; Eng. trans., *The Spaniards in Their History*, 1950), is a judicious essay on the formation of the Spanish character. A highly readable and provocative overview of the trends of Spanish history is given by JAIME VICENS VIVES in his *Aproximación a la historia de España*, 5th ed. (1962; Eng. trans., *Approaches to the History of Spain*, 2nd ed. rev., 1970).

Ancient Spain: For prehistoric Spain, see LUIS PERICOT GARCIA, *La España primitiva* (1950); or ANTONIO GARCIA Y BELLIDO, *La península ibérica en los comienzos de su historia* (1953). GARCIA Y BELLIDO is also the author of works on the Carthaginians and Greeks in Spain: *Fenicios y Carthagineses en occidente* (1942) and *Hispania Graeca*, 3 vol. (1948). A useful work in English on the Roman period is C.H.V. SUTHERLAND, *The Romans in Spain, 217 B.C.–A.D. 117* (1939).

Medieval Christian Spain: In addition to the general works cited above, the following may be helpful: LUIS G. DE VALDEAVELLANO, *Curso de historia de las instituciones españolas. De los orígines al final de la edad media* (1968), an excellent survey of medieval institutions; ALFONSO GARCIA GALLO, *Manual de historia del derecho español*, 3rd ed. rev., 2 vol. (1967), especially valuable for its collection of documents relating to government and law. The most up-to-date work on its subject is JAIME VICENS VIVES, *Manuel de historia económica de España*, 5th ed. (1967; Eng. trans., *An Economic History of Spain*, 1969). E.A. THOMPSON, *The Goths in Spain* (1969), is the most recent work on the Visigothic kingdom. YITZHAK BAER, *A History of the Jews in Christian Spain*, 2 vol. (1961; orig. pub. in Hebrew, 1959), treats one of the principal elements in medieval society and makes use of a wealth of sources. A classic work delineating the life of the Cid and placing him in the context of the 11th century is RAMON MENENDEZ PIDAL, *El Cid en la historia* (1921; Eng. trans., *The Cid and His Spain*, 1934). ROBERT I. BURNS, *The Crusader Kingdom of Valencia*, 2 vol. (1967), is a useful study of the settlement of Valencia after the Christian reconquest.

Muslim Spain: The best work in English is W. MONTGOMERY WATT, *A History of Islamic Spain* (1965); see also E. LEVI-PROVENCAL, *Histoire de l'Espagne musulmane*, 2nd ed., 3 vol. (1950), and *L'Espagne musulmane au Xe siècle* (1932).

1479–1700: JOHN H. ELLIOTT, *Imperial Spain, 1469–1716* (1963), is the only single work covering precisely this period, and it is perhaps the best modern synthesis in any language. JOHN LYNCH, *Spain Under the Habsburgs*, 2 vol. (1964–69), is based on the latest research, especially in economic history. It has useful up-to-date bibliographies. ROGER B. MERRIMAN, *The Rise of the Spanish Empire in the Old World and in the New*, vol. 2–4 (1918–34), covers the period from the Catholic Kings to Philip II. It is still useful as the most detailed political history of the period. The last volume, on Philip II, is especially good. ANTONIO DOMINGUEZ ORTIZ, *Crisis y decadencia de la España de los Austrias* (1969; Eng. trans., *The Golden Age of Spain, 1516–1659*, 1971), is a synthesis by one of the most distinguished contemporary Spanish scholars. HELMUT G. KOENIGSBERGER, *The Habsburgs and Europe, 1516–1660* (1971), attempts to see Spanish history in the perspective of European history. It consists of chapters from *The New Cambridge Modern History*, vol. 2 and 3, and from *The Age of Expansion*, ed. by H.R. TREVOR-ROPER (1968). This latter volume also contains a fine short chapter by H. KAMEN, "Spain's Dominion," on the reigns of Philip II and Philip III. FERNAND BRAUDEL, *La Méditerranée et le monde méditerranéen à l'époque de Philippe II*, 2nd ed., 2 vol. (1966; Eng. trans. of vol. 1, *The Mediterranean and the Mediterranean World in the Age of Philip II*, 1972), is a large-scale economic and historical geography of the Mediterranean basin during the 16th century, in which Spain plays a central role. This book has become a classic since it first appeared (1949) and has been the basis for all further work on the period and area. EARL J. HAMILTON, *American Treasure and the Price Revolution in Spain, 1501–1650* (1934), is a pioneering study whose price series have not been superseded; however, its statistical methods and interpretations are open to criticism, and the tables should be used with caution. WILLIAM H. PRESCOTT, *History of the Reign of Ferdinand and Isabella the Catholic*, 3 vol. (1838, many later editions), is still the fullest and most readable account in English. J.M. BATISTA I ROCA, "The Hispanic Kingdoms and the Catholic Kings," *The New Cambridge Modern History*, vol. 1 (1957), is an excellent short modern account. HENRY C. LEA, *A History of the Inquisition of Spain*, 4 vol. (1906–07), is still the fullest, and a very fair, account. PAUL J. HAUBEN (ed.), *The Spanish In-*

quisition (1969), prints short extracts from authors with different viewpoints on this very controversial subject. JOSEPH PEREZ, *La Révolution des "comunidades" de Castile, 1520–1521* (1970), is the most detailed account of both the events and the social and political background of the Comunero revolt. GARRETT MATTINGLY, *The Defeat of the Spanish Armada* (U.S. title, *The Armada;* 1959), is outstanding both for its scholarship and for its readability. JOHN H. ELLIOTT, *The Revolt of the Catalans: A Study in the Decline of Spain, 1598–1640* (1963), is the definitive work on this subject. JOHN H. PARRY, *The Spanish Seaborne Empire* (1966), is an excellent introduction to this subject.

Spain since 1700: The best general introduction to the problems of modern Spain remains GERALD BRENAN, *The Spanish Labyrinth*, 2nd ed. (1950). For a general outline of political, economic, and social developments, see RAYMOND CARR, *Spain: 1808–1939* (1966). SALVADOR DE MADARIAGA, *Spain* (1958), is a stimulating general study by a liberal Spaniard. For the War of the Spanish Succession, see HENRY KAMEN, *The War of Succession in Spain, 1700–15* (1969). The internal history of the reign of Charles III is well covered by RICHARD HERR, in *The Eighteenth Century Revolution in Spain* (1958). PIERRE VILAR, *La Catalogne dans l'Espagne moderne*, 3 vol. (1962), is a masterly treatment of the late-18th-century social and economic history of Catalonia. EARL J. HAMILTON, *War and Prices in Spain, 1651–1800* (1947), is a basic study of the economy. ROBERT SOUTHEY, *History of the Peninsular War*, 3 vol. (1823–32), is still useful. The military history of the war (from the British point of view) is covered in CHARLES OMAN, *A History of the Peninsular War*, 7 vol. (1902–30). Napoleon's policy in Spain is studied in detail in ANDRE FUGIER, *Napoléon et l'Espagne, 1799–1808*, 2 vol. (1930). For the intervention of the army in politics, see ERIC CHRISTIANSEN, *The Origins of Military Power in Spain, 1800–1854* (1967); and STANLEY G. PAYNE, *Politics and the Military in Modern Spain* (1967). For midcentury politics see V.G. KIERNAN, *The Revolution of 1854 in Spanish History* (1966); and for the history of the republic, see C.A.M. HENNESSY, *The Federal Republic in Spain* (1962). There is no adequate study of Restoration Spain in English. ANTONIO RAMOS OLIVEIRA, *Politics, Economics and Men of Modern Spain, 1808–1946* (1946), is a useful guide. The best introduction is contained in M. FERNANDEZ ALMAGRO, *Historia política de la España contemporánea* (1956). A Centre Left account of the republic is GABRIEL JACKSON, *The Spanish Republic and the Civil War, 1931–39* (1965). An excellent study of the politics of the Right is RICHARD ROBINSON, *The Origins of Franco's Spain* (1970). The history of the Civil War is fully treated in HUGH THOMAS, *The Spanish Civil War* (1961). For the internal history of the republic, BURNETT BOLLOTEN, *The Grand Camouflage* (1968), is excellent. FRANZ BORKENAU, *The Spanish Cockpit* (1937); and GEORGE ORWELL, *Homage to Catalonia* (1938, reprinted 1969), are splendid contemporary accounts. The politics of Nationalist Spain are treated in STANLEY G. PAYNE, *Falange* (1961). For the development of the Franco régime see J.W.D. TRYTHALL, *Franco* (1970); and GEORGE HILLS, *Franco* (1967).

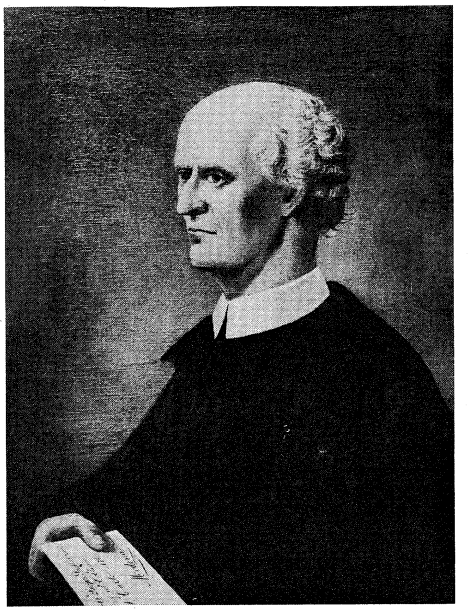
(A.G.yB./J.F.O.C./J.V.G./H.G.K./R.Ca.)

Spallanzani, Lazzaro

One of the great 18th-century physiologists, Lazzaro Spallanzani made important contributions to the experimental study of sense organs, regeneration, digestion, respiration, and animal reproduction. His investigations into the development of microscopic life in nutrient culture solutions paved the way for the research of Louis Pasteur. He also found time to make studies of physical phenomena, such as volcanoes, springs, the weather, and the depth of Lake Ventasso.

Spallanzani was born on January 12, 1729, at Scandiano, Modena. His father, a distinguished lawyer, sent him to the Jesuit college at Reggio, at which he received a sound education in the classics and philosophy. He was invited to join the order, but, although he was eventually ordained (in 1757), he declined this offer and went to Bologna to study law. Under the influence of his kinswoman Laura Bassi, a professor of mathematics, he became interested in science. In 1754 Spallanzani was appointed professor of logic, metaphysics, and Greek at Reggio College and in 1760 professor of physics at the University of Modena.

Although Spallanzani published in 1760 an article critical of a new translation of the *Iliad,* all of his leisure was being devoted to scientific research. In 1766 he published a monograph on the mechanics of stones that bounce when thrown obliquely across water. His first biological work, published in 1767, was an attack on the biological theory suggested by Georges Buffon and John Turberville Needham, who believed that all living things contain, in addition to inanimate matter, special "vital atoms" that are responsible for all physiological activities. They postulated that, after death, the "vital atoms" escape into the soil and are again taken up by plants. The two men claimed that the small moving objects seen in pond water and in infusions of plant and animal matter are not living organisms but merely "vital atoms" escaping from the organic material. Spallanzani studied various forms of microscopic life and confirmed the view of Antonie van Leeuwenhoek that such forms are living organisms. In a series of experiments he showed that gravy, when boiled, did not produce these forms if placed in phials that were immediately sealed by fusing the glass. As a result of this work, he concluded that the objects in pond water and other preparations were living organisms introduced from the air and that Buffon's views were without foundation.



Spallanzani, oil painting by an unknown artist, probably early 18th century. In the collection of the Università degli Studi di Pavia, Italy.
By courtesy of the Universita degli Studi di Pavia, Italy

Experiments on spontaneous generation

The range of Spallanzani's experimental interest expanded. In 1768 he published the results of his regeneration and transplantation experiments. He studied regeneration in a wide range of animals including planarians, snails, and amphibians and reached a number of general conclusions: the lower animals have greater regenerative power than the higher; young individuals have a greater capacity for regeneration than the adults of the same species; and, except in the simplest animals, it is the superficial parts not the internal organs that can regenerate. His transplantation experiments showed great experimental skill and included the successful transplant of the head of one snail onto the body of another. In 1773 he investigated the circulation of the blood through the lungs and other organs and did an important series of experiments on digestion, in which he obtained evidence that digestive juice contains special chemicals that are suited to particular foods. At the request of his friend Charles Bonnet, Spallanzani investigated the male contribution to generation. Although the spermatozoa had first been seen in the 17th century, their function was not understood until some 30 years after the formulation of the cell theory in 1839. As a result of his earlier investigations into simple animals, Spallanzani supported the prevailing view that the spermatozoa were parasites within the semen. Both Bonnet and Spallanzani accepted

Experiments on regeneration, transplantation, and digestion

Preformation and the male contribution to reproduction

the Preformation theory. According to their version of this theory, the germs of all living things were created by God in the beginning and were encapsulated within the first female of each species. Thus, the new individual present in each egg was not formed *de novo* but developed as the result of an expansion of parts the delineation of which had been laid down within the germ by God at the Creation. It was assumed that the semen provided a stimulus for this expansion, but it was not known if contact was essential nor if all the parts of the semen were required. Using amphibians, Spallanzani showed that actual contact between egg and semen is essential for the development of a new animal and that filtered semen becomes less and less effective as filtration becomes more and more complete. He noted that the residue on the filter paper retained all its original power if it were immediately added to the water containing the eggs. Spallanzani concluded that it was the solid parts of the secretion, proteinaceous and fatty substances that form the bulk of the semen, that were essential, and he continued to regard the spermatzoa as inessential parasites. Despite this error, Spallanzani performed some of the first successful artificial insemination experiments on lower animals and on a dog.

As Spallanzani's fame grew, he became a fellow of most of the scientific societies in Europe. In 1769 he accepted a chair at the University of Pavia, where, despite other offers, he remained for the rest of his life. He was popular with students and colleagues. Once a small group, jealous of his success, accused him of malpractice in association with the museum that he controlled, but he was soon vindicated. Spallanzani took every opportunity to travel, to study new phenomena, and to meet other scientists. The accounts of his journeys to Constantinople and Sicily still provide interesting reading. Toward the end of his life he conducted further research on microscopic animals and plants that he had started early in his career; he also began studies on the electric charge of the torpedo fish and sense organs in bats. In his last set of experiments, published posthumously, he attempted to show that the conversion of oxygen to carbon dioxide must occur in tissues, not in the lungs (as Antoine-Laurent Lavoisier had suggested in 1787). He died in Pavia in 1799.

BIBLIOGRAPHY. Most general histories of biology have a short section on Spallanzani. MICHAEL FOSTER, *Lectures on the History of Physiology During the Sixteenth, Seventeenth and Eighteenth Centuries* (1901, reprinted 1924), deals with Spallanzani's work on digestion and respiration. Two books by ELIZABETH B. GASKING are *Investigations into Generation, 1651–1828* (1967), which discusses Spallanzani's researches into regeneration, infusions, and his experiments on fertilization, and *The Rise of Experimental Embryology* (1970), a general history, with an account of Spallanzani's experiments and comments on the Preformation theory. Spallanzani's own works in translation are *Tracts on the Natural History of Animals and Vegetables*, trans. by J. DALYELL (1803); and *An Essay upon Animal Reproductions*, trans. by M. MATY (1769).

(E.B.G.)

Special-Interest Groups

In its widest definition, the term special-interest group (or its synonym interest group) refers to any aggregate of individuals that, on the basis of one or more shared concerns or wants, makes claims upon other groups in society or upon society in general in order to maintain or promote its position or objectives. Special-interest groups can be more specifically described and classified in terms of motivation, structure, and political activity.

Promotional groups

Categories defined by motivation. Within the broad definition of special-interest groups, two polar types are recognizable: first, interest groups proper, such as trade unions, farmers' unions, and employers' associations, which have as their primary purpose the enhancement of the advantage of their members; and, second, promotional groups, such as the societies for the prevention of cruelty to children or various voluntary relief agencies, which exist primarily or entirely to enhance the advantage not of their own members but of the population, even perhaps to the discomfort or disadvantage of their own members. Some of these, such as churches or various evangelizing groups, exist to promulgate a distinctive set of values to be applied to society as a whole. The distinction between interest groups proper and promotional groups is not sharp for two reasons. In the first place, most interest groups proper sincerely believe that in furthering their own material advantage they are also serving that of society as a whole—by promoting "free enterprise" or a healthy and wealthy body of farmers or a well-paid and enthusiastic corps of school teachers. But the propagation of such beliefs is certainly not the primary purpose of such organizations, thus distinguishing them from promotional groups. The second reason is that some groups—for instance, organized churches—fall between the two types; they can simultaneously pursue advantages for their own sect and seek to inculcate a distinctive set of values in a whole society.

Natural and associational groups

Categories defined by structure. In primitive or developing societies, the most prominent type of interest group is the natural (*i.e.*, primordial or communal) one—that is, one based on kinship, lineage, neighbourhood, or religious confession. In Western, industrialized societies, though such groups do sometimes retain influence (as do the nationalities in Yugoslavia or Czechoslovakia), the most prominent interest group is the associational (*i.e.*, secondary or factitious) type, like the trade union or Campaign for Nuclear Disarmament, which is deliberately created to serve defined purposes. Often, associational (or factitious) groups are created for the sake of the specialized purposes of the first type; the Indian Workers Union, for instance, was created from the community of Indian persons in Britain.

Within the class of associational groups, further qualifications must be made:

1. Not all associational groups possess formal structure. "Wall Street" in the United States, or the "City" in London, though consisting of a loose network of persons or functions, may nevertheless exert powerful collective pressures.
2. Some associational groups are collectivities; but others are single, discreet organizations. Thus General Motors Corporation exerts its own strong influence vis-à-vis the U.S. government irrespective of the influence of the more general employers' organization to which it possibly belongs, like the National Association of Manufacturers.
3. Some associational groups are temporary or *ad hoc*. Such are the so-called anomic groups, such as enraged French farmers who come together briefly to put logs across roads in order to draw governmental attention to their grievances.
4. Some associational groups are "latent." There may be a common interest among certain members of the public even though these individuals may not have combined into a formal or informal organization. The individuals may be unaware that they have a common interest or, if aware, see no reason to defend or promote it. Or, even if the members consciously wish to defend or promote an interest, the laws may restrict or control their ability to associate for such a purpose.

Pressure groups

Categories defined by political activity. If or when any interest group tries to influence the government in the pursuit of its aims, it becomes a "pressure group." Not all interest groups proper try to exert influence on governments; many (such as businesses and trade unions) do so only as a part of their more general activities of promoting their own interests. Promotional groups, on the other hand, spend much or all their time precisely in trying to influence the government to favour their aims. Groups such as an anglers' association may, therefore, turn into pressure groups for a time—the time when they seek to influence the government for one of their purposes—and then revert to simple interest-group status. A pressure group is therefore definable as "any interest group that is not a part of the government and does not itself seek to govern the country in its own name, but does seek to influence that government for its own purposes."

Two difficulties arise. First, some groups are neither a

governmental agency nor entirely a private group. In autocratic states, Communist or otherwise, for instance, trade unions are often controlled by the government or the governing party. In the Western liberal democracies, again, agencies like the Tennessee Valley Authority (in the U.S.) or the British Broadcasting Corporation, though subject to the overriding control of the government, enjoy substantial autonomy in certain brand areas.

The second difficulty lies in distinguishing a pressure group from a political party. The most fruitful approach is to stipulate that, whereas a party puts up candidates for election to public office and has a structure and carries out activities to secure their election, a pressure group seeks to influence both the government and the parties. The Catholic interest in Italy, for instance, is served both by the Christian Democratic Party and also by the Catholic Action Society. The first is unquestionably a party with its own program, mode of operation, and structures; and the second is unquestionably a pressure group. Again the "natural" group of Swedes in Finland serves itself by the Swedish People's Party; this is clearly a special-interest group, yet it puts up candidates and is represented in the legislature because it thinks its members' interest is best served by exerting pressure as part of a governmental coalition.

Thus, it may be that what begins as a pressure group winds up as a political party. The British Labour Party had its origins in 1900, when a collection of trade unionists formed a Labour Representation Committee in order to secure the election of working-class MP's to Parliament; later it broadened its membership and program to become the Labour Party of today. Again, in France, Pierre Poujade's pressure group, the Union of Small Shopkeepers, was involved in the 1956 elections and won over 40 seats, thus becoming a political party before retiring from the electoral arena in 1958 and reverting to pressure-group status. The reverse process can also occur, as when a political party engenders pressure groups; thus, the Communist parties in France and Italy have set up interest groups for women, youth, workers, and the like.

CHARACTERISTICS OF PRESSURE GROUPS

Power elements. Even in the most pluralistic of the Western liberal democratic states, it is wrong to picture public policy as simply the result of a parallelogram of group forces. The groups are themselves restrained by institutions, procedures, and public beliefs. Among the *institutions*, the most important are the public bureaucracy, the political parties, and the independent branches of the government (the executive, legislature, and judiciary —or some combination of these, depending on the structure of government). To the extent that each or all of these institutions have strong traditions, no interest group supplants them but instead has to deal with them. The governmental procedures in force also affect groups' performance. To the extent that public issues are traditionally subject to publicity and wide discussion, groups are limited in the kind of activities that they can pursue. They are constrained by public *beliefs* as to what is in "the public interest," by what is a proper or improper procedure, or by what causes are respectable and what are not. If any of these three entities—institutions, procedures, or public beliefs—become feeble, interest groups can intrude into them. By allying with or gaining power over the parties, the legislature, or the bureaucracy, for instance, groups can tend to substitute themselves for such organs in the decision-making process.

Inside every country the power that pressure groups wield against the government and that each group wields relative to its rivals depends on a combination of the following variables, some of which are mutually dependent:

1. Density is the ratio of actual membership to potential membership. The higher the density, the more "representative" is the group of all the people whose interests it purports to represent and the greater will be the inclination for governments to recognize and consult it. On the other hand, too wide a membership brings with it internal cleavages, so that on certain issues the organization may not be able to tender clear-cut advice or, indeed, may have to refrain from offering advice altogether.
2. Wealth not only helps a group to develop a skilled management and bureaucracy but also enables it to propagandize and to finance political parties.
3. Prestige lends a group the ability to get a favourable initial hearing from government and public, even though its numbers may be small and its wealth meagre.
4. Organization consists notably of the ability to brief legislators and administrators well and quickly, to mobilize members and the public rapidly, and to receive advance intelligence of likely trends in policy.
5. Socioeconomic leverage is strong among some groups (such as trade unions) that can disrupt social life and low among others (such as consumers associations).
6. Militancy consists of making an effective nuisance of oneself, so that, hopefully, governments will be willing to "buy" time or peace.
7. Specialized information and skills add the weight of authority.
8. Electoral strength refers to the power of some groups that, though poorly organized or having a low density, nevertheless command wide support in the electorate and so find themselves courted by rival political parties.

Strategy and tactics. Although all groups seek out the most influential agencies of government, the most influential may not be accessible; and resort must be had to the next most influential organ that *is* accessible. Because the Communist trade unions in Italy, for instance, are not welcome at the Roman Catholic controlled Ministry of Labour, they try to penetrate the legislature through political election. Indeed, in any system the influence of the various organs of government varies. In France the president and bureaucracy are more influential than the majority party and Parliament; in the United States, the president and Congress balance each other, with control of the bureaucracy shared between the two. Strategy, for an interest group, consists of determining and going to the most influential organ that is accessible.

Tactics to affect the executive

The tactics of an interest group range from the constitutional to the unconstitutional (or "direct action"), the legal to the illegal—all subject, as noted above, to the restraints of institutions, public procedures, and public beliefs. When the target of the tactics is the executive and the governmental bureaucracy, the following are the prime methods: *Constitutional operations* include (1) advice via advisory bodies attached to the ministries (as in Britain and France), (2) official hearings outside the ministries (as in Royal Commissions in the U.K. and presidential commissions in the U.S.) or inside the ministries (as in Germany or Sweden), and (3) ad hoc consultations and cooperation. *Semiconstitutional operations* include (1) recruitment of state civil servants into the private bureaucracies of the interest groups, or vice versa, and (2) bribery and favours, which need not necessarily mean passing of money but can take the form of providing entertainment, gifts, or meals. *Semidirect operations* include such measures as refusing to provide information or to cooperate in administering legislation and thus generally withdrawing from advisory functions. *Direct action* includes such measures as the withdrawal of labour (the British Medical Association, for instance, advised its members to withdraw from the state medical service in 1965 unless its demands were met) or even violence such as the roadblocking activities of enraged French farmers in the 1960s).

Tactics to affect the legislative

When the target is the legislature, the following are the chief kinds of tactics used by interest groups: *Constitutional operations* include (1) testimony before legislative committee hearings (extremely important in the U.S. and somewhat important in Sweden, the German Federal Republic, Japan, and Italy); (2) direct representation in the legislature, which involves having individual businessmen, financiers, farmers, school teachers, journalists, and other professionals representing the larger interests of their profession or sponsoring group (in Great Britain in 1955 some 30 percent of Labour's MP's were trade-union sponsored); and (3) lobbying, which refers to the activities of salaried or nonsalaried persons who try to promote their interest groups' aims by seeking personal contacts with

legislators, by sending communications or information to legislators or persuading others to do so, and by rendering campaign assistance to favoured legislators. (In the U.S., lobbying is conducted not only by private interest groups but also by certain members of the executive branch and the bureaucracy.) *Semiconstitutional operations* include subvention or payment of retainers as well as providing campaign contributions, secretarial help, office accommodations, and the like, which are restricted but not forbidden by law. *Semidirect action* includes mass lobbying, such as promoting mass demonstrations before legislators, ostensibly to relate grievances but in fact to impress public opinion. *Direct action* includes threatening legislators with a withdrawal of financial assistance, with physical violence, or with riots (as in Japan, where snake dances and riots become a means of protest).

When political parties are powerful, they may prove a highly preferred target. Tactics in such cases include the following: *Constitutional operations*, such as, (1) exacting election pledges, which is most important and fruitful in countries (such as the U.S.) where there is relatively little national-party discipline, where representatives owe a certain allegiance to local-party groups and constituents, and where, most important, representatives do possess legislative power; (2) financing the party, which in all systems is perhaps the most effective operation and which in most Western parliaments—and in Western-style parliaments like Japan's—is systematically done by trade unions, employers associations, and individual firms; and (3) lobbying the party machinery, a method that is virtually nonexistent in the United States but is significant in such countries as Great Britain and Germany.

Tactics to affect political parties

A final target of interest groups is public opinion. The general opinion of the public at large can be of importance in issues where questions of "fairness" or "justice" arise (on which most people feel they can pronounce). These questions involve such matters as capital punishment, the use of nuclear weapons, or certain wars, about which large numbers of people feel very strongly.

ROLE OF INTEREST GROUPS IN VARIOUS POLITICAL SYSTEMS

The Anglo-American and Swedish patterns. In the United States, Great Britain, and Sweden, the general population is not irreparably divided on ideological or cultural lines. Although there arise from time to time some small extremist groups, the population at large is not split into irreconcilable blocks. Given this situation, the most influential and prominent groups are "secondary" associations that have highly specialized objectives, form freely without the requirement of prior governmental permission, and carry out their activities autonomously. With the exception of the affiliation of the trade-union movement to the Labour Party in Britain and in Sweden, the groups are politically neutral; and even the alignment of British and Swedish trade unions and cooperatives does not prevent them from cooperating with other parties nor hinder their free access to government departments. Latent groups are few, since freedom of speech and association and the right to petition the government permit most interests to organize, while good communication media facilitate such organization. Public participation in such groups is, by world standards, high. The pressure groups play highly specialized roles and are sharply differentiated from the political parties.

National differences

Although American and British means of influencing the executive are similar (by advisory committees, special inquiries, and day-to-day consultation and mutual assistance), these relations are more insitutionalized in Britain than in the U.S. British groups seek to influence the parties by affiliation or by securing their members' nomination as candidates, neither mode being employed in the U.S., and by financial assistance, which occurs in both states but in quite different ways owing to the very different legislation on this matter in the two countries. British groups seek to influence the legislature by securing direct representation through participation in party caucuses and by promoting amendments in the committee stage. In the U.S. the methods employed consist of lobbying influential congressmen and appearing at committee hearings, which are far more important in Congress than in the House of Commons. In influencing public opinion, American groups spend far more than their British counterparts—one reason being that purchase of time on radio and television networks is permissible in the U.S., whereas in Britain it is not. In addition, access to the presidency can prove decisive in the U.S.; and some groups such as the National Association for the Advancement of Colored People have scored notable successes by seeking court action.

In Sweden the tactics are somewhat different, owing to three factors: the existence of a multiparty system and the existence of coalition cabinets, the structure of the bureaucracy, and the much greater institutionalization of pressure groups. Sweden, Denmark, and Norway each possess a multiparty system; but when coalitions are formed, they are highly stable. Consensus is achieved in a three-tier operation—first at interparty (electoral) level, then at the interparty-coalition (legislative) level, and finally at the bureaucratic level. In Sweden groups seek to influence the parties by affiliation (the trade unions being affiliated to the Labour Party), by financing them (trade unions giving to the Labour Party, employers groups to the antisocialist parties), and by securing direct representation through the nomination of members as parliamentary candidates. The parties respond by trying to compose a balanced ticket of the spectrum of interests. Hence the groups can affect the legislature through their representatives inside each of the parties and may, on occasion, form a cross-party legislative pressure group in defense of their common interest; and they also seek access to the party caucus and influence the committee stage of legislation. In the latter case they are assisted by the standard Swedish procedure by which the "comments" of outside interests are attached to the government's bill, even if they contradict the bill's purposes. Groups have ready access to the legislature in the same ways that Anglo-American groups do, except that the procedure is highly institutionalized; the "Royal Commission" with representatives of the bureaucracy, the parties, and the interest groups seek to reach initial agreement on a bill to be presented to the legislature; and when this device is not used, "comments" appended to the bill serve to put the measure in the context of the avowed interests of various groups.

Not unlike the system in Britain, the U.S., and Sweden is the system in Japan, which, despite sharp ideological divisions, has a strong and stable two-party arrangement. This has led to the emergence of a pattern of highly institutionalized pressure groups of the "secondary" type, the most influential of these being business and organized labour. Business, on the whole, is aligned with the conservative Liberal-Democratic party, supplies the bulk of its funds, and has a large say in policy formation. Organized labour constitutes a powerful element within the Japanese Socialist Party. In an earlier day, rural farm groups carried considerable political weight—for the conservatives—but the increasing industrialization and urbanization of the country have tended to reduce its importance.

The French and Italian patterns. In France and Italy the political pattern is characterized by deep ideological cleavages. The Roman Catholic Church, for example, operates through both its own political party and its own interest groups, such as the Catholic Action Society and Catholic youth and women's group. By the same token, the highly ideological and sectarian Commmunist Party establishes its own pressure groups, such as trade unions, farmers associations, and youth movements. Several rival groups thus compete for the same clientele aligned or affiliated with rival political parties.

Under France's Fifth Republic, the executive branch currently controls the majority party in the legislature, and the role of the legislature has been reduced; whereas, under Italy's current system, feeble and fragile coalition cabinets have weakened executive leadership of the legislature, and the legislature itself is highly fragmented among numerous parties. Because contacts between interest groups on the one hand and the parties, civil service departments, and legislature on the other are poorly insti-

tutionalized, semidirect action or even direct action against the executive and the legislature are more common, and lawless activities of crowds and violent organizations have proved more widespread.

The means of influencing the chosen targets also vary. In France attention has shifted from the legislature (under the Fourth Republic) to the president and bureaucracy (under the Fifth); interest groups now act through the numerous advisory bodies, private contacts, and such official institutions as the Economic and Social Council. Interest groups secure contacts by putting former civil servants on their own payrolls. The fact that many civil servants and business executives come from the same schools facilitates private collusion.

In Italy interest groups direct their attention mainly to three targets. First, the bureaucracy, even though it is highly fragmented and provides few formal advisory channels of communication, is open to ad hoc and personal consultations (except to groups deemed ideologically distasteful, such as the Communists). Second, the parties are often twinned to interest groups, as noted earlier; and thus the groups seek and achieve direct representation in the legislature. Third, the legislature, because it is composed of interest groups, is heavily influenced in specialized legislative committees. In general, therefore, interest groups have tended to "intrude" into the organs of government, forming close alliances with certain departments and legislative commissions and inserting themselves into others via their membership in political parties. In short, interest groups are not "contained" as they are in Anglo-American and Scandinavian countries—political institutions are fragmented and weak, procedures are irregular or personal and not formal, and public beliefs are polarized into uncompromising ideologies.

Patterns in developing countries. Although there is no one typical pattern for all underdeveloped countries, which of course embrace over two-thirds of the world's sovereign states, there are a few common characteristics that can be suggested. First, political culture in each of these countries is fragmented, sometimes because of ideological differences, sometimes because of the mutual hostility of "natural" groupings based on kinship, lineage, tribe, or language. Second, nascent interest groups representing new modernizing and industrializing forces tend to range against rural natural groupings. The modernizers have defined and specialized objectives, whereas the traditionalists usually represent a spectrum of attitudes, beliefs, values, and interests. Both, in any case, tend to sponsor or generate their own political parties. Third, during preindependence days in some countries, trade unions and other movements (such as the Somali Youth League) carried on activities that would otherwise have been carried on by political parties; when independence came, these groups continued such activities, thus blurring the functional distinctions between parties and interest groups. Fourth, associations in these countries tend to have low membership densities; and in most of these countries the principal "interest"—that is, the peasantry—is hardly organized at all. Attempts to develop widely represented peasant movements have had only limited success. Fifth, although organization and established procedures may be poor, latent interest is usually high. In a particularly divisive and administratively feeble country, this interest latency may reveal itself in lawless and violent group activity—such as strikes, riots, guerrilla movements, and assassinations.

Countries with one-party or military governments

In some countries "parties" in elections are not really parties but simply electoral coalitions of various interest groups—the most influential groups consisting of cliques of landlords; leaders of religious, ethnic, or linguistic groups; and the like. Legislatures tend to reflect these divisions and temporary arrangements in society and so enjoy only feeble authority.

In countries with single-party regimes, such as Kenya, Tanzania, and Tunisia, the parties tend to lose their mass following and fairly well-defined organization in the years following independence. (A notable exception, however, is India with its Congress Party.) To enhance their authority, the party leaders try to limit or abolish the autonomy of such factitious interest groups as trade unions, cooperative movements, and youth and women movements and try to turn these groups into ancillaries of the party.

In some countries the military has taken over, either alone or with civilian elements, and imposed an authoritarian rule. In such cases, the most prominent groups are governmental—the military and the executive. Unless factitious interest groups enjoy a long history of organization (as does the Catholic hierarchy in Latin America), they are often turned into governmental ancillaries or suppressed. In these countries ruled by the military or single parties, the political process turns into a highly informal medley of pressures and personal contacts between the government, the military forces, the various "natural" groups, and some of the more strategically placed factitious groups (such as the church hierarchy or private industry). The "system" is constantly threatened with government coercion or violent outbreaks by sections of the population.

Patterns in Communist states. Like Western states, but unlike developing countries, Communist countries contain various associational groups (trade unions, cooperatives, and the like) but few, if any, promotional groups. In addition, such countries as the U.S.S.R., Czechoslovakia, Yugoslavia, Romania, and the People's Republic of China contain many ethnic groups; and Poland, Hungary, and East Germany contain influential organized churches. But unlike the Western states and like a good many of the developing countries, Communist countries grant such groups very little autonomy. The government is firmly controlled by a single party, is highly centralized, and tries to control a great deal of life. The Communist Party may allow other parties to exist in coalition under conditions that the party itself lays down; in Poland, for instance, the controlling party is the (Communist) United Polish Workers' Party, and it permits the peasant interest to be represented by the minor United Peasant Party and the professions by the even smaller Democratic Party. The leading party assures its power, however, by controlling the nomination (in all but minor cases tantamount to election) of the leadership of all minor parties and interest groups, and the only great interest group that is excluded is the leadership of the Catholic Church. In any event, no interest group or party can be legally constituted in a Communist country without the express permission of the authorities. And even when an interest is allowed to formally organize, it rarely can exert any influence on the government, for governmental leaders, having been nominated by the Communist Party, are responsible to the party alone, not to various interests.

One consequence of these restrictions is a high degree of interest latency in certain sectors; this latency is attested by the fact that within six months of the fall of Antonín Novotný in Czechoslovakia in January 1968 over 70 groups were petitioning the government for legalization.

In the U.S.S.R.

In the U.S.S.R. there exist large trade unions, cooperative associations, and professional societies; many scientific, defense, and sports societies, as well as some 100,000 house committees, sanitary teams, parent-teacher associations, and the like. All are party led and controlled; and they purportedly serve, in Stalin's words, as "transmission belts from the party to the masses." One must not infer, however, that such associations are entirely without other significance. In the first place, the party is not identical with the state bureaucracy, though it controls the latter; hence divergences and collisions can occur between an interest group (such as the trade unions) and a concerned ministry regarding the interpretation or implementation of a party's directive. The result often is a triangular meeting of interest group, officials, and party functionaries, with the latter mediating. Second, the constitution provides that these so-called mass organizations should be consulted on selected issues, as they indeed were in regard to the education act of 1958.

In other Communist states

Poland's groups include such economic interests as the trade unions and the Peasant Self-Aid and such professional ones as the General Technical Organization and the Writers' Union, as well as the Socialist Youth League; the Rural Youth League; the Women's League; and vari-

ous cultural-ethnic associations for Ukrainian, Lithuanian, or German minorities. In all of them, party leadership is assured (with minor exceptions), and they have relations with the bureaucracy and party mediators similar to those in the U.S.S.R. In Poland, however, there is the exceptional role of the Catholic Church, which has established some kind of equipoise with the party and has its own press, organizations, and clubs, even though its electoral freedom is severely restricted (only certain pro-Socialist Catholics are allowed to form "parties," such as the Znak, Pax, and the Christian Socialist Association, which participate in elections on the joint list with the Communist and other parties).

Yugoslavia has gone furthest in recognizing interest groups. The society contains many conflicting interests—nationalities; religious communities; a free-enterprise peasant sector; a decentralized network of workers' councils; and a large number of cultural, economic, and professional associations. The 1963 constitution gives institutional expression to these groups: in the five-chamber legislature, legislation is processed through the Federal Chamber, representing the population at large and individual republics; the Politico-Economic Chamber, representing management agencies and sociopolitical organizations; and the Chambers of Economics, of Education, and of Social Welfare and Health, representing their respective fields. The current factionalism inside the ruling "National Alliance," the centrifugal pulls of the peasant sector and the workers' councils, and the rivalry of the nationalities have combined to make these legal provisions much more than a formality.

BIBLIOGRAPHY. G.A. ALMOND and J.S. COLEMAN (eds.), *The Politics of the Developing Areas* (1960), a synoptic view of government in Latin America, Asia, and Africa that places special-interest groups in the setting of the politics of developing countries; K. VON BEHME, *Interessengruppen in der Demokratie* (1969), a classification of the types, strategies, and tactics of groups in the Western democratic states; A.F. BENTLEY, *The Process of Government*, ed. by P.H. ODEGARD (1967, orig. pub. in 1908), the major theoretical source for all subsequent Western democratic writing, advocating the theory by which all political activity is reduced to group activity; D.C. BLAISDELL, *American Democracy under Pressure* (1957), one of the best of many surveys of the role of groups in American society; S.E. FINER, *Anonymous Empire: A Study of the Lobby in Great Britain*, 2nd ed. (1966), a brief analysis of the types, strategies, and tactics of groups in Britain; H.W. EHRMANN, *Organized Business in France* (1957), a highly detailed account of the typical methods of groups in France, and (ed.), *Interest Groups on Four Continents* (1958), a collection of papers, together with summaries of the discussions based upon them, that is one of the best country-by-country accounts of the operation of groups; J. MEYNAUD, *Les Groupes de pression en France* (1958), a detailed, acute analysis of the entire field of group activity in France; J.G. LA PALOMBARA, *Interest Groups in Italian Politics* (1964), a highly detailed analysis, much of it based upon interview and behavioral data.

(S.E.F.)

Species and Speciation

Species are groups of organisms sharing many traits, or characteristics, in common; some are obvious, others subtle. Species come into being through the process of speciation. Most people recognize the distinctness of at least a few of the birds, fishes, insects, and plants that inhabit their immediate environment, and such knowledge is often extensive among so-called primitive peoples, who may be familiar with hundreds of kinds of plants and animals upon which they depend for food, drugs, and clothing or which they learn to avoid because of their stings, poisons, or distastefulness. The seemingly universal ability to recognize different species of organisms strongly suggests the reality of species distinctions.

THE DISTINCTIVENESS OF SPECIES

The readily observable distinctiveness of most species results from the fact that individuals belonging to different species seldom mate with one another, and the characteristics of one species are, therefore, seldom transferred to another. At the same time, the characteristics of each species result from the fact that the young inherit from their parents genes that determine their characteristics. Offspring differ in many ways from their parents because there are many ways in which the parental genes can be shuffled, or recombined, in the offspring. Characteristics essential to the survival of the individual, however, vary little between parent and offspring. In any event, the differences by which species are distinguished are generally characteristics common to all members of one species but not shared by individuals of a related or similar species. Common differences between species of birds may involve the shape of the feet or bill or the form of the wings. The advantage of a given shape is often obvious from the role played by the bird in its environment—a foot suitable for attacking and grasping prey may not be suitable for swimming or wading. Similarly, a bill useful for crushing plant seeds may be useless in extracting insect larvae from a tree. But not all differences between species are so obviously adaptive. Certain species of lizards are distinguished by the presence of enlarged scales along the back. The reason such scales should be advantageous to members of one species but not to those of another is not immediately clear. Apparently trivial differences often are discovered to be important for the species; for example, differences between the colours of the iris of the eyes in some species of gulls ensures that mating takes place only between members of the same species.

Many seemingly unimportant characteristics are readily observed in plants. Species may differ in texture, shape, or size of leaves. Some species produce seeds surrounded by a heavy, fleshy fruit; others shed seeds light enough to be blown by the wind. Many of these differences are understandable as adaptations that better ensure the survival and reproduction of the individual in a particular habitat.

A fuller understanding of the nature and origin of species is a focal point of research in systematic biology. It is also the considerable task of systematists to discover, name, and classify the multitude of plant and animal species. For some groups, such as birds, the task is by now essentially complete, and systematists speak with assurance about the total numbers of species in the world; in other groups, such as the insects, there may yet be thousands of unnamed species.

The task of naming species

In order to handle the task of naming and classifying species, a relatively simple system of nomenclature has gradually developed and has been agreed upon internationally. Under this system each new species is described in a publication in a standard manner and given a name consisting of two parts—a generic name, which encompasses a number of closely related species, and a species name, which narrows the identification to one species. The name of the Gila monster, a poisonous lizard of the southwestern United States, for example, is *Heloderma suspectum*. A systematist in another part of the world, even with no knowledge of the English language, would readily recognize that this species is a close relative of the beaded lizard of Mexico, *Heloderma horridum*. In a similar fashion, genera that share a number of characteristics are placed in the same family, related families in the same order, orders in classes, and classes in phyla (or divisions; see CLASSIFICATION, BIOLOGICAL).

Species in nature are complex. Some may evolve very slowly, others rapidly. Some may interbreed and produce hybrid offspring, others at least show no evidence of interbreeding. Some may be quite different in one part of their range in comparison with another part, others remarkably similar over their entire range. Before considering these and other phenomena based upon the modern knowledge of species and speciation, it is well to review briefly the history of species concepts in order to place the subject and theory in perspective.

SPECIES CONCEPTS

Historical background. Historically, man's preoccupation with the recognition of species may be almost as old as man himself. Originally, as now, it was important to know which species were useful for what purpose.

An informal basis for recognition of species was sufficient for small regions or small groups of people; but as

scientists began to attempt the naming and classifying of all living and fossil species, it became clear that a more universal system was necessary. An agreed upon system of naming species was essential for communication among scientists, as was a system of classification based upon the relationships between species.

One of the first persons to suggest a formal arrangement of species, John Ray, a 17th-century English naturalist, proposed that species be recognized on the basis of the differences in structure that are transmitted from generation to generation. A great Swedish naturalist, Carolus Linnaeus (Carl von Linné), developed a simple system for naming plants and animals; with modifications, it is still in use today. It features a generic and a specific name for each organism and is therefore called binary (or binomial).

Linnaeus believed that species are fixed, or unchanging, and that each is the result of a separate act of creation. To him, therefore, a single individual organism could characterize an entire species. In describing species of snakes, for example, he might have recognized a species with 180 belly scales, another with 160. But the simple system began to break down with the discovery that the number of scales is variable even within one species. Nevertheless, Linnaeus' system of naming and grouping organisms into categories has been able to handle an enormous number of species discovered since his time.

It is important to note that the term related species had an entirely different meaning to biologists of the 17th and 18th centuries than it does in the 20th. At that time "related" meant simply sharing characteristics. Thus, limbless animals (sharing the characteristic of "no limbs")—entirely unrelated in an evolutionary sense—might have been grouped together. With the publication (1859) of Charles Darwin's *Origin of Species*, the term related acquired exact meaning: descent from a common ancestor, as a result of which many characteristics are shared. Species thus are related because they look alike, and they look alike because they share many genetic traits.

Darwin's view of species

With the publication of Darwin's work, which occurred before anything was known about heredity, it became necessary for man to change his view of species, most of all, of his own, for if species evolved as Darwin insisted, then they should be found to exist in varying intermediate stages in a continuous process of speciation. Furthermore, if they evolved by natural selection, variability must exist between individuals; for it is on variability that selection acts, and natural selection was the central core of Darwin's theories.

Darwin's research turned the attention of biologists to a detailed comparison of species for the purpose of determining their evolutionary relationship. The sharing of features by different species acquired new meaning and significance. Detailed studies of many species showed clearly that they are variable and, in fact, that some of the most important characteristics separating species are behavioral and physiological rather than anatomical.

The rediscovery (1900) of Gregor Johann Mendel's pioneering work on the heredity of peas clearly had implications for the species concept. Species differences came to be understood as inherited characteristics, and the genetic mechanism by which the presence or absence of a certain trait is manifested in a plant or animal came to be understood in a general way.

Currently, a species is recognized as a population of organisms that share a number of traits because, through interbreeding, such traits, as expressions of genes, appear in the offspring of matings between individuals of that species. Since the thousands of genes carried by each parent individual can be recombined in countless ways in the offspring, no two individuals will be genetically identical (except for those of identical multiple births, such as twins, triplets, etc.). Though each human individual is sufficiently different to be recognized as distinct, it is still easy to recognize that all human beings are members of a species because of the numerous common traits that exist between individuals. Outside a particular species, individuals have fewer genes in common and, therefore, resemble each other less. The sharing of many genes by individuals within a species, coupled with some degree of reproductive isolation from individuals of another species, is the basis of the concept of biological species that has resulted from a synthesis of genetics, systematics, and ecology.

Major concepts of the species. *The typological species.* Linnaeus and other early naturalists, who assumed that species are invariable in space and time, characterized the species by a single type or "normal" set of characteristics. Today a range of variation of distinctive characteristics within a species is recognized—such as four to eight spines per fin or 17 to 25 bristles per antenna. Strict adherence to a typological concept would require the recognition of many "species" within a single population of interbreeding members.

The morphological species. The concept of morphological species requires that one population must differ in some morphological characteristic from another population in order to be described as a new species. That the majority of species can probably be described in this way is evidenced by the fact that species named on the basis of morphological differences also fit the biological species concept (see below). The concept of typological species, while it is based on morphological differences, is distinct in denying variability within species. The morphological concept holds, instead, that variability is the rule rather than the exception.

Similarity of morphological and typological concepts

The nondimensional species. The idea of a nondimensional species differs from the morphological concept in emphasizing all differences between species at a given time and place. It is, accordingly, the working definition of the local naturalist. A species in a single area may differ in colour, structure, and behaviour from the same species in other areas. Each species, however, lacks the dimensions of space and time; in other words, there is no recognition in the concept that the same species at another locality may show differences in habitat and life history. The nondimensional species concept involves a particular population, or local group of individuals, of a species; the entire species range may encompass hundreds of such local populations.

The biological species. The biological species is what most biologists have in mind when they speak of species. This multidimensional concept recognizes that a species may be spread over a wide area and that in different areas it may be dissimilar in appearance, may breed at different times, or, in the case of migratory species, may leave and arrive at different times. The central concern is with what it is that produces and maintains the distinctiveness of each species over its range. Attention is therefore focussed on the mechanisms that thwart successful interbreeding. Although the biological species concept is difficult to apply to all species, it seems to be the most suitable way of regarding species. Throughout the remainder of this article, the term species refers, unless otherwise stated, to the biological species.

APPLICATION OF THE BIOLOGICAL SPECIES CONCEPT

The biological species concept has wide application to the majority of organisms. It focusses attention on the really important criterion of reproductive isolation, which is the mechanism by which species integrity is maintained. Because reproductive isolation is often brought about as a result of differences in behaviour, ecology, and genetics, much attention has been given to these characteristics; this has resulted in enrichment of the field of systematic biology and a better understanding of the nature of species.

It is useful to present the exceptions—examples of situations in which the concept is difficult to apply—for these have been the objects of intense study and have resulted in much new biological information.

Allopatric species. The biological species concept is difficult to apply to allopatric species—*i.e.*, populations that appear to be closely related but differ in enough characteristics to be called different species and are geographically isolated from one another. The barrier of geography prevents natural interbreeding, but sometimes

the two problematic groups can be artificially bred in the laboratory by taking eggs from the female of one "species" and fertilizing them with sperm from the male of another "species." Normal development of the hybrid offspring is justification for stating that the two "species" are closely related but does not provide enough evidence for considering them a single species. Even if they came into contact naturally, they might still behave as two species and not interbreed because of reproductive isolation, as a result of differences that preclude the production of normal offspring. Similar experiments with plants, in which one species is pollinated with pollen from the flower of another species, suggest little about the natural forces that might prevent cross-pollination or discourage germination of the hybrid seed.

The decision as to whether to consider two allopatric populations as species often can be made on a fairly reasonable basis. If two such populations are as different from one another as related species that are known not to interbreed, then they may be considered separate species.

Evidence from experiments

The failure of artificial crossing would be further confirmation of the presumption that the species are distinct; the two populations have indeed reached a point at which interbreeding would not be possible even if their ranges were rejoined. Similarly, if the two allopatric populations differ in characteristics that would make mating difficult should their ranges meet, they may be considered separate biological species. Differences in time of flowering, in mating calls, or in soil requirement might prohibit or greatly restrict interbreeding. Generally then, the biological species concept has been extended to allopatric populations by emphasizing that a species is a group of populations that actually—or potentially—exchange genes and that are reproductively isolated from other groups of populations.

Hybridization. Hybridization between populations that are considered on all other grounds to be a species is a common occurrence in some frogs and toads, some freshwater fishes, and many plants. In some instances the hybrids that result are abnormal and soon die. In some cases the hybrid matures but is sterile, or incapable of reproduction. In still other instances the hybrid may mature and reproduce, but subsequent breeding with either parental species results in "hybrid breakdown," in which later generations of hybrids rather than the first suffer from gene incompatibility. In these instances of hybrid inviability, sterility, or breakdown there is effective reproductive isolation: little or no transfer of genes occurs between the parental types, and they are therefore distinct species even though interbreeding occurs.

There are instances, however, of interbreeding between organisms that are generally considered to be separate species and in which there appears to be fairly free gene exchange in the area of species overlap. In these cases the gene flow is limited to a narrow geographic band or is associated with unnatural disturbances that dissolve the barriers to interbreeding or provide habitats in which hybrids can survive; nevertheless, the entities still maintain their distinctiveness as two species over the vast majority of their ranges.

Speciation is a dynamic process, and there exist populations in all stages of becoming species. Some populations are completely isolated reproductively in zones of overlap, and there is no question about the status of these as species. Others hybridize but produce unsuccessful offspring, again raising no doubts about their status as species. Others interbreed rather freely in some areas but not in others, perhaps because the process of speciation has not been completed.

Variants within a single species. Variations within a species are common in both animals and plants. In some instances there may be distinct colour or pattern differences within a single population with no intermediates between the types. Such instances are called polymorphisms. Males and females of some species are so different in appearance that they were originally described as separate species. Other polymorphisms are not so evident and may involve such characteristics as chromosome structure and number. If such differences pose no barrier to successful interbreeding, then they offer no real difficulty to the application of the biological species concept.

Polymorphism

In many species, striking differences occur between populations in different parts of the geographic range. At the boundaries between these geographic groups—called races or subspecies in animals and varieties in plants—are zones of intergradation in which the characteristics of one group are replaced by those of another. In some instances it may be difficult to decide whether there is sufficient restriction of interbreeding in these intergradation zones to call the populations involved species.

Asexual species. If a species is defined as a group of organisms reproductively isolated from other groups but with freely interbreeding individual members, then every individual of an asexual species is potentially a new species. Asexual species of animals usually produce only females; in such species as bees the males are produced asexually, the females sexually. In other species there may be alternation between sexual and asexual stages of reproduction in the life cycle. Such an alternation of generations is the rule among plants.

Among plants, many types of asexual reproduction are grouped under the term apomixis. This category includes vegetative reproduction by the rooting of broken stems or specialized buds, self-pollination of a flower by pollen from the same flower, or development of the egg without fertilization. All such mechanisms tend to restrict recombination of genes to a greater or lesser extent, and the resulting offspring resemble the parent more closely than is true in unrestricted sexual reproduction.

The biological species concept cannot really be applied to organisms that reproduce asexually. In such cases comparative morphology of such species is relied upon to reach a decision concerning its status. Fortunately, most such species have characteristics that are shared by the majority of individuals within the species and are different from those of other asexual species.

Fossil species. It is difficult to apply the biological concept to species that have a time dimension. Living species exemplify various points in the evolutionary process. It is relatively easy, for example, to discern three separate living species; however, the distinctions between these species become fewer and fewer as the fossil record approaches the common ancestor from which they arose. They begin to lose their separate identities.

It is difficult to apply the biological species concept to fossil species in general. Often the species is represented only by fragments, but even given complete fossil material the species can be defined only on a morphological basis. If fossil species show striking structural differences, however, it is generally a safe presumption that they were incapable of interbreeding. Were this not the case it is unlikely that the differences would have been maintained through the course of evolution.

THE PROPERTIES OF SPECIES

Distinctiveness. The ability of sexes to recognize mates within their own species and to distinguish them from sexes of another species is exceedingly important; an error may produce sterile or defective offspring.

The importance of mate recognition

Isolating mechanisms are employed to decrease the likelihood of such untoward events. Fireflies use the flash as a means of species identification; different species differ in the timing of the flash signal produced. In frogs and toads, the distinctive call of the males attracts the females to the proper mate. In some animals, such as moths, mate attraction depends upon volatile scents produced by another individual of the same species. In some fishes, the pattern of movements made by the male during courtship allows the female to discriminate between a male of her own species and one of another.

Isolating mechanisms in plants are less obvious. Differences in flower structure between related species, however, may serve the same function as do differences in mating behaviour between animal species. Insect pollinators transfer pollen grains from the male reproductive organs of one plant to the female organs of another plant of the same species. Differences between related plant species in flower colour, shape, and fragrance may result

in the attraction of different pollinators and reduce the possiblity that an insect may carry pollen to the "wrong species." Even if such unproductive transfers of pollen do occur, the reproductive organs differ sufficiently between the two plant species to preclude pollen germination.

Great emphasis has been placed on the necessity that the female discriminate between potential mates. Correct species identification is not usually equally important for the male. He generally can produce enormous numbers of sperm, and thus he loses little as a result of occasional mismatings if, in the process, he also fertilizes a number of females of his own species. The female, however, generally expends a great deal of energy on the production of a relatively small number of nutrient-packed eggs. If she errs in her choice of a mate, she may lose an entire year's —or even a lifetime's—production of young.

Isolating mechanisms are often grouped into classes such as premating and postmating. For practical purposes, premating (or prezygotic) mechanisms serve to prevent interbreeding or at least to prevent cross-fertilization between species. These mechanisms include differences in habitat preference of species that occur in the same area, differences in time at which mating occurs, differences in various aspects of the mating process, physical incompatibility of the reproductive organs, or failure of fertilization.

Postmating (or postzygotic) mechanisms interfere with the successful exchange of genes between species even though mating is possible. These include deficiency of the hybrid young, sterility of the hybrids, or hybrid breakdown.

Gene flow. Many species are spread over thousands of square miles and may look somewhat different in one part of their range than they do in another. But they still possess certain common characteristics by which their single species status can be recognized throughout the range. The maintenance of these features in common over vast areas may be the result of a cohesive force known as gene flow. Offspring produced from a mating are likely to move some distance away from their parents, and the genes they received will move with them. In the next generation the reproductive process will result in a similar "flow" of genes from one place to another. Certain species characteristics are essential regardless of the environment in which the species occurs; such traits spread easily because of gene flow and will be maintained throughout the population because of their survival value to the individual possessing them. It has not yet been possible, however, for biologists to state with certainty the relative roles played by gene flow and by selection in maintaining species characteristics over large areas.

The movement of genes through populations

Variability. *Genetic variation.* As already indicated, there is a great deal of variation between individuals in a species. These may be differences in colour patterns, in structural details, or in body chemistry. It is apparent that all human beings belong to the same species because interbreeding between them is demonstratably possible. Humans also share many characteristics in common, including a relatively large brain, longer legs than arms, the same number and shape of teeth, and the ability to communicate by symbols. In other characteristics, however, individual humans differ considerably; they may have different blood types, skin colour, hair colour, tolerance to cold or heat, and so forth. Sexual reproduction provides a great opportunity for shuffling genes at each generation, so that the offspring produced are never identical to the parents or, for that matter, to other individuals of the family or the species. Differences in blood type are but one familiar example of the fact that many important gene differences are unnoticeable. These "hidden" variations offer significant alternatives upon which selection can act.

In most animal species there are striking genetic differences between the sexes, and individuals are clearly male or female in some characteristics while sharing most of their other characteristics. Other differences, as in colour, shape, or life stage, may simply be polymorphic variations.

Finally, there is a genetic variability between populations of most species. This becomes obvious when it involves characteristics such as colour; for example, rodents that live on lava flows may have darker fur than individuals of the same species living on the pale soils and sand surrounding the lava flows. The genes that produce dark coat colour in the animals on the lava flows are favoured there because they render the rodents carrying these genes less subject to predators. On the paler ground, genes producing paler coloration are favoured for the same reason. Such variations present in the species will be acted upon by the selection to favour different series of phenotypes (obvious variants) in different parts of the range.

Environmental and age-related variations. It is often difficult to separate variations produced by genes from those produced by the environment, since most characteristics of organisms are a result of the interaction between a gene and its environment. Although genes control skin colour in man, for example, exposure to sunlight also has an effect. Thus, the distinction between which aspects of a trait are the result of differences in genes and which are the result of differences in environment is sometimes nebulous. Generally, when a certain characteristic is said to be environmentally produced, it is meant that most of the variability in that characteristic is produced by the environment.

Relative influence of nature and nurture

In many plants certain differences in colour and leaf shape are caused primarily by the conditions under which the individual plant grows. If it is shaded, its leaves may be larger to provide greater surface area for sunlight. Plants that grow in areas of poor soil and high wind velocity may be dwarfed and bent compared to individual plants growing in richer soil and a protected situation.

Among animals, differences in size may simply reflect the adequacy of diet, which varies depending upon the habitat. Similarly, the number of eggs laid by the female may reflect the adequacy of the diet available to her. Fish grown under crowded conditions but with adequate food may be stunted.

The aging process also produces obvious differences between individuals. In wild, hoofed mammals, defective joints are more common in old than they are in young individuals. In turtles, the edges of the shell often show a ragged appearance in old animals, and the distinct lines between the scales on the shell, very regular in young animals, may be quite erratic in older individuals. Some of the most obvious effects of aging are seen in man.

Many of these characteristics are probably primarily under direct genetic control. A common North American snake, the blue racer, shows a distinctly blotched pattern when young; the adult, however, shows no pattern at all. The timing of appearance of the difference is probably under genetic control. Similar features include the change in pattern of deer from the spotted fawn to the uniform colour of the adult, the development of horns or antlers in older animals, and the different stages in the life cycle of insects and frogs.

To the extent that any trait has some genetic variability, it can evolve, or change, in time as certain genes in given environments are either favoured or not favoured by natural selection. An athlete with large, firm muscles and a large chest capacity does not pass such characteristics on to his offspring, but if there is a genetic variability in the capacity of muscles and lungs to respond to stress, then such traits may be more easily developed through exercise in the offspring of outstanding athletes than in the offspring of less exceptional individuals.

Geographic variation in species. As already has been stated, many wide-ranging animal species show distinctly different characteristics in one part of the range than in another. Each variant type occurs over a fairly discrete geographic area, but at certain places in the range many intermediate types are found. Such areas, called zones of intergradation, represent subspecies boundaries. These intergradation zones indicate that there is no restriction to interbreeding between the variants. The presence of subspecies, called races among animals, confirms that there are stages in the speciation process that can readily be observed in most wide-ranging species. In many cases

such geographic variants were called different species before their true status was determined. Such geographic races are common in plants as well, but they are generally called varieties by botanists. The problems presented by these populations are the same in plant as in animal species.

Sibling, or cryptic, species. Species that look remarkably similar but nevertheless possess all of the characteristics of biological species are spoken of as sibling, or cryptic, species. For some reason, the genetic differences that exist between them are not manifested in their appearance. Detailed studies, however, have often revealed subtle morphological differences between sibling species.

Several related and similar-appearing species of frog have long been called leopard frog. Crosses between these species often result in young that are abnormal or have quite different mating calls. The species may differ in certain details of internal anatomy as well as in the number of external spots, though superficially they may look quite similar. Even more similar species are known among insects. Two common species of fruit fly (*Drosophila*) are exceedingly difficult to distinguish; their behaviour, however, is so different that there is little or no mismating between them under natural conditions.

THE EVOLUTION OF SPECIES

The process of species formation—speciation—evidently takes a very long time. Consequently, the origin of a new species cannot be observed. Changes that occur within a species, however, can be observed, as in the case of industrial melanism of the peppered moths (see EVOLUTION). The process of speciation has been deduced from data gathered by comparative examination of evidence and not by experiment. Speciation can conveniently be divided into species formation through geographical isolation, or allopatric speciation, and species formation in the absence of geographic isolation, or sympatric speciation.

Allopatric speciation. In the process of allopatric speciation, the range of a species becomes fragmented so that some populations are geographically isolated from others. This geographical isolation comes about in different ways. One obvious method, the invasion of islands by some individuals of a species, often happens when individuals are rafted on flood-borne debris; the castaways are then effectively isolated from others of their species. The cutting of large canyons over geological time spans disrupts the ranges of some species, as do changes in climate. The development of grasslands where once there were forests, encroachments by the sea, the reduction of a stream to a series of isolated pools, and the formation of caves may result in the isolation of some populations of a species. During the history of the earth, such disruptions of the ranges of species have been occurring continuously. Regardless of the manner in which geographic isolation actually occurs, the process of allopatric speciation begins with some degree of geographic isolation of populations within a single species.

The degree of geographic separation required to isolate a population depends upon the species concerned. A change sufficient to isolate small spring-dwelling salamanders would be of less significance to a flying insect, bird, or mammal. Plants with effective seed-dispersal or pollen-drift mechanisms will not be easily isolated by narrow disruptions of their ranges.

The significance of isolation

The evolutionary significance of geographic isolation is that populations, when separated, evolve independently. Natural populations cannot become highly specialized to their particular environment as long as there is a free flow of genes from adjacent populations. The gene flow is stopped or enormously reduced by geographic isolation, but this stoppage of gene flow does not explain why differences arise between the now-isolated populations.

Differences between isolated populations may arise in several ways. First, differences may arise by chance. Perhaps an island that is invaded and eventually populated by a few individuals of a species may not be reinvaded by others of that species for a long time. The small number of founders cannot possibly carry all the genes found in the large parental population from which they came. Consequently, there will be some chance differences between island and mainland populations. Second, differences may arise because of differences in the environments of isolated populations. As a result, selection may favour slightly different gene arrangements in one population than in another. The soils may be of different composition, predators may be different or may be present in different proportions, or competing species may be different; all these environmental variables affect the way in which selection acts. Thirdly, differences may arise by the process of mutation, which changes the arrangement or structure of genes. Finally, genetic differences may already have existed between the populations prior to their isolation.

Differences between isolated populations increase with the passage of time. One of the facts that struck Darwin was the enormous differences between breeds of domestic animals produced by deliberate selection on the part of man if such selection was consistently applied over several generations. Natural selection has the same effect but over longer periods of time in isolated populations.

Just as populations become isolated in time, they also can become reunited. Changes in the environment can permit formerly isolated populations to come into contact; their individuals then may freely interbreed, thereby tending to break down the differences accumulated during isolation. Some characteristics, however, may still be preserved as differences between overlapping populations, which are then recognized as subspecies. If the genetic differences accumulated during isolation are sufficiently great or if the differences affect important characteristics such as mating or embryological development, then interbreeding between the reunited populations may not occur at all or may occur and result in inferior offspring. If the two formerly isolated populations are effectively isolated by such genetic barriers and their differences are irreversible, speciation has occurred: two biological species have evolved from an original single species.

The geographic distribution of closely related species often shows that they are allopatric or that their ranges only partially overlap; both types of data suggest that isolation or former isolation was important in the formation of such species. Furthermore, it is often found that entirely unrelated animals, such as mammals and reptiles or insects and plants, contain pairs of species with similar ranges, suggesting that their isolation and subsequent speciation was the result of a common geographic barrier.

The selection of isolating mechanisms

It is important to recognize that isolating mechanisms are not favoured by natural selection because they prevent interbreeding, for if such mechanisms evolve in isolated populations there would be no chance for interbreeding, anyway. Even if two formerly isolated populations are reunited, selection may yet act to reduce the interbreeding between these populations. Earlier it was said that isolating mechanisms are generally divided into prezygotic mechanisms, which prevent fertilization, and postzygotic mechanisms, which operate against the offspring resulting from crosses between species. Clearly natural selection cannot usually favour postzygotic isolating mechanisms in zones of overlap between species. If, for example, the offspring between two species are sterile, there is no way for selection to strengthen the infertility. If the offspring die, selection cannot favour "increased" death. This is because the advantage will be to those offspring that are the least liable to die (for they will survive in greater numbers) and to those hybrids that are the least sterile (for they will produce more offspring). Thus, selection would tend to reduce rather than enhance the effectiveness of postzygotic isolating mechanisms.

It is, however, possible for selection to favour prezygotic mechanisms and to enhance the strength of such isolating mechanisms in zones of overlap. In time, a prezygotic mechanism could be produced in the zone of overlap that would be effective in reducing interbreeding and thus avoiding the waste of energy in the production of inferior hybrid offspring. In other words, reinforcement of reproductive isolation has occurred in the zone of overlap. It is

unlikely that such reinforcement could occur if there were no pre-existing postzygotic incompatibility between the two populations; otherwise there would be no disadvantage to interbreeding and thus no opportunity for reinforcing selection. The two populations would be subspecies or less. Such examples of reinforcement of isolating mechanisms are extremely rare, a fact that constitutes a weakness in speciation theory.

It may be questioned whether or not distinct geographic races, or subspecies, are stages in the speciation process. If differences accumulating in allopatry do not interfere with reproduction between individuals of different populations when the populations are reunited, then speciation has not occurred. This does not mean, however, that differences cannot be recognized between the populations. If, for example, the area in which the populations were rejoined was along a forest–grassland border, then the colour patterns best suited to animal species of the grassland might not be the ones best suited to species of the forest. Although interbreeding between the populations occurs, the movement, or flow, of genes affecting body colour may be interrupted along the border. This may occur with no incompatibility between the populations, so that speciation will not take place. Instead, there exist populations that can be recognized as subspecies with perhaps a narrow zone of intergradation between them in which there is considerable variability in body colour or pattern. This intergradation zone may persist over a long period of time.

Subspecies in general, therefore, are not stages in speciation because the process will go no further. On the other hand, some populations—particularly allopatric ones—that are recognized as subspecies have the potential for becoming different species. Such subspecies may very well be stages in the speciation process.

Sympatric speciation. Speciation that occurs without geographic isolation and while populations are in continuous contact with one another is called sympatric speciation. There has been considerable controversy over whether sympatric speciation is possible at all and, if so, how important it has been in the formation of species.

Advocates of the theory of sympatric speciation must explain in a plausible way how a single population can differentiate into different species without some barrier to gene exchange between them. A few examples will suggest how such differentiation may happen. Thus, a species of moth in Britain lays eggs on apple trees and hawthorn trees, and the larvae hatching from these eggs feed on these two trees. There is a tendency for moths reared on apple trees to mate with other moths reared on apple trees and to lay their eggs there. A similar predisposition affects the moths reared on hawthorn trees. Despite these habitat differences, the moths look alike and can interbreed; they represent a single biological species. At the same time, however, they suggest a mechanism by which sympatric speciation might occur in time. Genes that promote the success of moths on apple trees can increase in frequency because these moths tend to breed with other moths reared on apple trees; such is also the case with the hawthorns and their moths. In time, interbreeding between apple moths and hawthorn moths might result in offspring less capable of utilizing either food plant.

Disruptive selection

Then selection would be spoken of as disruptive—favouring either an apple specialist or a hawthorn specialist but not an intermediate type. This disruptive selection coupled with nonrandom mating—*i.e.*, a tendency to mate mostly with individuals reared on the same food plant—would result in reproductive isolation. Increased genetic divergence and speciation could then possibly occur just as if the moths were geographically isolated.

Sympatric speciation has been suggested for two North American field crickets with nearly, but not quite, congruent ranges. These two species differ primarily in their life cycles. In one, the adults breed in the spring; the other breeds in the fall. Because the adults are short-lived, there can be little or no interbreeding between them. It has been suggested that an ancestral species of cricket had both spring and fall breeders in a single population (such crickets are known among the relatives of those under consideration). A series of harsh winters could have resulted in the death of all but the resistant eggs and nymphs, effectively dividing the ancestral population into fall-breeding and spring-breeding strains between which there was essentially no genetic difference at the outset; however, there was opportunity for such differences to evolve because of reproductive isolation.

It is possible to construct theoretical models of ways in which sympatric speciation might occur, and it is also possible to promote genetic divergence sympatrically in laboratory populations with appropriate selective forces. Furthermore, there are species in nature whose characteristics and ranges suggest no history of geographic isolation. For these reasons sympatric speciation is not impossible or even improbable in some instances.

Sympatric speciation initiated by hybridization has evidently been of major importance in plant speciation. A similar process has occurred in animals but is far less common, presumably because hybridization between animal species is seldom successful. Most examples of hybridization seem to be associated with a disturbance of the habitat, the breaking down of some of the barriers to interbreeding, and provision of new sorts of habitats in which hybrid young may flourish.

Flooding by rivers, fires in forests or prairies, lava flows, and soil erosion have always altered plant environments.

How hybrids survive

The offspring produced by hybridization, if they manage to survive, are most likely to cross with one or the other parental species, a process known as backcrossing. This occurs because the parental types greatly outnumber the hybrids. Offspring from a backcross will contain predominantly parental types, which are already adapted to a particular environment; for this reason, backcrossing hybrids are more likely to be successful than the hybrid at producing viable offspring. Through backcrossing, at least some of the characteristics of one parent species will be transferred by way of the hybrid into the other species. This process, called introgression, is widespread in plants and occurs in a few animals. Introgression introduces new gene combinations into both parental species, providing new material upon which selection can act. Since the new combinations are enormous compared with the small number introduced by mutation, introgression can be of considerable importance in the evolution of some species.

One common problem in hybrids between species is sterility of the hybrids resulting from failure of the chromosomes to pair properly, a critically important process in the formation of sex cells. Plant species have often circumvented this problem by a process called polyploidy, in which the chromosome number in hybrids is altered. In some hybrid plants, a doubling of the chromosome number has occurred to produce a polyploid hybrid in which pairing between the identical duplicated chromosomes allows normal sex-cell formation and successful reproduction. The hybrid, however, must breed with other hybrids like itself; backcrosses to either parent species are unsuccessful. Thus, polyploid hybrids are reproductively isolated from either parental species, and speciation has occurred sympatrically. Comparative study of closely related species of plants indicates that many close relatives differ by multiples of some chromosome number, indicating that polyploidy has occurred in the evolutionary history of many of the species. It has been estimated that more than one-fourth of the flowering plants may be polyploid species.

In recent years polyploidy has been discovered in the vertebrate animals. At least one frog and several lizards are known to be polyploid species. The lizards are particularly interesting because, for the most part, they are triploids, meaning that they have one set of chromosomes derived from one parent species and two sets from the other. The triploid species are parthenogenetic; *i.e.*, only females are produced, and the eggs are able to undergo development without fertilization.

Stasipatric speciation. This relatively new concept in speciation theory was proposed to account for the distribution of various "species" of Australian grasshoppers.

Stasipatric speciation contains aspects of both allopatric and sympatric speciation. The grasshoppers concerned differ from one another in chromosome structure. In one type there evidently has been fusion of chromosomes. In other types there have been translocations of parts of chromosomes. These chromosome differences presumably arose by mutation, but it is difficult to understand how they spread, since the mutant containing one of each chromosome type would have been at a disadvantage. In quite small populations within the range of the grasshoppers, however, the new chromosome might become established in the population by chance, a process known as genetic drift. Once the new chromosome was present as the only type in one of these populations, it could spread through the parental species. Stasipatric speciation thus is visualized as a process by which differences arise in isolated populations within the range of a species and, once established, spread and displace the ancestral type in part of its range. It results in a number of closely related species with ranges that abut on one another or overlap only slightly.

The role of chance in spreading chromosomal differences

There seems little doubt that such a model of speciation is possible, but how different it is from allopatric or sympatric models is not clear. The grasshoppers just mentioned evidently have limited powers of movement and prefer specific habitats, so it is not difficult to accept the possibility that one population in a favourable patch of habitat might be geographically isolated from another even if it were in the centre of the range of the species. Thus, the differences between these species actually develop in allopatry. It is also possible that the chromosomal differences arise in isolated peripheral populations at the edge of the range of a species. Again, in this instance, allopatry serves to explain the situation.

THE RELATIONSHIP BETWEEN SPECIES AND OTHER TAXONOMIC CATEGORIES

The artificiality of higher taxa

The evolutionary relationship between species and subspecies or varieties has already been dealt with, but little consideration has been given to the relationships between species and genera or families, the higher taxonomic categories. All of the upper categories of taxonomy have been constructed in an attempt to show evolutionary relationships between species, between genera, between families, and so forth. All of these more inclusive levels of classification originated because of a speciation event at some time in the distant past.

Two closely related living species are distinct points in a continuous evolutionary process from a single ancestral species. With the passage of long periods of time—millions of years—these two species in turn may have given rise to several descendant species. Two groups clustering similar species constitute two genera. Since both genera had a remote common ancestor, they are placed in a common family.

There exists no complete record of any species through time to confirm the systematists' notions of relationships. Nevertheless, it reasonably can be assumed that species sharing a large number of characters in common are closely related and should be placed in a genus distinct from another group of related but significantly different species. Without direct knowledge of the history of the evolution of species, errors in the placement of species into higher categories are bound to occur. The category of species, however, appears to be the most stable and universal of all taxa. Beginning with the species that comprise the vast assemblage of organisms now on earth, systematists attempt to place them into genera, families, orders, classes, and phyla (or divisions) in a way that reflects the latest conception of their evolutionary history; for example, snakes and lizards are both considered orders in the class Reptilia, a grouping that implies that the two orders shared a common ancestor in the past; it in no way means that any living snake evolved recently from any living lizard. The speciation event that split these animals into different lines occurred millions of years ago; the animals are the end points from which the evolutionary history of the group is deduced. It is upon this history that a system of classification is imposed.

BIBLIOGRAPHY. ERNST MAYR, *Animal Species and Evolution* (1963), a masterful but technical reference book on species and speciation; JOHN MAYNARD SMITH, *The Theory of Evolution*, 2nd ed. (1966), perhaps the most readable, scholarly book on evolution and on speciation available in English; G. LEDYARD STEBBINS, *Processes of Organic Evolution*, 2nd ed. (1971), a concise and easily understood account of the process of evolution, particularly useful for an understanding of speciation in plants and the attendant phenomena of hybridization and polyploidy; M.J.D. WHITE, "Models of Speciation," *Science*, 159:1065–1070 (1968), a brief paper that sets forth the essential details of White's model of stasipatric speciation; BERNHARD RENSCH, *Neuere Probleme der Abstammungslehre*, 2nd ed. enl. (1954; Eng. trans., *Evolution Above the Species Level*, 1959), an analysis of patterns of evolution, primarily in animals, with consideration of the origins of higher categories.

(D.W.T.)

Spectroscopy, Principles of

Spectroscopy is the study of the absorption and emission of light and other radiation, as related to the wavelength of the radiation. Light from natural sources (such as the Sun, stars, fireflies, flames, and lightning) and from man-made sources (such as incandescent lamps, arcs, lasers, and fluorescent lamps) is generally composed of many colours. This composite nature of light, discovered in 1666 by the English physical scientist Sir Isaac Newton, is not usually evident to the casual observer, but it can be shown by passing light through a prism or other device that splits it into its constituent colours. A display or description of these colours emitted by a source of light is called the spectrum (plural: spectra) of the source (see Figure 1).

Spectroscopy is the science that deals with the sources, measurements, analyses, and uses of spectra. It has practical application in almost every technical field, especially for identifying constituents and processes in any source that emits light. With spectroscopic methods, one can analyze in the laboratory, for example, the composition of a small amount of material with an accuracy and speed that cannot be achieved by chemical processes. Samples can be compared to test their identity. A continuous flow of products can be monitored for variations or pollutants. Apparatus to measure temperatures in controlled thermonuclear fusion (the uniting of certain atoms so that part of their mass is converted to energy) is another special application of spectroscopy. Still another type of spectroscopic observation makes possible detailed studies of atomic and molecular structures. In astronomy, solar magnetic storms can be predicted and the chemical constitution of stars can be analyzed by spectroscopy.

The major divisions of this article are:

I. Introductory survey of optical spectroscopy
  General principles
    Basic features of light
    Basic properties of atoms
    Historical survey
    Applications
  Practical considerations
    General methods of spectroscopy
    Processes in spectral sources and types of spectra
    Laboratory sources of spectra
    Methods for dispersing spectra
    Detectors
II. Foundations of atomic spectra
  Basic atomic structure
  Hydrogen atom states
    Principal quantum number
    Orbital quantum number
    Magnetic quantum number
    Spin-orbit interaction
  The periodic table
  Transitions
  Perturbations of levels
III. Molecular spectra
  Experimental principles
  Theory of molecular spectra
    Origins of molecular spectra
    Quantization of molecular energy
  Applications
  Fields of molecular spectroscopy
    Microwave spectroscopy

Infrared spectroscopy
Raman spectroscopy
Visible and ultraviolet spectra
Electron spectroscopy
Molecular-beam spectroscopy
Double and multiple photon effects
IV. X-ray spectroscopy
Relation to atomic structure
Production methods
Analysis
X-ray detectors
Applications
V. Radio-frequency spectra
Origins
Methods
VI. Spectral analysis
Analyses of photographic data
Analyses with scanned spectra

I. Introductory survey of optical spectroscopy

GENERAL PRINCIPLES

Basic features of light. The term electromagnetic radiation describes the transfer of energy through space by electric and magnetic fields without the transfer of matter. Light is properly designated as that range of electromagnetic radiation visible to the human eye, although the term sometimes includes regions which are invisible.

Light, as well as other electromagnetic radiation, has a wave character, and specification by wavelength (distance between successive crests of a light wave) can be made far more precisely than specification by colour. In early spectrometers, each wavelength absorbed or emitted by a source was displayed as the image of a slit, with images side by side for many wavelengths; such images are called lines, and so each line represents a wavelength or frequency, and the series of lines is called a spectrum. Some modern spectrometers use electrical devices for detectors, and the spectrum is displayed as a series of peaks on a graph of line intensity versus wavelength; these peaks by convention are also called lines.

Meaning of spectral lines

Because the shortest wavelengths of electromagnetic radiation are on the order of atomic diameters, special units are employed to specify them. The angstrom unit, symbolized by Å, named after a Swedish physicist, Anders Jonas Ångström, is equal to 10^{-10} metre; and one nanometre, abbreviated nm, is equal to 10^{-9} metre. The micrometre, symbolized by the Greek letter mu, μ, is equal to 10^{-6} metre (equal to 10,000 Å or 1,000 nm) and is frequently used in the infrared region. Visible light has a wavelength range of about 400 nanometres (violet) to 700 nanometres (deep red).

It is often convenient to specify an electromagnetic wave by the frequency at which it oscillates, or vibrates, rather than by wavelength. Frequency (ν) and wavelength, symbolized by the Greek letter lambda, λ, are inversely proportional to one another, and their product is equal to a constant ($\nu\lambda = c$), in which ν is frequency or number of oscillations per second, usually expressed in hertz (Hz), one hertz being one cycle per second, or megahertz (1,000,000 Hz). The constant, c, is equal to the speed of light *in vacuo*, 2.997925×10^8 metres per second. Electromagnetic phenomena with their approximate wavelength and frequency ranges are shown in Table 1.

Basic properties of atoms. An isolated atom can exist as a stable structure only in certain discrete (distinct) states, called energy states or stationary states, and a measurement of the energy of the atom will result in one of the values characteristic of that atom. In principle, the so-called allowed states for any atom can be found from the laws of quantum mechanics. Under ordinary conditions, an atom is almost always in its lowest possible energy state (its ground state, in which it possesses the minimum possible energy). Energy can be added only in

By courtesy of John O. Stoner, Jr.

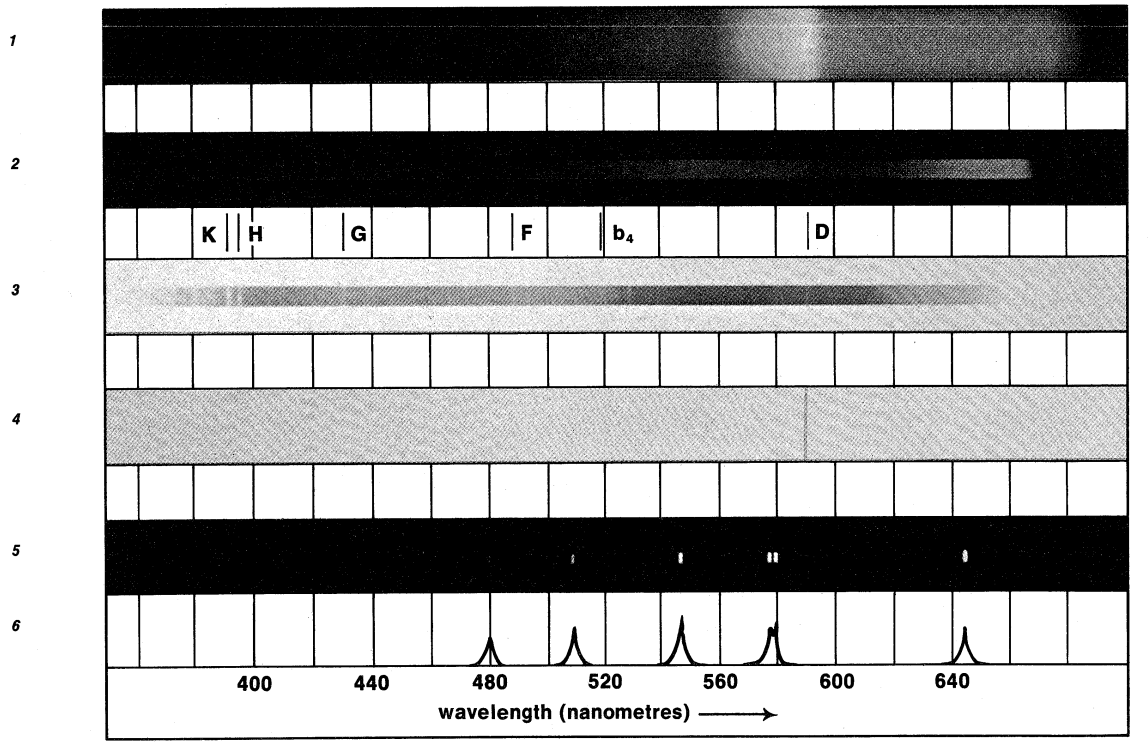


Figure 1: *Emission and absorption spectra of various light sources.*
(1) Continuous-emission spectrum from an ordinary incandescent lamp. (2,3) Spectrum of the Sun showing an absorption (dark-line) spectrum characteristic of the cooler outer layers of the Sun superimposed upon the continuous-emission spectrum of the hotter inner layers. The Fraunhofer lines (marked by letters) are characteristic of sodium (D), iron and magnesium (b_4), hydrogen (F), iron and calcium (G), and calcium (H,K). On the negative film (3), the film is darker where it has been more exposed; hence an absorption line appears white. (4) Emission-line spectrum of sodium showing a closely spaced pair of lines in the yellow. Other lines in this spectrum are much weaker than these and hence do not appear. (5,6) Emission-line spectrum of a discharge lamp containing mercury and cadmium as registered on colour positive film (5) and schematically as registered by a scanning spectrometer in which the height of the curve at any particular wavelength is the line intensity at that wavelength (6).

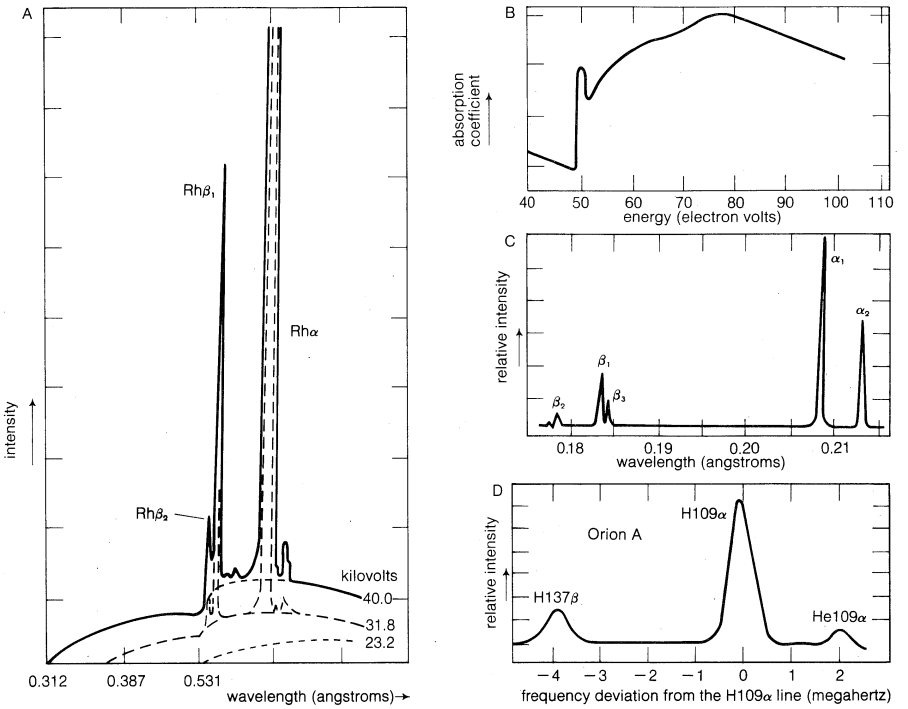


Figure 2: *Emission and absorption spectra of various X-ray and radio-frequency sources.* (A) X-ray emission from an X-ray tube with a rhodium anode operated at three different voltages, showing characteristic spectrum (α and β lines) of rhodium superimposed upon broad continua due to bremsstrahlung. (B) The absorption of magnesium versus incident X-ray photon energy, showing the abrupt increase in absorption as the energy becomes sufficient to remove an electron from the *L* shell. (C) Characteristic fluorescence line spectrum of tungsten, recorded with a bent crystal spectrometer. (D) Radio-frequency spectrum of centre of Orion Nebula; centre of largest peak is at frequency of 5008.9 MHz, which corresponds to a hydrogen transition.

From (A,B) S. Flugge (ed.), *Encyclopaedia of Physics*, (D) P.G. Mezger and P. Palmer, "Radio Recombination Lines: A New Observational Tool in Astrophysics," *Science*, vol. 160, pp. 29–42 (April 5, 1968). Copyright 1968 by the American Association for the Advancement of Science

those amounts that will just raise the energy to one of the other allowed values. The atom will not absorb an amount of energy that would make its final energy some intermediate value. An atom that has absorbed energy and is in one of its higher states is said to be in an excited state. The energy levels are identical for identical atoms; allowed energies of a particular atom of silver are the same as those for any other atom of the same isotope of silver (isotopes are all atoms of an element that have unequal mass but the same atomic number.

Table 1: Electromagnetic Phenomena

	approximate wavelength range (metres)	approximate frequency range (hertz)
Radio waves	10–1,000	3×10^5–3×10^7
Television waves	1–10	3×10^7–3×10^8
Microwaves, radar	1×10^{-3}–1	3×10^8–3×10^{11}
Infrared	8×10^{-7}–1×10^{-3}	3×10^{11}–4×10^{14}
Visible light	4×10^{-7}–7×10^{-7}	4×10^{14}–7×10^{14}
Ultraviolet	1×10^{-8}–4×10^{-7}	7×10^{14}–3×10^{16}
X-rays	5×10^{-12}–1×10^{-8}	3×10^{16}–6×10^{19}
Gamma rays (γ rays)	1×10^{-13}–5×10^{-12}	6×10^{19}–3×10^{21}
Cosmic rays	less than 1×10^{-13}	greater than 3×10^{21}

Other small systems, including molecules, ions (charged atoms or molecules), atomic nuclei, and elementary (subatomic) particles, have discrete allowed energies.

If an atom in its ground state is given some amount of energy so that it then exists in an excited state, it will give off that extra energy spontaneously as it moves back into lower states, ending in the ground state again. The energy is often emitted as electromagnetic radiation, hence it appears as light if the wavelength is appropriate. In the process of a single transition from some upper state to a particular lower state the energy given off is found from the combination principle: emitted energy (E) equals upper state energy minus lower state energy; this energy is usually carried by a single group of light waves (a photon) having a frequency (ν) in which photon energy (E) is equal to a constant times the frequency, or, $E = h\nu$, in which h is Planck's constant, a unit of action (work times time), 6.6256×10^{-34} joule second. This relationship determines frequencies (and wavelengths because $\lambda = c/\nu$) of light emitted by atoms if the energies of the states are known; conversely, the relationship allows finding the energy states of an atom from measurements on the spectrum of that atom. The allowed energies of an atom are discrete and therefore its spectrum consists of discrete wavelengths. Because of the appearance of a given wavelength when seen through a prism or grating spectroscope, these discrete emitted (or absorbed) wavelengths are called spectral lines (see Figures 1 and 5).

Historical survey. The basis for spectroscopic research and for analytical spectroscopy is the discovery, made by a German physicist, Gustave Kirchhoff, in 1859, that each pure substance has its own characteristic spectrum. Another German physicist, Joseph von Fraunhofer, repeating more carefully an earlier experiment by an English scientist, William Hyde Wollaston, had shown in 1814 that the spectrum of the Sun does not grade smoothly from one colour to the next but is crossed by many dark lines, indicating that light is missing at certain wavelengths because of absorption (see Figure 1 above; see also the Plate in the article LIGHT). These lines are still called Fraunhofer lines, and the spectrum containing them is called a darkline, or absorption, spectrum. The spectra of some metals and compounds when heated in flames and the spectra obtained from electric discharges in gases were studied by many scientists during the 18th and 19th centuries; these spectra were seen to be composed of numerous bright discrete lines indicating that only certain wavelengths were present in the emitted light. These are called brightline, or emission, spectra. Although the possibility of the existence of characteristic spectra, a spectrum unique for each chemical element, had been considered by numerous investigators, the early studies were hampered by the difficulty of obtaining rela-

The combination principle

tively pure substances and the fact that even slightly impure substances could yield many spectra simultaneously. By using carefully purified substances, Kirchhoff demonstrated characteristic spectra and initiated the technique of spectroscopic analysis of the chemical composition of matter. The technique was applied by Kirchhoff and his colleague Robert Bunsen in 1861 to the analysis of the solar spectrum and the identification of the chemical elements in the Sun.

Before the 20th century, there was no theory that could explain satisfactorily how spectra are formed or why different elements have different spectra. Some of the principal discoveries leading to present understanding of spectra were: the discovery of infrared radiation by an English astronomer, William Herschel, in 1800 and of ultraviolet radiation by a German physicist, Johann Wilhelm Ritter, in 1801; the recognition of a series in the spectrum of hydrogen in 1885 by a Swiss scientist, Johann Jakob Balmer; the discovery of the electron, the basic negatively charged constituent of atoms, by a British physicist, J.J. Thomson, in 1897; the beginnings of quantum theory by a physicist in Germany, Max Planck, in 1900; the formulation of the combination principle; the experimental proof that atoms are composed of tiny, massive positively charged nuclei surrounded by electrons, by a British physicist, Ernest Rutherford, in 1911; the first theoretical calculations of the discrete energy states of atoms and the wavelengths of light given off by atoms, by the Danish physicist Niels Bohr in 1913. Quantum mechanics was developed by the physicists Erwin Schrödinger and Werner Karl Heisenberg of Germany to describe the detailed properties of atoms. This theory, extended by the physicists Wolfgang Pauli of Austria and P.A.M. Dirac of England, and others, now provides methods for calculating energy states and spectral lines emitted by atoms, ions, and simple molecules.

Applications. In chemistry and physics, spectroscopy is used as a tool for studying the structures of atoms and molecules. The large number of wavelengths emitted by these systems makes it possible to check theories of their structure in detail.

Use of spectroscopy as an analytical tool

Spectroscopy provides a precise analytical method for finding the constituents in material having unknown chemical composition; in typical analyses, a concentration of a few parts per million of a trace element in a material can be detected when the material is vaporized and caused to emit light (as in an electric arc) to produce an emission spectrum, the bright lines that identify the element causing the spectrum. The absorption spectra of solutions containing inorganic or organic chemicals show dark lines and bands (close groups of lines) characteristic of the molecules, radicals (molecular fragments), and ions present in the solution. Emission and absorption spectra of the atmosphere are thus used for monitoring trace elements and compounds (*e.g.*, mercury or the compound dichlorodiphenyltrichloroethane [DDT]) in pollution control, the sensitivity of these techniques sometimes approaching a few parts per 1,000,000 in concentration.

In astronomy, the spectra of the Sun and other stars show lines characteristic of the atoms, ions, and radicals in them, plus lines characteristic of materials in the nearby empty space between the source and the Earth, and materials in the Earth's atmosphere. Spectra of the light reflected from planets other than the Earth show, in addition, lines caused by absorption by their atmospheres. Thus chemical analyses can be made of these distant objects.

Data on spectroscopic properties of atoms, ions, and molecules are used in the design of lasers (devices that operate on the principle of stimulating the emission of stored energy in atoms), which have wide scientific, industrial, and military application. Fluorescent lamps make practical use of the absorption of ultraviolet radiation with subsequent emission of visible light. High-purity spectral lines are used in metrology (the science of measurement), particularly for measurement of fundamental physical constants, such as the electron's charge and mass; the fundamental metric unit of length, the metre, is in fact defined as 1,650,763.73 wavelengths of a particular unperturbed line in the spectrum of krypton gas. Spectroscopic analyses are used in criminology, archaeology, medicine, agriculture, the biological sciences, and engineering.

PRACTICAL CONSIDERATIONS

General methods of spectroscopy. Production and analysis of a spectrum usually require four components: (1) a source of light (or other electromagnetic radiation), (2) a disperser to transmit different wavelengths differently, (3) a detector to sense the presence of light where it appears after dispersion, and (4) a recorder to make a spectrogram, a temporary or permanent copy of the spectrum.

An apparatus intended to accept light, separate it into its component wavelengths, and detect the spectrum is called a spectrometer; a spectrograph is a spectrometer that also records spectra. A spectroscope is a combination of disperser and associated optics for visual observation of spectra. A monochromator consists of a disperser and associated optical components arranged to accept light from an external source and transmit only a small range of wavelengths. A spectrophotometer is a spectrometer used to record quantitatively the amount of light emitted or absorbed by a sample of material according to the wavelength.

Processes in spectral sources and types of spectra. Two kinds of spectra can be distinguished: emission spectra, which show bright lines or bands on a dark background, and absorption spectra, which have a continuously bright background except for dark lines (Figure 1).

A spectral line is emitted when atoms (or other particles) are excited by some means to the upper energy state responsible for the transition and then decay to the lower energy state. The required initial excitation can be produced in a variety of ways. A collision between an electron with high energy (energetic electron) and an atom can give energy to the atom to raise it to an excited state. Electron impact is the principal basis for excitation in electric discharges (see below *Laboratory sources of spectra*). Excitation of atomic states can also be accomplished by irradiating atoms with electromagnetic radiation that they can absorb. Fluorescence spectra are produced when atoms become excited by absorbing radiation (often in the ultraviolet region) and then emit radiation as the atoms decay to lower states. Impacts on atoms by energetic atomic nuclei, ions, and molecules produce excited atomic states in special applications.

Electron excitation

An incandescent liquid or solid, such as a lamp filament, emits a broad, continuous range of wavelengths. In these sources, the atoms may be excited both by electron and by atom impact. Because the atoms are packed closely, their individual energy levels merge together into overlapping levels; the emitted lines then overlap and merge into a continuous—*i.e.*, nondiscrete—spectrum. Similar phenomena occur in the interiors of stars and in high-pressure arcs, in which broadening of spectral lines occurs because of the high collision rates.

So-called bound–free or free–free transitions occur, with resulting emission of radiation, when an electron makes a transition in its energy level to an unbound state from a bound or from another unbound state. These transitions also give rise to continuous spectra.

Before radiation reaches an observer, it may pass through a medium, usually consisting of a vapour of unexcited atoms (or ions or molecules), in which specific wavelengths are absorbed; these absorbed waves correspond to transitions by the vapour atoms from their ground states to higher states. As a result, the absorbed wavelengths are missing in the spectrum of radiation transmitted through the vapour. This phenomenon is the origin of the Fraunhofer lines in the solar spectrum; light emitted by the hotter, deeper layers within the Sun is partially absorbed by atoms in the cooler, outer layers before it reaches Earth. Flames are often used as sources of spectra. Temperature and composition vary throughout a flame; hence flames can be used as absorption sources or as emission sources.

An emitting atom moving toward an observer (or in-

strument) compresses the light waves emitted in that direction. The light then has shorter wavelengths than it would have if the emitter were at rest; *i.e.*, its spectrum is shifted toward the blue. Similarly, light emitted by atoms moving away from the observer has longer wavelengths, hence appears redder. This phenomenon is called the Doppler effect. As a result, light that would have a well-defined wavelength if emitted by atoms at rest has a spread or broadening caused by their ordinary random thermal motion toward and away from the observer. This Doppler broadening restricts the useful resolving power (ability to distinguish between two adjacent lines) of spectroscopic instruments. The average speeds of atoms in a source increase with temperature; so the widths of spectral lines indicate the temperature of the source. The shifts of spectra with relative velocities of source and observer are sometimes used to measure velocities of sources, especially in astronomical spectra.

Laboratory sources of spectra. In atomic spectroscopy, a few sources of spectral emission lines serve for most purposes. These sources are: the hollow cathode; the Geissler tube invented by Heinrich Geissler, a German physicist; carbon arc; sparks; electrodeless lamps; and flames (see Figure 3).

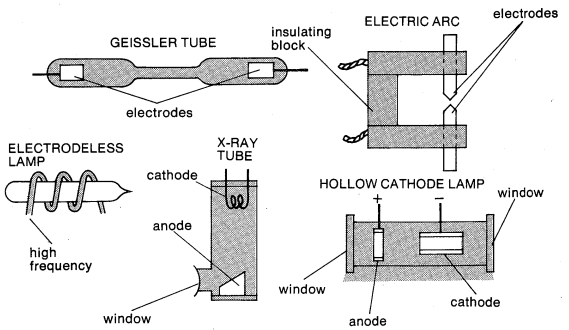


Figure 3: *Sources of laboratory spectra.*
All of these sources, except for the X-ray tube, generate visible and ultraviolet light.

The hollow cathode discharge lamp

The hollow cathode discharge lamp (Schüler tube after Hermann Schüler, a German physicist) consists of two electrodes mounted inside a glass or ceramic envelope. The cathode (negative electrode) is usually formed as a cup or hollow cylinder constructed of, or coated with, the material the spectrum of which is to be investigated. The gas within the envelope is usually one of the inert gases —helium, neon, or argon—at a pressure of roughly 1 torr (the torr is a unit commonly used in specifying gas pressures; sea-level atmospheric pressure is approximately 760 torr). A direct electric current, typically one to 100 milliamperes, at several hundred volts, is passed through the tube; the spectrum appears most brightly within the cathode.

The Geissler discharge tube, the archetype of the neon signs used extensively in advertising, consists of a glass tube having an electrode at each end and filled with a gas at pressure typically 0.1–10 torr. Often a portion of the glass tube is constricted to provide greater current density (hence greater brightness) in that region. Transformers operating typically at voltages from 1,500–10,000 volts and currents from 10–50 milliamperes, alternating current (ac), power these tubes. A modification of the Geissler tube is used to provide the metric standard of length.

Electrodeless lamps consist of glass tubes, usually about one centimetre in diameter and a few centimetres long, filled with a gas at 0.1–10 torr pressure. Spectra of materials other than gases can be obtained by adding to the gas a few milligrams of a volatile compound containing the element of interest; *e.g.*, uranium chloride for the spectrum of uranium. The power source is a high-frequency oscillator, typically 25–2,500 megahertz. (Current within the lamp may be induced by wrapping wires leading from the oscillator around the lamp or by placing the lamp inside an appropriate resonant cavity connected to the oscillator.)

For rapid spectroscopic processing and analysis of solid materials, arcs are convenient sources. Graphite and copper are typical electrode materials. A small amount of the material of interest may be powdered, mixed with powdered electrode material, and pressed into a shape convenient for use as one electrode of an arc. Other methods can be used to incorporate the material within an electrode. Arcs can be run with a wide range of voltages (perhaps 10–100 volts) and currents in the range of 1 to 100 amperes, in an atmosphere of air or some other gas. Usually the current is direct current (dc), with the powdered material in the cathode. Effective temperatures in arcs often range from 3,500° to 5,000° C (6,300° to 9,000° F). At such temperatures, little ionization of the atoms occurs, and mainly the spectra called the first, or neutral, spectra of elements are observed.

Spark sources are often used when higher excitation is required than is available in an arc. The simplest spark source consists of the discharge of a strong current at a potential of around 5,000 volts across a gap between two electrodes. Effective temperatures for these sources can range into the tens of thousands of degrees Celsius; the materials within the spark are often ionized, sometimes highly so, and the dominant spectra are the second (II, singly ionized, or one electron missing from each atom), third (III, doubly ionized, or two electrons missing from each atom), etc., depending upon the power released in the spark, the materials and spacings of the electrodes, and the surrounding atmosphere.

The first man-made spectroscopic sources consisted of flames to which small amounts of substances were added, to be vaporized and excited by collisions with the constituents of the flame. Flame sources are used extensively in analytical chemistry; in these, the materials are formed into an aerosol so that they can be injected into the flame conveniently at a controlled rate.

Transparent chambers or containers with windows at both ends serve as absorption cells for the production of absorption spectra. Light having a continuous distribution of wavelength is passed through the cell. When a gas or vapour is introduced into the cell, the change in the transmitted spectrum can be interpreted as the absorption spectrum of the gas. Often, absorption cells are enclosed in ovens because many materials of spectroscopic interest vaporize significantly only at high temperatures.

Methods for dispersing spectra. *Refraction.* Historically, glass prisms were first used to break up or disperse light into its component colours. The path of a light ray bends (refracts) when it passes from one transparent medium to another; *e.g.*, from air to glass. Different colours (wavelengths) of light are bent through different angles; hence a ray leaves a prism in a direction depending upon its colour (see Figure 4). The amount of

The refracting prism

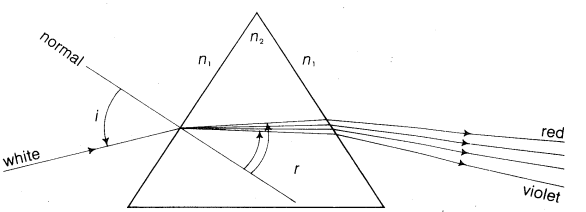


Figure 4: Refraction of light by a prism having index n_2 immersed in a medium having refractive index n_1. The angles i and r that the rays make with the normal are the angles of incidence and refraction. Because n_2 depends upon wavelength, the incident white ray separates into its constituent colours upon refraction, with deviation of the red ray the least and the violet ray the most.

bending of a ray at each interface can be found from Snell's law (Willebrord van Roijen Snell of Holland, *c.* 1600), which states that if n_1 and n_2 are the refractive indices of the medium outside the prism and of the prism, respectively, and the angles i and r are the angles that the ray of a given wavelength makes with a line at right angles to the prism face, as shown in Figure 4, then the equation $n_1 \sin i = n_2 \sin r$ is obtained for all rays. The refractive index of a medium, indicated by the sym-

bol n, is defined as equal to the speed of light *in vacuo* divided by the speed of light in the medium. Typical values for n range from 1.0003 for air at 0° C (32° F) and atmospheric pressure, to 2.3 for thin layers of zinc sulfide.

For typical optical glasses, the refractive index is in the neighbourhood of 1.5–1.9 and varies by 2 to 4 percent across the visible spectrum, with the refractive index greater for violet light than for red light. Because the variation of the refractive index is small, the spectrum is spread over a relatively small range of angles; thus prism instruments are generally used in applications for which low spectral resolution (ability to separate wavelengths) is sufficient.

In order for a prism to produce a large number of exit light beams having different, well-defined wavelengths, it is necessary that the incident light beam composed of these same wavelengths falls upon the prism as a set of parallel rays. This condition is conveniently accomplished by a collimator, a combination of lens and aperture (in this case a slit). Light rays from the source pass through the slit at the focus of the collimator lens, which then renders the rays nearly parallel (see Figure 5). After

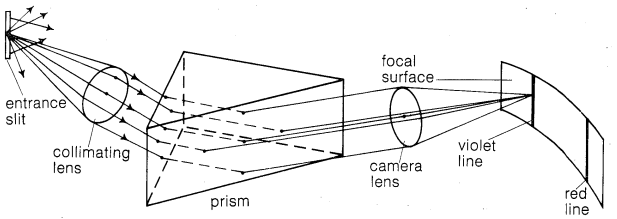


Figure 5: *Principles of a prism spectrograph.*
Light from a point on entrance slit (illuminated by a light source) is formed into parallel rays by the collimating lens. The light is refracted by the prism, then focussed by the camera lens. Violet light from the point is brought to focus; the slit is imaged as a line at the focal surface. Red light from slit is imaged as a different line. A spectrum thus appears at the focal surface.

the light is refracted by the prism, it is focussed by a lens, usually called the camera lens; if a viewing screen is placed at the focus, an image of the slit is seen for each wavelength of light present. Because the image of a slit is a line, each wavelength forms its own line. The lines can be observed by the eye or by placing photographic or photoelectric detectors at the focal position.

There are other devices besides prism instruments that can be used to break up light into its constituents. A collimated beam of light may be considered as a collection of parallel wave fronts advancing in the direction of travel of the light. Just as a water wave breaks up into ripples moving in all directions when it strikes a barrier, a light wave breaks up into waves travelling in all directions when it meets a barrier. This is the phenomenon of diffraction. Where the sizes of optical components approach the wavelength of light, diffraction phenomena become predominant.

Diffraction. Diffraction gratings are composed of closely spaced transmitting slits on a flat surface (transmission gratings) or alternate reflecting and nonreflecting grooves on a flat or curved surface (reflection gratings). A typical grating might be ruled on an aluminum coating on a piece of glass having dimensions from 2 × 2 centimetres (0.8 × 0.8 inches) up to 15 × 30 centimetres (6 × 12 inches) and consist of equally spaced parallel lines, as few as 100 lines per millimetre of width or as many as 6,000 lines per millimetre. The close spacing of lines requires that the wave picture of light be used to describe the interaction of light with a grating.

Secondary waves

If collimated light falls upon a transmission grating, the wave fronts, or crests, successively pass through and spread out as secondary waves from the transparent parts of the grating. Most of these secondary waves, when they meet along a common path, interfere with each other destructively, so that light does not leave the grating at all angles. At some exit angles, however, secondary waves from adjacent slits of the grating are delayed by exactly one wavelength, and these waves reinforce each other when they meet; that is, the crests of one fall on top of the other. In this case, constructive interference takes place and light can leave the grating. Constructive interference occurs also for delays of 0, 2, 3, 4 . . . wavelengths. The grating equation describes this phenomenon for both the transmission and reflection grating and gives the relationship between the incident angle, i; exit, or reflection angle, r; spacing between grating slits, d; wavelength, λ; and the integer, m (usually called the order of interference): the product of order and wavelength is equal to the product of slit spacing and the difference between the sines of the angle of incidence and exit; *i.e.*, $m\lambda = d(\sin i - \sin r)$. If light having several constituent wavelengths falls upon a grating at a fixed angle, i, different wavelengths are diffracted in slightly different directions and can be observed and recorded separately. Each wavelength is also diffracted into several orders (or groupings); gratings are usually blazed (engraved) so that a particular order will be the brightest. A lens or concave mirror can then be used to produce images of the spectral lines.

A simplification is obtained when a concave grating is used as a disperser. This device is formed by ruling on a concave mirror parallel lines equally spaced along a geometrical chord of the surface. Because a concave mirror has focussing properties, it can replace both collimator and camera lens (see Figure 6). Henry Augustus Rowland, a physicist in the United States, showed (1882) that if entrance slit, grating, and photographic plate lie on a circle the diameter of which is equal to the radius of curvature of the grating, all spectral lines can be made to focus. Many large, high-resolution grating spectrographs have been built on the basis of the Rowland circle, and most tabulated data on wavelengths have been obtained with such instruments.

The Rowland mount

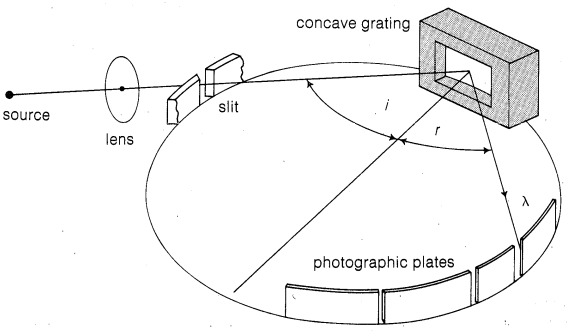


Figure 6: *Essentials of a concave-grating spectrograph based on the Rowland circle.*
Light from source is imaged onto the entrance slit by the lens. Light having wavelength λ will be diffracted by the concave grating shown in its holder, at angles r fixed by the grating equation and imaged on the photographic plates arranged in an arc around the circle to accept one of the various orders.

In any grating spectrometer, if a slit aperture is moved along the surface where the spectral lines are focussed, the lines are transmitted successively through the aperture. A detector placed behind the aperture then responds to the spectrum, line by line; the output of such a detector can then be amplified and recorded. These scanning techniques have some advantages over photographic recording: because the detection efficiencies of some electrical detectors (see below *Detectors*) are greater than efficiencies of photographic detectors, the method provides greater sensitivity and greater speed. Many photoelectric detectors respond nearly linearly—*i.e.*, with response proportional to the amount of light received—so the recorded brightness is simply related to the characteristics of the source.

In some arrangements, it is more convenient to tilt the grating (or prism) about its axis that is parallel to the slit than to move the slit. Typically, the grating may be rotated so that the angles of incidence and diffraction change, hence changing the wavelength of light reaching the fixed exit slit.

Interference. A third class of devices for dispersing spectra consists of interferometers, which divide a wave front by semitransparent surfaces, allow the beams thus produced to travel different paths, and then recombine

the beams. Variants of this method are used in testing of optical components. In spectroscopy, the principal interferometers are those developed by a United States physicist, A. A. Michelson (1881), in an attempt to find the luminiferous ether—a hypothetical medium then thought to pervade all space—and by two French physicists, Charles Fabry and Alfred Pérot (1897), specifically for high-resolution spectroscopy (see Figure 7).

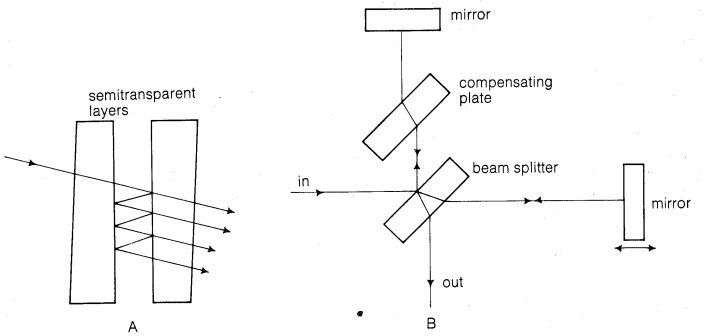


Figure 7: *Interferometers often used in spectroscopy.* (A) Fabry–Pérot interferometer; the plates are wedged to aid in the removal of spurious reflections. (B) Michelson interferometer, showing the path of a light ray through a beam splitter, which transmits and reflects approximately equal amounts of light, and a compensating plate to make the paths in glass equal for both arms of the instrument.

In the Michelson interferometer, an incident beam of light strikes a tilted semitransparent mirror and divides, weaker waves being reflected and transmitted. These continue to their respective mirrors, are reflected, and return to the semitransparent mirror. The two waves finally moving toward the observer have travelled different path lengths, hence interfere and are transmitted preferentially only if their path difference is 0, 1, 2, 3, . . . wavelengths. This device then acts as a filter that transmits preferentially certain wavelengths and reflects others back to the light source, resulting in a visible interference pattern. The Fabry–Pérot interferometer (Figure 7A) consists of two parallel glass plates coated with highly reflecting, slightly transparent films on their inner faces. When the interferometer is placed in a beam of light from a spectrometer, multiple reflections take place between the two faces, producing concentric interference rings. Even if there are two spectral lines so close together that they cannot be resolved by an ordinary spectroscope, the Fabry–Pérot interferometer will produce rings of slightly different radii, depending on the wavelengths of the two lines, which can be distinguished from one another.

Instruments for obtaining high resolving power often use interferometers in series with grating instruments. The grating preselects the line or lines of interest, which are further analyzed by the interferometer.

Multiplex spectrometers

Beginning about 1960, instruments known as multiplex spectrometers found considerable use, especially in recording astronomical infrared spectra. The prototype of this kind of instrument is the Michelson interferometer (Figure 7B), with one mirror mounted upon a carriage so that length of the light path in that branch of the light path can be varied. A spectrum is obtained by recording photoelectrically the light intensity of the interference pattern as the carriage is moved. Little preselection of wavelength is done, and so the resulting signals, containing information simultaneously about many wavelengths, are complicated. A mathematical transformation (called a Fourier transform) must be performed on the data to yield the spectrum (see OPTICS, PRINCIPLES OF). The principal advantage of this method is that the entire spectrum is recorded simultaneously with one detector. With the use of these Fourier spectrometers and other types of multiplex spectrometers, the spectra obtained are less affected by electronic noise originating in the detectors than are the spectra obtained by conventional scanning spectrometers.

Detectors. The principal detection methods used in spectroscopy, to be described below, are: photographic (plates and films), photoemissive (photodiodes, photomultipliers, image tubes), radiometric (bolometer, thermocouple, thermopile, Golay cell), and semiconductor (photoconductive and photovoltaic cells).

In most investigations before about 1940, and in many current applications, a strip of photographic film or a photographic plate records the pattern of spectral lines. This method has several advantages over other methods of recording: the source brightness need not be constant; information over the entire spectrum is recorded simultaneously; a permanent record results, detailed studies of which can be made later.

Photoemissive detectors have replaced photographic plates and films in many applications. When a photon (quantum of light) having sufficient energy strikes a metal surface, it can cause the ejection of an electron from the surface. A stream of photons releases a series of electrons, hence produces an electric current. A photoemissive diode (photodiode) consists of a surface (the cathode) appropriately treated to permit the ejection of electrons and a separate electrode (the anode) on which electrons are collected, both often sealed within an evacuated glass envelope. A photomultiplier (or multiplier phototube) has a cathode, a series of electrodes (dynodes), and an anode sealed within a common evacuated envelope. Some photoemissive detectors use continuous-strip or continuous-tube dynode structures. Appropriate voltages applied to the cathodes, dynodes, and anode cause electrons that are ejected from the cathode to make collisions with the dynodes in succession, knocking out more electrons upon each collision. The current is thus multiplied by the tube and is then more easily amplified and recorded. A single photon ejecting a single electron from the cathode can be detected as a pulse of 100,000 to 10,000,000 electrons at the anode; for some applications, these single pulses are counted directly. Photomultipliers are useful primarily in the wavelength range 1,050–12,000 angstroms, from far ultraviolet to near infrared; the lower wavelength limit is set by light-transmission properties of optical materials, whereas the upper limit is set by the requirement that the photon must have enough energy to eject an electron from the cathode.

Image intensifier tubes contain a sensitive photocathode and provision for accelerating the ejected photoelectrons so that they strike a phosphor screen, causing many photons to be emitted; alternatively, the electron may pass through a thin film, ejecting more electrons, which are accelerated in turn and detected, or may strike a photographic plate, causing its exposure. Such devices are often used in night-vision devices. They have found some application in spectroscopy because they maintain to some extent the sensitivity of photoelectric detection while preserving the information about the location at which the initial photon arrived.

A bolometer is a device the electrical resistance of which changes when it becomes warmer as a result of its absorbing light. A thermocouple junction is a junction between dissimilar metals. If an electrical circuit is made using two dissimilar metals, and if the junctions are at different temperatures, current flows in the circuit. A junction that absorbs light can thus be used as a detector, because its temperature generally increases when it absorbs light. A thermopile is a set of thermocouples mounted in series in order to obtain sensitivity greater than that obtained with a single junction. Golay cells detect light by the expansion of a gas when it is warmed by incident radiation. The four detectors described here have the common characteristic of responding to incoming energy of any wavelength; hence they can be used for detection in the far infrared where photographic and photoemissive detectors are insensitive.

Photoconductive cells contain thin semiconductor layers in which electrical resistance varies with illumination. Geiger–Müller tubes, proportional counters, and ionization chambers that depend on the release of electrons by incoming photons are used for detecting radiation from far ultraviolet to gamma rays (see below *X-ray detectors*).

When detection of spectra is done photographically, the

Recording of spectra

spectrogram serves also to record the results. In order to produce a nonphotographic permanent record, as with a scanned spectrum, the output signal current or voltage from the detector may be amplified and fed into an auxiliary output device. The simplest such device is a galvanometer that deflects an amount proportional to the input signal; a chart recorder can be used to draw an ink record of deflection versus wavelength on a moving strip of paper, or the amplified signal may be digitized and printed or recorded on magnetic tape, paper tape, or punched cards for later handling by computer.

II. Foundations of atomic spectra

BASIC ATOMIC STRUCTURE

Because the emission of spectral lines depends on the electron structure of atoms, it is apropos here to show how the periodic table of elements is organized. This will require definitions of the following quantum numbers: principal, orbital, spin, and magnetic. An atom, the smallest stable particle of neutral matter, has a diameter of approximately 2×10^{-10} metre. It consists of a core, or nucleus, which contains most of the mass of the atom, and a collection of electrons surrounding the nucleus. The nucleus contains a certain number (Z) of protons and a generally different number (N) of neutrons; the diameter of a nucleus depends upon the number of protons and neutrons and is typically in the range 1.5 to 10 $\times 10^{-15}$ metre (.000015 to .0001 angstrom); *i.e.*, much less than an atomic diameter. Each proton has one unit of positive electric charge ($+e$). Because each electron carries the basic unit of negative electric charge, and because opposite charges attract one another, in a neutral atom the total charge of the nucleus, $+Ze$, is balanced by Z electrons surrounding the nucleus. An atomic ion is an atom having either more or fewer electrons than protons (*i.e.*, more or fewer than Z electrons); an ion thus has a net negative or positive electric charge.

The chemical and spectroscopic properties of atoms and ions are primarily determined by their electronic structures; *i.e.*, by the numbers and arrangements of electrons surrounding their nuclei. Typical energies of electrons within an atom range from a few electron volts (eV) to a few thousand electron volts. (One electron volt, equal to 1.60×10^{-19} joule, is the kinetic energy gained by an electron when it is accelerated by a one-volt potential difference.) Chemical reactions and other processes occurring in spectroscopic sources usually involve energy exchanges of this order of magnitude. Processes that occur within nuclei to produce electromagnetic radiation and other forms of radioactivity typically involve energies ranging from thousands to millions of electron volts; hence nuclei are nearly unaffected by the usual processes occurring in chemical reactions, light absorption, and light sources.

The forces holding an atom together are primarily electrostatic—*i.e.*, the mutual attractive electric force between the positive charges in the nucleus and the negative charge of each electron. Because like charges repel one another, there is a significant amount of electrical repulsion of each electron by the others. Also, two principles not usually apparent in the macroscopic world of everyday experience become important in atomic phenomena. The Pauli exclusion principle (1925) states that no two electrons can exist in the same energy state. The Heisenberg uncertainty principle (1925) predicts that it is impossible to measure accurately both the position and momentum of an electron simultaneously. A more useful approximate picture of an atom is that of a massive nucleus concentrated at a point surrounded by an electron mist or cloud, the density of which is proportional to the probability that an electron can be found there.

To a rough approximation, the electrons within an atom or ion move independently of one another. It is often possible first to calculate the properties of an atom on the basis of independent motions of the electrons rather than to make corrections for the correlations in electron motion caused by their forces on each other. Calculations begin with finding the possible states of atoms (and ions) that have just one electron outside a closed shell (see below *The periodic table*)—*i.e.*, hydrogen-like atoms and ions. The electron states in a complex atom are built up of electrons in these "one electron" states, and then corrections are made for the interelectronic forces.

HYDROGEN ATOM STATES

Principal quantum number. A hydrogen atom is composed of a nucleus (a proton) and a single external electron. The electron has wave properties in addition to particle properties. (Experiments show that atoms, molecules, and other microscopic particles also have wave properties.) Its wavelength is specified by the de Broglie relation (1924): particle wavelength, symbolized by the Greek letter lambda, λ, is equal to Planck's constant, h, divided by momentum (mass times velocity), or $\lambda = h/p$, in which p is the momentum of the particle. This wave then travels around the nucleus of the atom. In order that the wave not cancel itself, it must travel at the appropriate distance from the nucleus so that the circumference of the orbit is equal to n times the electron wavelength, in which n is a positive integer equal to 1, 2, 3, ... The electron wave will then be stable. The motion of the electron is also limited by the requirement that the force on the electron is equal to the electron's mass times its acceleration (Newton's second law). The result is that the electron can exist stably only at certain average distances from the nucleus, and these distances are specified by giving n, the integer mentioned above. The electron's average distance from the nucleus is equal to the square of this integer times a constant, or n^2a_0, in which a_0 (equal to .525 angstrom) is known as the first Bohr radius (because Niels Bohr first calculated it on the bases outlined here). The number n is usually called the principal quantum number for the orbit. The energy of an electron in a hydrogen-like atom or ion is equal to minus the atomic number (Z) squared times Planck's constant (h) times the speed of light (c) times a constant divided by the principal quantum number squared—*i.e.*, $-Z^2hcR/n^2$. The constant R is called the Rydberg constant, and its product with Planck's constant and the speed of light (hcR) is approximately 13.6 electron volts. The energy is negative, indicating that the electron is bound to the nucleus. The algebraic value of the energy increases as n increases; hence the principal quantum number n equals one for the ground state of hydrogen. As the integer n takes on larger values toward infinity, the energy approaches zero. As this happens, the electron's average distance from the nucleus increases without limit and the atom becomes more nearly ionized (an ion approaches further ionization). The energy that is added to an atom or ion in its ground state just to cause the removal of an electron is called the ionization potential.

De Broglie waves

The first Bohr radius and the Rydberg constant

The combination principle suggests that the energies, frequencies (ν), and wavelengths (λ) of photons emitted by hydrogen atoms ($Z = 1$) should be expressed by $E = hf = hcR(1/n_1^2 - 1/n_2^2)$ and $1/\lambda = R(1/n_1^2 - 1/n_2^2)$, in which n_1 and n_2 take on the values of the principal quantum number for the final and initial states, respectively. The latter formula (with $n_1 = 2$) was found empirically by Balmer in 1885 to fit the wavelengths of the visible spectrum in hydrogen if $n_2 = 3, 4, 5, \ldots$. This was the first discovery of a spectral series. It was found in 1890 by the Swedish physicist Johannes Robert Rydberg that numerous data for the other spectra of atoms could be fitted by series formulas that are a slight modification of Balmer's formula, $1/\lambda = R[1/(n_1 - a)^2 - 1/(n_2 - b)^2]$, in which n_1 is held fixed for each series, n_2 takes on integer values greater than n_1, and a and b are nearly constant numbers for a given series and are called quantum defects. In each series, as n_2 becomes much larger than n_1, the wavelength approaches the series limit given by the wave number, $1/\lambda$, equal to $R/(n_1 - a)^2$.

Orbital quantum number. The second quantum number to be considered is the orbital quantum number. Although classically an electron's orbit may be elliptical with any ratio of major to minor axes, solution of the quantum mechanical equations shows that only certain distributions of the electron cloud exist in space and these depend on the angular momentum of the electron. Angu-

lar momentum is a concept applied to a body that is moving in an orbit or rotating on its own axis; it depends on the angular velocity and distribution of mass about an axis; it has the property of direction, being at right angles to the plane of motion. The orbital angular momentum of the atom is equal to the square root of a positive integer times the next successive integer, multiplied by the quantity Planck's constant divided by two pi—*i.e.*, $\sqrt{l(l+1)}(h/2\pi)$, in which l is a positive integer and h is Planck's constant. The number l, called the orbital quantum number, is restricted to be less than the principal quantum number; so if n is equal to one, l can only be zero, and hence the orbital angular momentum is equal to zero. If $n = 2$, l can be 0 or 1, yielding orbital angular momentum $= 0$ or $\sqrt{2}(h/2\pi)$. If $n = 3$, l can be 0, 1, or 2, etc.

Magnetic quantum number. The third quantum number is the magnetic quantum number. The tilt of an electron's distribution will be restricted to certain values when it is surrounded by a magnetic field. This phenomenon, the restriction of allowed directions for the atomic angular momentum, is known as space quantization and was first demonstrated in 1922 by two German physicists, Otto Stern and Walther Gerlach. The component (value) of the angular momentum of the atom along any given direction is equal to integral units of a quantity, Planck's constant divided by 2π—that is, $m_l(h/2\pi)$, in which m_l is another integer, being positive or negative so long as its magnitude does not exceed l. For instance, if $l = 1$, m_l can have the values $+1$, 0, or -1. The number m_l is usually called the magnetic quantum number.

Space quantization

Spin quantum number. The fourth quantum number is concerned with the spin of electrons. Experiments show that an electron has a further characteristic, its spin, which affects its motion and hence its possible energies. The electron behaves as if it were a spinning top, with its own intrinsic angular momentum having magnitude that can be obtained from the expression for orbital angular momentum by substituting ½ for l—*i.e.*, $\sqrt{(\frac{1}{2})(\frac{1}{2}+1)}(h/2\pi) = \sqrt{\frac{3}{4}}(h/2\pi)$—one says that the electron has spin $s = \frac{1}{2}$. Whether this angular momentum is directed toward or opposite a particular direction is indicated by giving the component of s along that direction. The allowed values are equal to $m_s(h/2\pi)$, in which m_s is either $+\frac{1}{2}$ or $-\frac{1}{2}$.

Spin–orbit interaction. These four quantum numbers, n, l, m_l, m_s, specify the state of a single electron completely; each number implies a particular wave function (equation) describing that aspect of the electron's motion. The description of an electron's motion can be done entirely mathematically using these numbers and the functions they represent.

The energies of the electron in a hydrogen-like atom or ion are fixed primarily by the mutual electrostatic attraction of the electron and the nucleus; there are, however, significant magnetic effects affecting the energies. An electron has an intrinsic magnetic dipole moment; that is to say, it behaves in some ways as a small bar magnet aligned along its spin axis. Also, because of its orbital motion within the atom, the electron creates a magnetic field in its vicinity. The interaction of the electron's magnetic moment with the magnetic field created by its motion (the spin–orbit interaction) modifies its energy to depend upon n, the principal quantum number, and another quantum number, j, representing a combination of its orbital-angular-momentum quantum number, l, and its spin-angular-momentum quantum number, s. In keeping with the requirement that energy and angular momenta can have only discrete values within an atom, the quantum number, j, can take on positive values only between $l + s$ and $l - s$ in integer steps; because s is ½ for the single electron in a hydrogen-like atom, j is ½ for $l = 0$ states, $j = \frac{1}{2}$ or $\frac{3}{2}$ for $l = 1$ states, $j = \frac{3}{2}$ or $\frac{5}{2}$ for $l = 2$ states, etc. The magnitude of the total angular momentum of the atom, then, can be expressed in the same form as for the orbital and spin momenta: $\sqrt{j(j+1)}(h/2\pi)$, and the component of angular momentum along any direction is a multiple of Planck's constant divided by 2π, namely, $m_j(h/2\pi)$, in which m_j can have any value between $+j$ and $-j$ in integer steps (*e.g.*, $\frac{5}{2}$, $\frac{3}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$, $-\frac{3}{2}$, $-\frac{5}{2}$). The quantum numbers n, l, j, m_j are often used to specify a state in the hydrogen atom. Small differences in energies of levels because of the spin–orbit interaction sometimes cause complexities in spectral lines that are known as the fine structure of spectral lines.

Fine structure

THE PERIODIC TABLE

Although there seems to be no simple justification for the fact, it is found that in any atom no two electrons have the same set of quantum numbers n, l, m_l, m_s (or n, l, j, m_j), defined above. Even when interactions between the electrons occur, and when magnetic interactions between an electron's spin and its orbital motion occur so that a new state of an electron is specified by a different set of four quantum numbers, no two electrons in an atom have the same set. This is the Pauli exclusion principle; its most direct application involves the explanation of the periodic table of the elements first developed by two chemists, Dmitry Ivanovich Mendeleyev of Russia and Lothar Meyer of Germany (*c.* 1869). This table (see Figure 8) lists the elements in rows in order of increasing atomic number Z; it is found that the elements in the same column have similar chemical properties.

	group Ia												IIIa	IVa	Va	VIa	VIIa	0
1s	1 H	IIa															1 H	2 He
2s, 2p	3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
3s, 3p	11 Na	12 Mg	IIIb	IVb	Vb	VIb	VIIb	— VIII —		Ib	IIb		13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4s 3d, 4p	19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5s 4d, 5p	37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6s, 4f 5d, 6p	55 Cs	56 Ba	57 La*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7s 6d, 5f	87 Fr	88 Ra	89 Ac‡	104 *Rf*	105 *Ha*													

* 4f, 5d	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
‡ 5f, 6d	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Figure 8: Periodic table of the elements. Left column indicates the subshells that are being filled as atomic number Z increases. The body of the table shows element symbols and Z. Elements with equal numbers of valence electrons—and hence similar spectroscopic and chemical behaviour—lie in columns. In the interior of the table, where different subshells have nearly the same energies and hence compete for electrons, similarities often extend laterally as well as vertically.

To see how elements fit into the periodic table, it is helpful to start with a hydrogen atom—a singly charged atomic nucleus and one electron—then add exterior electrons and positive charges to the nucleus one by one for successive elements. Each electron will make transitions to its lowest possible energy state. For hydrogen, this is the state with $n = 1$, $l = 0$, $m_l = 0$, $m_s = \frac{1}{2}$ or $-\frac{1}{2}$; these numbers specify the resulting configuration, or arrangement, of electrons of a hydrogen atom in its ground state. If a positive charge is added to the nucleus and a second external electron is added, it too will go to its lowest state, again $n = 1$, $l = 0$, $m_l = 0$, and with m_s different from that of the first electron (otherwise both electrons would have the same set of quantum numbers n, l, m_l, m_s and this would violate the Pauli exclusion principle). The resulting configuration is that of helium in its ground state. There are only two possible electron states with $n = 1$, and these are now occupied by electrons. It is said that the $n = 1$ shell (or orbital) is filled or closed. This closed shell is relatively stable and difficult to excite; helium is the first of the inert, or noble, gases.

The helium configuration

If a third electron is added (to make a lithium-like atom), it cannot occupy the $n = 1$, $l = 0$ shell with the others, because its quantum numbers would then be identical with those of one of the other electrons. This electron has its lowest possible state with $n = 2$. For this value of n, the orbital quantum number l can be either 0 or 1; a state for $l = 0$ has slightly lower energy. The quantum numbers of this electron are then $n = 2$, $l = 0$, $m_l = 0$, $m_s = +\frac{1}{2}$, or $-\frac{1}{2}$. The inner $n = 1$ shell is

Table 2: Ground Configurations of the Elements

	1s	2s	2p	3s	3p	4s	3d	4p	5s	4d	5p
1 H	1										
2 He	2										
3 Li		1									
4 Be		2									
5 B		2	1								
6 C		2	2								
7 N	$(1s^2)+$	2	3								
8 O		2	4								
9 F		2	5								
10 Ne		2	6								
11 Na				1							
12 Mg				2							
13 Al				2	1						
14 Si	$(1s^2$			2	2						
15 P	$2s^22p^6)+$			2	3						
16 S				2	4						
17 Cl				2	5						
18 Ar				2	6						
19 K						1					
20 Ca						2					
21 Sc						2	1				
22 Ti						2	2				
23 V						2	3				
24 Cr						1	5				
25 Mn						2	5				
26 Fe	$(1s^2$					2	6				
27 Co	$2s^22p^6$					2	7				
28 Ni	$3s^23p^6)+$					2	8				
29 Cu						1	10				
30 Zn						2	10				
31 Ga						2	10	1			
32 Ge						2	10	2			
33 As						2	10	3			
34 Se						2	10	4			
35 Br						2	10	5			
36 Kr						2	10	6			
37 Rb									1		
38 Sr									2		
39 Y									2	1	
40 Zr									2	2	
41 Nb									1	4	
42 Mo									1	5	
43 Tc									2	5	
44 Ru	$(1s^2$								1	7	
45 Rh	$2s^22p^6$								1	8	
46 Pd	$3s^23p^63d^{10}$									10	
47 Ag	$4s^24p^6)+$								1	10	
48 Cd									2	10	
49 In									2	10	1
50 Sn									2	10	2
51 Sb									2	10	3
52 Te									2	10	4
53 I									2	10	5
54 Xe									2	10	6

	6s	4f	5d	6p	7s	5f	6d
55 Cs	1						
56 Ba	2						
57 La	2		1				
58 Ce	2	1	1				
59 Pr	2	3					
60 Nd	2	4					
61 Pm	2	5					
62 Sm	2	6					
63 Eu	2	7					
64 Gd	2	7	1				
65 Tb	2	9					
66 Dy	2	10					
67 Ho	2	11					
68 Er	2	12					
69 Tm	2	13					
70 Yb	2	14					
71 Lu	2	14	1				
72 Hf	2	14	2				
73 Ta	2	14	3				
74 W	2	14	4				
75 Re	2	14	5				
76 Os	2	14	6				
77 Ir	2	14	7				
78 Pt	1	14	9				
79 Au	1	14	10				
80 Hg	2	14	10				
81 Tl	2	14	10	1			
82 Pb	2	14	10	2			
83 Bi	2	14	10	3			
84 Po	2	14	10	4			
85 At	2	14	10	5			
86 Rn	2	14	10	6			
87 Fr					1*		
88 Ra					2		
89 Ac					2		1
90 Th					2		2
91 Pa					2	2	1
92 U					2	3	1
93 Np					2	4	1
94 Pu					2	6	
95 Am					2	7	
96 Cm					2	7	1
97 Bk					2	9	
98 Cf					2*	10	
99 Es					2*	11	
100 Pm					2*	12	
101 Md					2*	13	
102 No					2*	14	
103 Lr					2*	14	1
104 Rf					?		
105 Ha					?		

*Unverified predictions or not wholly confirmed data.
Source: Data compiled by J.E. Mack with the help of C.E. Moore.

relatively stable and remains inert in chemical processes; the chemical and spectroscopic behaviour of this atom is similar in many ways to that of hydrogen, since lithium and hydrogen both have one external electron free to move (valence electron).

Addition of the next electron to produce a beryllium atom completes the partial shell or subshell $n = 2, l = 0$. A subshell consists of all electrons of the same principal and orbital quantum numbers. The atom is analogous to helium in that both atoms have two outer electrons. The $n = 2$ shell is not filled, however, because an electron with $n = 2$ can have $l = 1$. Outside the inner shell $n = 1$, there are six possible electron states with $l = 1$ because an electron can have any combination of $m = 1$, 0, or -1, and $m_s = +½$ or $-½$. As successive electrons are added to yield boron, carbon, nitrogen, oxygen, fluorine, and neon, the electrons take quantum numbers $n = 2$, $l = 1$, and all possible different combinations of m_l and m_s, until a total of six with $n = 2$, $l = 1$ have been added. This completes the $n = 2$ shell, having a total of eight electrons in two subshells. The resulting atom is neon, the second of the noble gases.

When the next (11th) electron is added, it cannot go into any state with $n = 1$ or $n = 2$ because these are already filled by electrons. The lowest available state is now $n = 3$, $l = 0$, $m_l = 0$, $m_s = ½$ or $-½$. This atom is sodium; its outer structure is analogous to those of lithium and hydrogen, and so it has similar spectroscopic and chemical properties. Addition of another electron, which takes quantum numbers $n = 3$, $l = 0$, $m_l = 0$, $m_s = ½$ or $-½$ (m_s opposite to that of the 11th electron) yields magnesium, similar in structure and properties to beryllium, which lies above it in the periodic table. Addition of another electron yields aluminum, similar to boron. As more electrons are added, the elements silicon, phosphorus, sulfur, chlorine, and argon are produced. The spectra, structures, and chemical behaviour of these are similar to those of the elements carbon through neon, respectively, because the numbers of outer (valence) electrons are the same. In a similar fashion, the remainder of the periodic table is constructed, with some irregularities caused by different subshells having nearly the same energies (see Table 2).

Electron nomenclature

As a shorthand method of indicating the electron configurations of atoms and ions, the letters *s, p, d, f, g, h,* . . . are used to denote electrons having, respectively, $l =$ 0, 1, 2, 3, 4, 5, ... (The letters *s, p, d, f* originate from spectral series that had the designations sharp, principal, diffuse, and fundamental before their origins were fully understood.) A number prefixed to the letters gives the value for n, and a superscript to the right of each letter indicates the number of electrons with those values of n and l. The configuration $2s^1$ represents a single electron with $n = 2$, $l = 0$ (the superscript 1 is usually omitted). The configuration $1s^22s^22p^3$ represents two electrons with $n = 1$, $l = 0$, two electrons with $n = 2$, $l = 0$, and three electrons with $n = 2$, $l = 1$. The ground state configurations for the known atoms are given in Table 2.

For the atoms in the first third of the periodic table, one can often specify both the total orbital angular momentum and the total spin angular momentum in a given state. The total orbital angular momentum is the sum of the orbital angular momenta from all the electrons, taking into account their quantized values and their signs (plus or minus); it has magnitude $\sqrt{L(L+1)}(h/2\pi)$, in which L is an integer. The possible values for L depend upon the individual l values and the orientations of their orbits for all the electrons comprising the atom. The total spin momentum has magnitude $\sqrt{S(S+1)}(h/2\pi)$, in which S is an integer or half an odd integer depending on whether the number of electrons is even or odd. The possible value of the total spin angular momentum can be found from consideration of all the possible orientations of electrons within the atom. (Usually, in taking the L and S values, only the valence, or outer shell, electrons are considered, because in a closed subshell, there are as many electrons quantized in one direction as there are in the opposite direction, with the result that their orbital and spin momenta add up to zero. Thus only the valence electrons contribute angular momenta to the whole atom.) The total of orbital plus spin angular momentum has the magnitude $\sqrt{J(J+1)}(h/2\pi)$, in which J can take any positive value from $L+S$ to $L-S$ in integer steps; *i.e.*, if $L=1$ and $S=\frac{3}{2}$, J can be $\frac{5}{2}$, $\frac{3}{2}$, or $\frac{1}{2}$. The remaining quantum number, m_J, specifies the orientation of the atom as a whole; m_J can take any value from $+J$ to $-J$ in integer steps. A term is the set of all states with given configuration, L, S, and J. If this specification can be made, it is called LS, or Russell–Saunders coupling (1925), after an astronomer, Henry Norris Russell, and a physicist, Frederick A. Saunders, of the United States. Energy levels are plotted as bars on a vertical energy scale. Atomic levels are labelled S, P, D, F for $L=0$, 1, 2, 3, ... The designation of such a level is written as ^{2S+1}L; *i.e.*, the level $^3D_1$ has $S=1$, $L=2$, $J=1$. Sometimes the term is preceded by the principal quantum number to indicate in which shell the state occurs—*e.g.*, $5\,^3D_1$. A multiplet is the set of all transitions between two terms.

For heavier atoms, magnetic interactions among the electrons often contrive to make L and S poorly defined, although J and m_J still remain constant quantities for a given state of an atom, a coupling scheme known as j–j coupling.

Principle governing internal structures

A simple unifying principle underlies all the internal structures of atoms: atoms or ions having the same number of electrons outside the same, closed electron shell, called an isoelectronic sequence, have quite similar spectra. This similarity is illustrated in Figure 9, which shows drawings to scale of some of the energy levels of one electron near the ground states of a hydrogen nucleus ($Z=1$), a helium nucleus ($Z=2$), and a calcium nucleus ($Z=20$). Similarities in spectra for ions having the same number of electrons are sufficiently close that the spectral lines from a highly ionized (multiple loss of electrons) element can often be identified by extrapolation from known isoelectronic spectra.

TRANSITIONS

An isolated atom or ion in some excited state spontaneously transfers to a lower state with the emission of one or more photons, thus ultimately going into its ground state. The time required for such a transition varies greatly according to the wavelength of the photon to be emitted; for a transition in which a photon of visible light (wavelength typically 500 nanometres or 0.5 micrometre) is emitted, a characteristic average emission time is 1–10 nanoseconds ($1\text{–}10\times10^{-9}$ second). The emission time is proportional, approximately, to the cube of the wavelength; *i.e.*, a transition giving rise to a photon having wavelength 5 micrometres will occur in a time characteristically 1–10 microseconds.

Not all conceivable transitions, however, are possible within an atom. Theoretical and empirical selection rules restrict the kinds of transitions that may occur. The first selection rule found (1924, by Otto Laporte, then a physicist of Germany) states that all states of an atom can be divided into two sets: odd and even. A transition may take place from an odd to an even state or vice versa, but not between states in the same set.

Another selection rule states that during a transition, only one electron changes values of n and l. Other selection rules state that J can change only by unity or zero in a transition, and that if L and S are well defined within the atom, L can change by not more than unity ($\Delta L=0,\pm1$) and S cannot change at all during a transition.

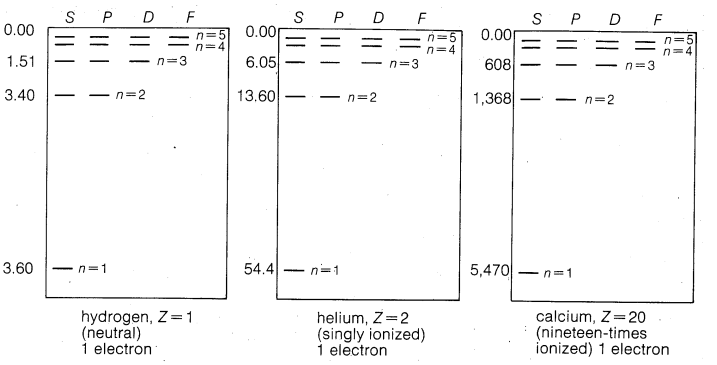


Figure 9: Level diagrams (Grotrian diagrams) for isoelectronic atom and ions, showing similarities in their level structure. The values for the principal quantum number n are for the electron in each case. The numbers at the left of the levels are the binding energies for the outermost electron, in electron volts; the electron is freed from the atom (or ion) at binding energy=zero.

Although these rules are not exact, and violations of them occur, the corresponding transitions for the violations are usually relatively slow and the spectral lines resulting are relatively weak. Such transitions are called forbidden transitions. For atoms in about the first third of the periodic table, the L and S selection rules provide useful criteria for the classification of unknown spectral lines. In heavier atoms, greater magnetic interactions among electrons cause L and S to be poorly defined, and these selection rules are less applicable. Occasionally, excited states are found that have lifetimes much longer than the average, because selection rules prohibit all transitions between levels. Such states are called metastable.

Metastable states

The transition rates from one state to another in all but the hydrogen-like atoms can be calculated only with difficulty. Accurate theoretical values can be obtained only with sophisticated computer routines, paying careful attention to the detailed structure of the atom. Experimental values for state lifetimes are of great importance in checking the accuracy of atomic theory.

PERTURBATIONS OF LEVELS

The energies of atomic levels are affected by external magnetic and electric fields in which atoms may be situated. The effect of a magnetic field (the Zeeman effect, after Pieter Zeeman, a Dutch physicist) is to cause an atomic level to split into its states of different m_J, each with slightly different energy. The result is that each spectral line separates into several closely spaced lines. The number and spacing of such lines depend upon the J values for the levels involved; hence the Zeeman effect is often used to identify the J values of levels in complex spectra. The required magnetic field is usually provided by a massive electromagnet that nearly surrounds the light source.

The corresponding effect of line splitting caused by the application of a strong electric field is known as the Stark effect (after Johannes Stark, a German physicist).

Small modifications to electronic energy levels arise because of the finite mass and nonzero volume of the atomic nucleus and the distribution of charges and currents within the nucleus. The resulting small energy changes, called hyperfine structure, are used to obtain information about the properties of nuclei and the distribution of the electron clouds near nuclei. Systematic changes in level positions are seen as the number of neutrons, N, in a nucleus is increased; these effects are known as isotope shifts. The principal isotope effects are: for light atoms (Z less than about 20), increased binding of electrons to

nuclei occurs, and decreased effects are seen caused by correlations in electron motion as the number of neutrons increases. For heavier atoms, a decreased binding of the electrons to the nucleus results as the neutron number increases, because the nuclear volume increases. The isotopic effects are so small that they are of no concern in molecular spectroscopy. The nucleus may behave as a small magnet because of circulating currents within; the magnetic fields produced in this way may affect the levels slightly. If the electric field outside the nucleus differs from that which would exist if the nucleus were concentrated at a point, this difference also can affect the energy levels of the surrounding electrons (see section V).

(J.O.S.Jr.)

III. Molecular spectra

The principal divisions of molecular spectroscopy are nuclear quadrupole resonance, nuclear magnetic resonance (NMR), electron paramagnetic resonance (EPR), gas phase microwave spectroscopy, electron spectroscopy, Raman spectroscopy, and infrared, visible, and near and far ultraviolet spectroscopy. The first three divisions are discussed in the article MAGNETIC RESONANCE, and the remainder are covered in the text below.

EXPERIMENTAL PRINCIPLES

The experimental arrangements for observing absorption spectra will normally involve a source of radiation; a device for selecting a narrow wavelength range, unless the source itself is monochromatic—*i.e.*, emits a single wavelength that can be changed at will; a cell for containing the sample of matter to be studied; and a radiation detector for measuring and recording the amount of radiation transmitted, both with the matter present and with it removed.

In the different spectral ranges, the sources of radiation for absorption spectroscopy will vary as will the detectors. In the microwave and radio-frequency regions, powerful monochromatic sources of controllable frequency have been available for some time. With the advent of the tunable laser (a monochromatic source of variable frequency in the optical regions), such sources became available in the early 1970s for the production of infrared, visible, and near ultraviolet. Otherwise, sources having a natural broad range of wavelengths combined with some optical device for selecting the desired monochromatic wavelengths, such as a prism or diffraction grating, are needed.

The photographic plate, the eye, photocells and photomultipliers, thermal detectors, and silicon rectifier crystals are among the detectors employed.

The molecular spectra of various substances are observed under a variety of conditions. Depending in part on the type of spectroscopy, the substances may be studied as gases, liquids, or solids, or in liquid or solid solutions, as single crystals, polycrystals, glasses, or as solid mixtures with other crystals. Similarly, circumstances make it advantageous to work in certain temperature ranges, sometimes at extremely low temperatures. Interesting work has been done with materials in unusual, possibly transient states, such as, for example, molecules that are ionized, parts of molecules that have been broken (*i.e.*, free radicals), highly reactive species trapped and preserved at low temperatures in inert, unreactive media such as frozen argon, special impurity centres in crystals, and molecules passing through a vacuum as a molecular beam. Each of these special circumstances confers advantages in particular cases.

THEORY OF MOLECULAR SPECTRA

Origins of molecular spectra. The spectra emitted by molecules can be much more complicated than those emitted by atoms because there are not only energy contributions from the electrons but also from the relative motions of the atomic nuclei. Molecules are made up of two or more atoms and hence consist of atomic nuclei surrounded by clouds of electrons. The moving electrons are attracted by the nuclei and thereby bind the latter together.

The energy of a molecule can be usefully divided into parts. The electrons have kinetic energy of motion and potential energy because of their attraction to the nuclei and repulsion by each other. Another potential energy originates in the mutual repulsions of the positive nuclei. These parts together constitute the electronic energy of the molecule, energy that is changed under electron or molecular impact, or when X-rays, ultraviolet, or visible light is absorbed, processes called excitation. X-rays primarily affect the inner atomic electrons; visible light or impact energy mainly excites the outer, so-called valence electrons.

Electronic energy of a molecule

The nuclei also can move, particularly in a back and forth vibrational motion with respect to one another. A molecule with several nuclei can display various patterns of vibration. Radiation in the infrared region may excite some of these vibrations.

Molecules also rotate about their centres of mass and thus possess rotational energy, at least when the substance exists in the gas phase. Finally, in a gas they are normally moving about as a whole—*i.e.*, executing translational motion.

The magnitude of these energies can be ordered, with the electronic energy of the inner electrons the greatest, that of the outer electrons next, followed by changes in vibrational energy, and, finally, by those in rotational energy. Changes in translational energy usually are not involved in spectra except for certain effects connected with the breadth of spectral lines.

There are some interactions, or interdependence, between these different kinds of energies, but the separation into kinds of energy, or energy parts, is nevertheless a concept of great practical importance. It is also true that changes in the smaller energy terms usually accompany changes in the larger ones. Thus, absorption of ultraviolet or visible light will normally change the vibrational and rotational energies as well as the electronic energy of a molecule. Similarly, in gases at least, rotational changes usually accompany vibrational changes brought about by absorption of infrared radiation.

Quantization of molecular energy. One consequence of quantum mechanics is that the energies of a molecule (except the translational energy) are quantized; *i.e.*, a molecule is not allowed by nature to possess an arbitrary amount of electronic, vibrational, or rotational energy; instead, only certain restricted discrete values of the energy are allowed.

Figure 10 shows the allowed values—that is, the energy levels—of the vibrational energy of a molecule composed of two atoms. The lower levels, designated as E_v, are nearly equally spaced as is predicted by quantum mechanics for a hypothetical case in which the force between the atoms of a molecule increases linearly (at a constant rate) with the increase in internuclear distance. For such a vibrating molecular system, called a harmonic oscillator, theory gives an energy for each level equal to a factor of some integer plus ½ (by substituting for v its values 0, 1, 2, etc., one gets ½, 3/2, 5/2, etc.) times a constant (ν_0), the vibration frequency calculated according to ordinary or classical mechanics and Planck's universal constant (h): $E_v = (v + ½)h\nu_0$, $v = 0, 1, 2, \ldots$, in which v, an integer, is called the vibrational quantum number. The actual levels converge upon one another as this quantum number increases, because the chemical bond (the force holding the molecule together) weakens and ultimately breaks when it is stretched sufficiently—*i.e.*, the constituent atoms are separated. Beyond the dissociation energy, when the bond breaks, all values of the relative translational kinetic energy of the atoms are allowed without any restriction from quantum mechanics.

Allowed energies for a harmonic oscillator

When a molecule is exposed to radiation, it is found from both theory and experiment that it will not efficiently absorb radiation and thereby undergo a transition from one energy level to another unless the energy of a quantum of radiation (which has the value $h\nu$, the product of the frequency of radiation ν, and Planck's constant h) matches the energy difference between the energy level initially occupied by the molecule and some higher level. If, on the other hand, the value of $h\nu$

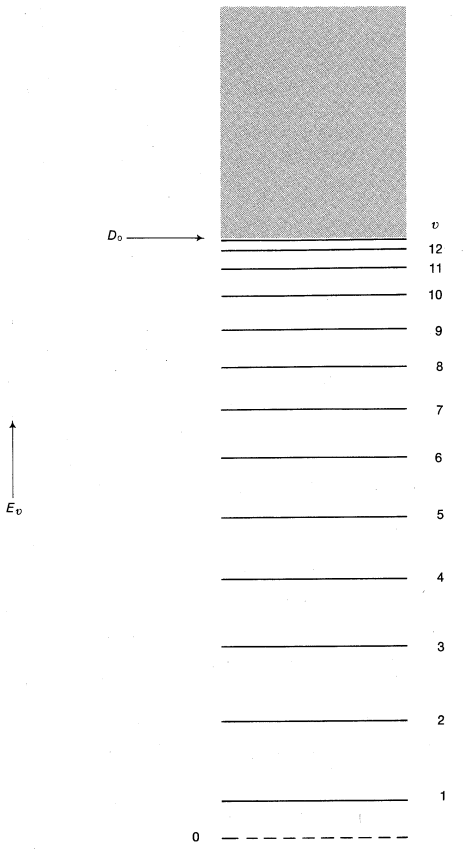


Figure 10: Allowed values of the vibrational energy E_v of a diatomic molecule. Levels are labelled with the quantum $v = 0, 1, 2, \ldots$. The zero energy (dashed line) is not an allowed value. The point D_0 is the dissociation limit, above which the molecule breaks into two separate atoms.

matches the energy difference to a lower energy level, the molecule may be caused to emit the frequency ν. This phenomenon is called induced emission.

Selection rules

Not only must energy be conserved (*i.e.*, $\pm h\nu = E_{\text{final}} - E_{\text{initial}}$) but also there may be additional restrictions, called selection rules, on what changes of energy level may occur. Thus for a polar diatomic molecule (a molecule with an unsymmetric charge distribution), vibrational transitions from one energy level to the adjacent level symbolized as $\Delta v = 1$ are much more probable than changes in the quantum number v by 2, 3, etc. These rules have a sound theoretical foundation, and often it is possible to make firm predictions regarding selection rules and intensities.

APPLICATIONS

The most obvious application of many of the branches of molecular spectroscopy is for characterizing and identifying specific chemical species—*i.e.*, for chemical analysis, either qualitative or quantitative. The infrared spectra of materials in the gaseous, liquid, or solid state at moderate resolution are highly characteristic of individual species—a sort of molecular fingerprint that, once taken, usually permits the same substance to be identified if encountered again. Raman spectra (after Sir Chandrasekhara Venkata Raman, a physicist of India) are similar in their utility. Pure rotational spectra (microwave spectra) are even more characteristic but are obtainable for far fewer species. Visible and ultraviolet spectra are usually considerably less structured than infrared spectra but are easy to observe, quantitatively measurable, often obtainable on tiny samples, and have been of great importance in chemistry. In a similar way, the closely related fluorescence spectra are of wide utility for analysis. X-ray spectra are mostly useful for identifying the atoms that are present. One of the most useful branches for these chemical applications is that of nuclear magnetic-resonance spectroscopy, which is discussed in the article MAGNETIC RESONANCE.

Besides these important analytical uses—important in controlling manufacturing processes as well as in the laboratory—spectroscopy has been one of the major sources of scientific information about the geometry of molecular structure, including much of the data on the way electrons are distributed in molecules and how they bring about chemical binding of the atoms. Some of these applications will be discussed below.

FIELDS OF MOLECULAR SPECTROSCOPY

Microwave spectroscopy. (See also section V.) Microwave radiation is electromagnetic radiation with wavelengths of a few millimetres to several centimetres (see Table 1). The region is differentiated from its neighbours (the far infrared at shorter wavelengths, ultrahigh frequency radio at longer wavelengths) only by the practical techniques of generation, transmission, and detection used. Microwaves can be generated by special electronic vacuum tubes—*e.g.*, klystrons—or by backward wave oscillators, and more recently by certain solid-state devices such as the Gunn effect diode. These sources can be made highly monochromatic.

Microwaves can be transmitted through hollow metal tubes, often of rectangular cross section (wave guides, see ELECTROMAGNETIC RADIATION) and can be detected by a silicon diode rectifier crystal connected to a suitable amplifier and recording or display device.

If a wave guide with mica windows at the ends is arranged to form an absorption cell and then evacuated and filled to a low pressure (perhaps one hundred thousandth of atmospheric pressure) with a suitable gas, such as carbonyl sulfide (COS), certain frequencies in the microwave region are weakly but sharply absorbed by the gas. Figure 11 shows the signal strength versus frequency of *p*-bromotoluene. In practice, some form of modulation of the absorption is generally used in order to increase the

By courtesy of Dr. L.H. Scharpen, Hewlett-Packard Co., Palo Alto, California

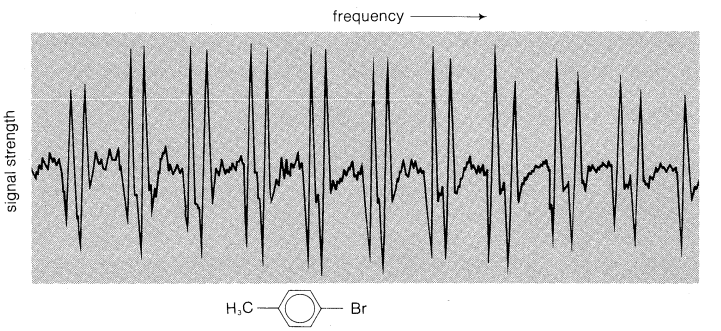


Figure 11: Microwave spectrum of *p*-bromotoluene at low resolution. Signal strength in absence of electric field is upward, in presence of field downward. Note pairs of peaks arising from two isotopes of bromine.

sensitivity. In almost all such cases, the molecule absorbs a photon or quantum of microwave energy and converts it to increased molecular-rotational energy; *i.e.*, the molecule undergoes a transition from one specific, quantized discrete rotational level to another.

Rotational levels

In order to display a sharp microwave spectrum, a substance normally must have a nonvanishing, permanent electrical dipole moment (an electric dipole moment means that the electric charge is not symmetrically distributed over the molecule); it must be in the vapour state at a low pressure; and its molecules must not be too large and complicated. Many thousands of kinds of molecules have already been found suitable for producing microwave spectra.

Detailed information about individual molecules that can be obtained from microwave spectroscopy includes quantitative molecular structures, electrical dipole moments, magnetic moments, nuclear-electrical quadrupole moments that measure the deviation of the nuclear charge from sphericity, nuclear spins, low vibration frequencies, barriers to internal rotation, molecular quadrupole moments, and information about chemical bonds. The

spectra are highly characteristic of each kind of molecule and hence can be used for analysis. Some transient species, such as the free radicals hydroxyl (OH), sulfur monoxide (SO), bromine monoxide (BrO), and others, have been studied.

Astronomers have used special radio telescopes to discover the presence in interstellar space of certain molecules—hydroxyl, water, ammonia, formaldehyde, and others—by means of characteristic, recognizable microwave lines.

A simple example of how structures can be obtained is provided by diatomic molecules, consisting of two atomic nuclei and linear molecules, for which an adequate model is provided by the linear-rigid rotor—a rotating rigid body made up of a set of mass points (each representing a nucleus) arranged in a straight line. The orientation is measured with the two polar angles of the axis, and the only parameter, or characteristic quantity, required is the moment of inertia (I) specified as a sum of terms, each of which is equal to the mass of a nucleus (electrons are not considered because of their relatively negligible mass) times the square of its distance along the axis from the centre of mass; *i.e.*, $I = m_1z_1^2 + m_2z_2^2$, in which m_1 and m_2 are the masses and z_1 and z_2 are their distances. For this model system, quantum mechanics leads to the result that the energy is inversely proportional to the moment of inertia (I) and dependent on a quantum number (J) that takes on integral values; *i.e.*, $E_J = J(J + 1)\ (h^2/8\pi^2I)$, in which $J = 0, 1, 2, \ldots$ for the allowed values of the rotational energy E_J. Here h is Planck's constant.

Although this system provides a good model for the rotational energy levels of diatomic and linear molecules, it is not perfect; for example, molecules are not really rigid but stretch because of the centrifugal force of the rotation. This characteristic and other interactions of rotation and vibration are observable but not large. In addition, electrons contribute a small amount to the moment of inertia.

From A.P. Cox and J.M. Riveros, *The Journal of Chemical Physics*, vol. 42, p. 3109 (1965)

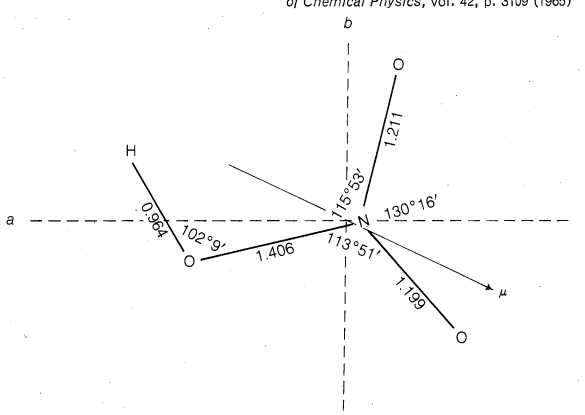


Figure 12: Structure of nitric acid molecule, determined by microwave spectroscopy. Principal axes *a* and *b*, and direction of dipole moment μ are also shown. Interatomic distances are in units of 10^{-8} centimetre (see text).

Quantum mechanics also provides selection rules (*i.e.*, prescribes the energy levels involved) for transitions in this case. If the molecule has a permanent dipole moment, emission or absorption of radiation can occur between adjacent levels; that is, the selection rule is obeyed by a difference in quantum numbers $\Delta J = 1$ for a pure rotational transition. Consequently, the spectrum should consist of a series of lines, equally spaced in frequency ν_J, which follows from the energy equation and selection rule: $h\nu_J = E_{J+1} - E_J = [(J + 1)\ (J + 2) - J\ (J + 1)]\ h^2/8\pi^2I = 2\ (J + 1)\ h^2/8\pi^2I$, in which the term $(J + 1)$ takes on the values 1, 2, 3, etc. This equation means that the transitions have frequency spacings $2h/8\pi^2I$. From a measured frequency of known quantum number J, the moment of inertia, I, for the particular kind of molecule can be obtained. For a diatomic molecule consisting of two nuclei of masses, m_1 and m_2, with internuclear distance r, the moment of inertia expression reduces to the product of the masses times the separation squared, divided by the sum of the masses, or, $I = m_1m_2\ r^2/(m_1 + m_2)$; so that a knowledge of the moment of inertia obtained from spectral measurements leads to a value for the internuclear separation. For nonlinear molecules, molecules having their nuclei arrayed other than in a straight line, the expressions for the rotational energy levels and transitions in terms of the moments of inertia are more complicated, but they have been developed and are widely applied.

Nonlinear molecules

For molecules with more atoms, the structure involves more parameters (internuclear distances) and the information from one microwave spectrum is not sufficient. The internuclear distances of molecules that differ only in that one or more nuclei are replaced by other isotopes are found to change but slightly, however, so that the microwave spectra of a number of such isotopic species can be used in order to provide sufficient information for the complete structure of a simple molecule. Figure 12 shows an example.

Infrared spectroscopy. Infrared radiation has a range of wavelengths from about 8×10^{-4} millimetre (.8 micron) to about 10^{-1} millimetre (100 microns)—*i.e.*, from visible light to microwaves. It is customary to divide the infrared region into near, medium, and far infrared domains. Photography can still be used out to about .0012 millimetre (1.2 microns) wavelength (the near infrared), and various special photoconductive detectors are available for part of the medium region, but at longer wavelengths one must use much less sensitive thermal detectors, such as thermocouples, bolometers, or Golay-cell detectors, all of which use the radiation to produce a temperature rise in some temperature-sensitive element.

Heated objects were the primary sources of infrared radiation before the discovery of lasers. The latter became available in the early 1970s for many specific sharp frequencies (narrow lines), whereas the thermal sources used theretofore gave radiation spread continuously over a broad range of wavelengths.

Present-day infrared spectrometers still use thermal sources, a prism of special material or a diffraction grating to select out narrow ranges of wavelength, mirrors instead of lenses for focussing the radiation, and either photoconducting or thermal detectors. Depending on the wavelength range, prisms and windows may be made of rock salt (sodium chloride), potassium bromide, quartz, or other materials chosen because they are transparent to the desired rays. Glass does not transmit except in the visible and adjacent infrared. The difficulty of finding suitable transparent materials for the remainder of the infrared region explains the use of concave mirrors instead of lenses and also favours the use of gratings over prisms. Grating instruments are commonly used in infrared studies (see *Methods for dispersing spectra*).

A different type of spectrometer than those employing prisms and gratings uses the interferometric principle and has the advantage of large aperture; *i.e.*, it accepts a large cone of light from the source and hence is particularly advantageous in the infrared where sources are weak and detectors are of low sensitivity.

Lasers are highly promising sources, and some spectroscopy is being done with them, but lasers tunable over an adequate range became feasible only in the early 1970s.

The main impact of infrared spectroscopy is probably in the medium region, say from 0.003 millimetre to perhaps 0.025 millimetre (or 3 to 25 microns), because most of the vibration frequencies of not too complicated molecules fall in this range.

The vibrational behaviour of a molecule is quite well simulated by a model consisting of point masses (the atomic nuclei) held together by weightless springs (representing the chemical-valence bonds). These binding forces are really caused by the electrons, which, being many thousands of times lighter than the nuclei, can be treated as approximately massless. If a macroscopic, mechanical model of a molecule made of steel balls (nuclei) and springs (bonds) is vibrated at a slowly increasing frequency, it will display a series of resonances (*i.e.*,

Resonant frequencies

certain frequencies at which the model will respond fairly sharply and vibrate in a particular way) with specific ratios of the amplitudes of motion of the different balls. Each resonance involves a different pattern of motion, these patterns being the normal modes of motion, each with its own frequency (see Figure 13).

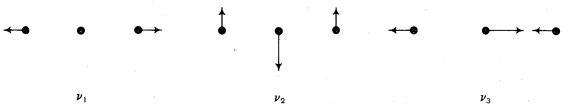


Figure 13: Normal modes of vibration (ν_1, ν_2, ν_3) of the linear molecule carbon dioxide, CO_2. Arrows represent the relative displacements of the three atoms in each of the modes.

Such mechanical models are not actually used, because classical mechanics permits the ready calculation of the frequencies and modes of motion of an idealized model with point masses and weightless springs. A molecule is much closer to this idealization than any achievable mechanical model. Actually, of course, a molecule obeys quantum mechanics and not classical mechanics. Fortunately, the quantum-mechanical results for the molecular model are closely related to the classical ones. Calculations for the infrared spectrum show absorptions only at frequencies equal to certain of the classical resonant frequencies discussed above. The frequencies that can appear are those for which the electrical dipole moment of the molecule changes during the normal mode of motion.

Further refinement of the calculations shows that other frequencies may occur weakly, for example, at twice those of certain classical frequencies. In addition, these infrared absorptions are not extremely sharp and for light gases under high resolution are seen to consist of bands of many sharp lines. This fine structure arises from simultaneous changes in rotational energy.

The infrared spectrum of a molecule is highly characteristic and hence is widely used by chemists to identify substances by comparing the substance to be identified with an authentic sample (see Figure 14). Moreover, it is found that certain groups within the molecule, such as the hydroxyl (OH), lead to the appearance of frequencies that are rather constant from one molecule to another and hence can be used to indicate the presence of the group in the molecule. Other groups that show their own frequencies include carbon–hydrogen (CH), sulfur–hydrogen (SH), double bonded carbon–oxygen C = O, and carbon–carbon double bond C = C. A great deal of empirical information of this nature has been accumulated.

From L. Bellamy, *The Infrared Spectra of Complex Molecules*, 2nd ed. (1958); Barnes & Noble Publishers, New York, and Methuen & Co. Ltd.

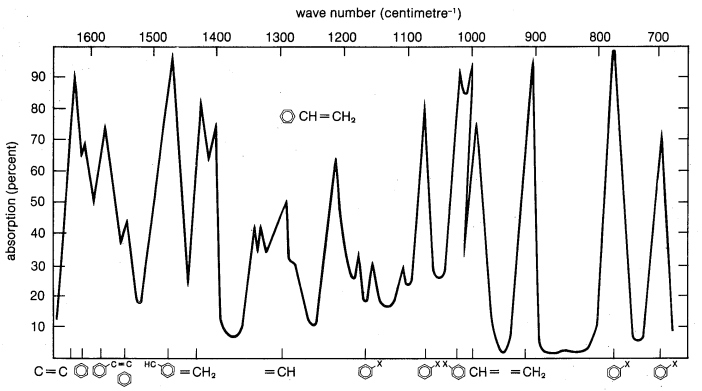


Figure 14: Vibrational infrared spectrum of styrene. The individual peaks are characteristic of the various structural features of the styrene molecule shown along the bottom edge of the figure. The hexagons represent the benzene ring of six carbon atoms present in the structure; X represents the entire group of carbon and hydrogen atoms attached to the benzene ring.

From a more theoretical viewpoint, infrared vibrational frequencies can be used to determine the forces between the atoms within a molecule, and they can also provide information about molecular symmetry.

The rotational fine structure, when observable, provides much the same kind of structural information as does the pure rotational spectra studied in the microwave region. The frequency resolving power in the infrared region is a great deal lower than in the microwave regions at present; on the other hand, nonpolar (uniform-charge distribution) molecules can be studied.

Raman spectroscopy. Raman spectroscopy makes use of the fact that photons can be scattered by molecules (whether in the gaseous, liquid, or solid state), with or without a change in frequency—*i.e.*, loss or gain in energy. Most scattered photons continue on with the same energy; a small percent lose some of their energy to the scattering molecule as additional vibrational or rotational energy; the remainder of the photons gain energy at the expense of the vibrational or rotational energy of the scattering molecule. Because the vibrational energy of a molecule is quantized (as is the rotational energy of a gas molecule), the changes in energy or frequency that occur are limited to a few discrete values. When using photons initially of a single frequency, the scattered radiation contains a small number of discrete frequencies corresponding to energies that have been shifted by rotational or vibrational energy intervals.

Laser beams for Raman scattering

Experimentally, modern Raman spectrometers use a strong monochromatic, visible laser beam to illuminate the sample, which can be quite small. The light scattered, usually at right angles to the incident beam, is passed into a spectrometer (normally of a grating type) that is scanned over the desired spectral range, and the spectral output is detected with a phototube and recorded as a function of frequency and intensity (see Figure 15).

If a Raman line has frequency ν resulting from incident radiation of frequency ν_0, the frequency shift $\nu - \nu_0$ (or $\nu_0 - \nu$ if ν is less than ν_0) is the frequency of a transition, usually vibrational, in the sample. The Raman effect therefore provides information about vibration frequencies, as does infrared spectroscopy. If the scattering molecule has any symmetry, however, the selection rules for the Raman and infrared transitions may differ.

Pure rotational Raman spectra have been observed in gases for small molecules and have provided structural information. For the Raman effect, the molecule can be polar or nonpolar.

Visible and ultraviolet spectra. Coloured substances absorb radiation having frequencies in the visible portion of the spectrum. Thus a solution that looks blue when viewed by transmitted white light does so because the material absorbs yellow light from it. Other materials may be transparent in the visible region but may or may not absorb in the near ultraviolet region. More and more substances show absorption bands as the region considered is moved farther into the ultraviolet.

These particular absorptions are almost always due to the excitation of electrons. As in atoms, the electronic energy in a molecule is quantized, so that certain discrete values are allowed. If the quantum of radiation is energetic enough to remove an electron or an atom from the molecule, however, the resulting system can have any relative translational energy (nonquantized) so that part of the spectrum can show regions of continuous absorption.

An approximate way of describing the behaviour of the electrons in a molecule has been found to be useful, though not exact. Whereas each electron is actually influenced in its motion by the instantaneous position of every other electron, an approximation is made by replacing this complicated interaction by an interaction that has been averaged over the motions of the other electrons. In this way, it is possible to treat the complex problem in a certain sense as if the electrons acted independently of each other. In quantum mechanics, the location of an electron in a molecule is determined by specifying the probability of its presence at any point. Thus a plot of the probability would have the appearance of a cloud. This probability distribution in space of a particular electron in the present approximation is then governed by a single-electron function called a molecular orbital (see CHEMICAL BONDING). For a given molecule there exists a whole set of these molecular orbitals (of various cloud-

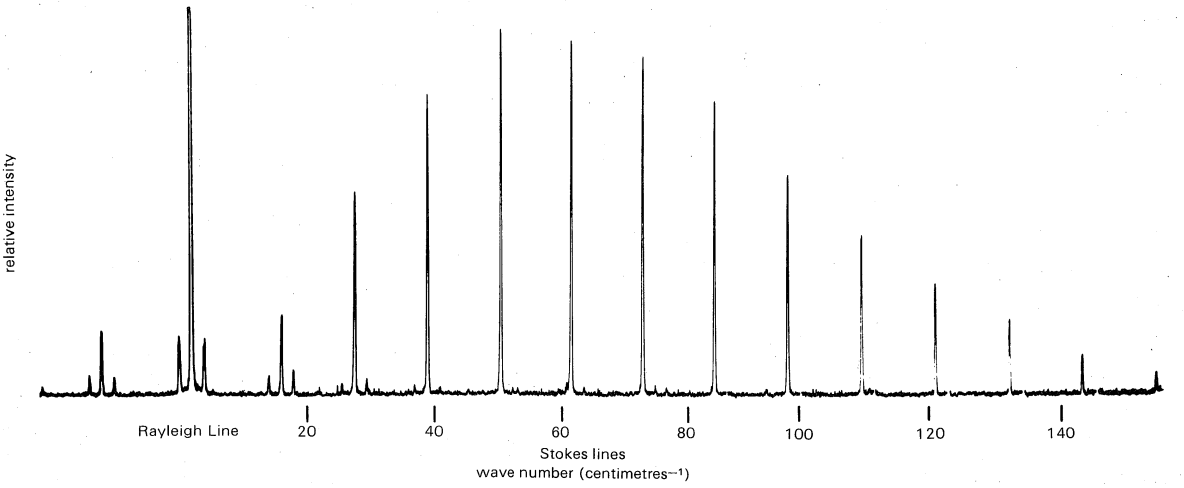


Figure 15: Pure rotational Raman spectrum of gaseous oxygen, O_2. Rayleigh line is light scattered without change of rotational or vibrational energy. Stokes lines involve an increase in rotational energy of the molecules.
By courtesy of the Pennsylvania State University, University Park

like shapes); and for the state of lowest energy for the whole molecule (called the ground state), the electrons are assigned to the molecular orbitals of lowest energy, two at a time. Two electrons can occupy a single molecular orbital if the two electrons have opposite electron spin orientations; the Pauli exclusion principle forbids occupancy of an orbital by more than two. In the simplest form of the theory, the electronic energy is then considered to be twice the sum of the energies of these occupied molecular orbitals.

Besides the filled molecular orbitals, there are others that are not filled when the molecule is in its ground state. A molecule, however, may be raised to an excited electronic state by absorption of radiation or by a collision. The excited states can be approximately described by assigning one or more electrons to previously unfilled molecular orbitals (MO's) of higher energy, leaving a corresponding number of holes among the previously fully occupied lower MO's. The transition of lowest frequency would involve transferring one electron from the highest MO occupied in the ground state to the lowest empty MO. See Figure 16.

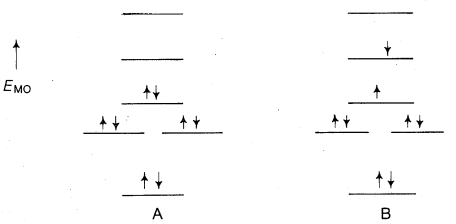


Figure 16: Example of an energy occupancy diagram for molecular orbitals (MO's). (A) For ground electronic state; each lower MO is occupied by a pair of electrons with opposite spins (arrows). (B) For lowest excited singlet state, in which an electron has been excited to a previously empty MO. Vertical position of bars indicates energy, E_{MO}, of individual MO's. Second level shows two MO's of same energy (degeneracy).

This method of labelling and classifying excited states and electronic transitions is of enormous value, but the mathematical approximations related to it do not always give good quantitative results. The missing energy correction, called the correlation energy, arises from the difference between the actual instantaneous repulsions of the electrons and the averaged interaction, and this energy can be large.

Correlation energy

Frequently, the additional approximation is made by describing a molecular orbital as a combination of two or more atomic orbitals from different atoms of the same molecule. The electronic orbitals of lowest energy are usually in fact describable as essentially pure inner-shell atomic orbitals. On the other hand, the highest occupied molecular orbitals, in a covalently bound molecule (each atom contributes one electron to the mutual bond) at least, will involve outer-atomic orbitals of more than one atom. Some of the MO's are largely responsible for the chemical binding energy of the molecule because they have lower energy than the atomic orbitals into which they transform when the atoms separate.

Molecules with simple, covalent electron-pair bonds with all inner-electron orbitals fully occupied tend to be transparent to visible radiation. In other words, there is a fairly large energy gap between the completely filled MO's and the lowest unfilled one.

There are two classes of molecules, on the other hand, that tend to be coloured (*i.e.*, to absorb in the visible region). One class consists of compounds of such transition-group metals as iron, nickel, platinum; that is, those with partially filled atomic orbitals. In many cases, electrons can undergo transitions to low-lying (inner-electron) orbitals of the same shell. (When the probability distributions in space of all the electrons of an atom in the ground state are considered, it is found that the electrons are grouped more or less in shells around the nucleus.) Another class contains molecules for which the simple bonding diagrams are inadequate; but there are several alternate ways of drawing in the lines representing chemical bonds, and the more correct representation (called conjugation or resonance) is a kind of quantum-mechanical superposition of the several bond structures. The two classic Kekule structures of benzene (Figure 17) provide a good example of resonance. The six carbon atoms of the molecule form a closed, six-sided structure that has six-fold symmetry (all sides equal, all angles equal) with all carbon–carbon bonds identical and with properties intermediate between those of double and single bonds. Theory shows that molecules displaying resonance should have low-lying empty MO's and therefore

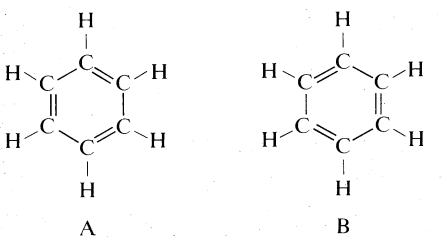


Figure 17: Two classic Kekule structures for benzene, C_6H_6, which show two equivalent ways of drawing alternate single and double bonds between neighbouring pairs of carbon atoms. In the actual molecule all six bonds are equivalent and the bonding is approximately the average of the two Kekule structures.

the possibility of visible or near-ultraviolet absorption bands. Organic dyes are in this category. Sometimes the MO's responsible for absorption are quite localized in part of the molecule. Such groups of atoms are called chromophores.

The fact that an electron is endowed with an intrinsic angular momentum (spin) that can display only two orientations, roughly speaking, in space gives rise to a number of spectroscopically interesting effects. The Pauli exclusion principle requires that if two electrons are in the same molecular (or atomic) orbital, they must have opposite spin orientations, cancelling their spin angular momenta, so that a majority of molecules exist in ground states with the electron spins completely paired. If one of these electrons is excited to another orbital that happens to have a vacancy, the atom or molecule will be excited to one of four energy states, grouped as a singlet state or a triplet level. The singlet state arises when the excited electron and its original partner are antiparallel (that is, have spins in the opposite direction); the triplet level, on the other hand, results when the excited electron and its original partner are parallel, having spins in the same direction. The combined two spins can have any one of three overall orientations, or more correctly, three possible values of the component of spin-angular momentum, versus only two for a single electron spin. There are many orbitals to which an excited electron may be raised, and consequently there exists a series each of singlet and triplet levels; an atom or molecule will thus lose energy by radiation by making transitions either from one singlet state to another singlet state or from one triplet set to another triplet set, transitions between singlet and triplet states being forbidden.

Singlet and triplet states

Many organic molecules, mostly so-called aromatic molecules, absorb in the visible or near ultraviolet and return rapidly to the ground state with the emission of nearly the same frequency as was absorbed. This phenomenon is fluorescence. The efficiency may be quite high; *i.e.*, nearly all the radiation absorbed is rapidly (in about 10^{-8} second) reradiated. Much of fluorescent light is of somewhat lower frequency than that absorbed, because the molecule can return, not only to the lowest state but also to one or another of the excited vibrational levels of the ground electronic state (see Figure 18). In some cases, this phenomenon leads to a fluorescence spectrum with a vibrational pattern that is approximately a mirror image of the pattern of the absorption spectrum.

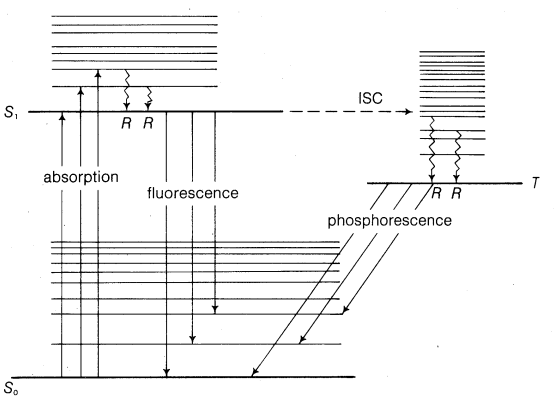


Figure 18: Lower vibrational and electronic energy levels of a complex molecule, showing ground electronic level S_0, the lowest excited singlet electronic state S_1, the lowest triplet electronic level T, plus associated vibrational levels for each. Zigzag arrows R show nonradiative loss of vibrational energy by collisions. ISC represents an intersystem crossing.

With some molecules another process competes with fluorescence. Absorption raises the molecule to an excited singlet state, perhaps with some vibrational excitation also. Collisions cause a rapid loss of vibrational energy but not of electronic energy. The molecule then has a certain probability of intersystem crossing—*i.e.*, of undergoing a transition, without radiation—to a highly excited vibrational state of the triplet manifold. Collisions again remove this vibrational energy, leaving the molecule in the triplet electronic state. From there it can return to the ground state either by radiation—*i.e.*, phosphorescence—or by another intersystem crossing to a high vibrational state of the ground, singlet, electronic manifold, with subsequent loss of vibrational energy. In various molecules, one or another of these pathways is used.

Triplet states generally have much longer lifetimes than singlet states, especially in solid glassy media at low temperatures, because the radiative transition to the ground singlet state is a forbidden one. Its occurrence is explained in terms of certain small interactions, in particular that between the magnetic moments of the electron spin and electronic orbital motion.

These phenomena are of chemical interest, for some compounds undergo different reactions in their excited states depending upon whether they are in the excited-singlet or excited-triplet state. Furthermore, these excitation energies can in suitable circumstances be transferred from one molecule to another, of the same or of different species.

One important application of visible and ultraviolet spectra is to the accurate determination of the energy required to dissociate a diatomic molecule into its constituent atoms. The energy-level pattern for the ground electronic state consists of a sequence of vibrational states, the lower members of which are approximately equally spaced (see equation in section above on quantization of molecular energy). As the vibrational energy is increased, the energy levels are more and more closely spaced and ultimately converge to a limit, which is the maximum vibrational energy possible without actually tearing apart the molecule into two atoms (see Figure 19).

Sequence of vibrational states

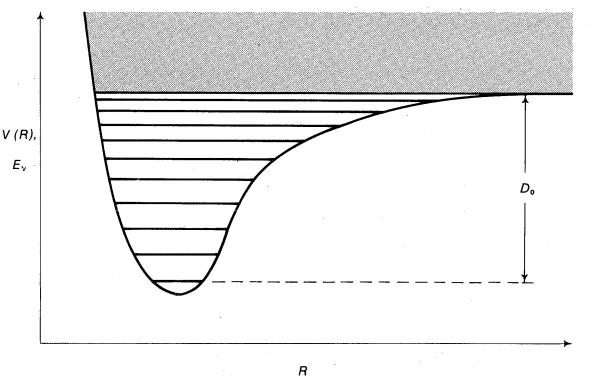


Figure 19: Potential energy function $V(R)$ and vibrational energy levels (E_V) for a diatomic molecule, showing dissociation limit at energy D_0 above lowest vibrational level.

In favourable cases, it is possible to observe an emission spectrum consisting of a series of lines that converges to a limit in the high-frequency direction. These lines are transitions from an excited electronic state to the various vibrational levels of the ground electronic state, as shown in the Figure. Planck's constant (h) times the difference in frequency between that of the limit and that of the transition to the lowest vibrational level is the energy of dissociation (D_0) and can sometimes be obtained with great accuracy. It is naturally necessary to know the states of the resulting atoms, because they need not be in their lowest states.

Electron spectroscopy. In most forms of spectroscopy, the fractions of the radiation for various wavelengths absorbed by the sample are measured. These data are then often used to determine the energy levels of the matter. There are other ways, however, of obtaining the same or similar information. In one of these, monochromatic radiation of quite high energy—in the X-ray or ultraviolet regions—is passed into the matter. Because of its high energy, it can cause the complete removal of an electron from the irradiated molecule, leaving the latter in the form of a positive ion (*i.e.*, an electrically charged molecule). The ejected electron will carry away as kinetic energy the excess energy, and this kinetic energy can then

be rather accurately measured by recently improved instruments. For example, Planck's constant (h) times the frequency (ν_0) of the incident photon is its energy ($h\nu_0$); if E_{elec} is the kinetic energy of the ejected electron, then the energy of the resulting ion, relative to the initial state of the neutral molecule, is equal to their differences, or $E_{ion} = h\nu_0 - E_{elec}$, by conservation of energy. In a typical experiment to determine an ionization energy, a gas may be irradiated by the 585-angstrom line of a helium discharge and the velocities of the electrons measured to give E_{elec}. Normally there will be several peaks in a curve plotted to show numbers of electrons versus energies. Each peak will represent one quantum state of the ion (provided that all the molecules were initially in their lowest energy level). Consequently, this method provides information about the states of the ion.

The ionic states can be approximately described in terms of the molecular orbital, now empty, from which the electron came. With ultraviolet excitation only outer electrons can be removed.

With X-ray excitation, electrons from the inner atomic shells (K or L electrons) can be ejected. The energies of these electrons are primarily characteristic of the particular atom involved; but there are small shifts from one molecule to another molecule, especially for molecules in which the chemical binding or environment of the particular atom differs markedly. These electron spectra (giving E_{elec} in the above equation) are therefore of use in identifying the atoms present and in obtaining some information about their environment. These two versions of electron spectroscopy were in the development stage in the 1970s, but these versions appeared likely to become quite important.

Related to electron spectroscopy are various experiments in which energy levels and other types of information are obtained by collisions between the molecule of interest and beams of electrons, neutrons, ions, or other particles. Various quantities can be measured, such as the energy loss of the stream of particles, or the angular and velocity dependence of this loss. One advantage of some of these collisional techniques is that they can sometimes detect transitions forbidden in optical spectroscopy, so that energy levels that are difficult to locate by conventional methods can be found.

Forbidden transitions

Molecular-beam spectroscopy. In the long-wavelength regions, the minimal width of a spectral line, and hence the achievable resolving power, is often limited by collisions between molecules, or between molecules and a wall of the container, or by the effect of the motion of the molecules along the direction of propagation of the radiation (Doppler effect). All of these effects can be largely eliminated by examining the molecules in the form of a narrow stream moving through an evacuated chamber in a direction at right angles to the direction of propagation of the radiation. Such a stream is called a molecular beam and can be produced by allowing a gas to flow through a small opening into an evacuated vessel.

In principle, extremely narrow spectral lines are thereby made possible, but the difficulty is that there are so few molecules in the path of the radiation that it is difficult to detect the absorption. A physicist, Isidor Rabi of the United States, solved this problem by reversing the usual procedure; instead of measuring the photons absorbed from the radiation beam by the molecules, he devised a scheme for measuring the number of molecules that have absorbed or emitted a photon. Figure 20 is a schematic diagram of a typical molecular-beam spectrometer, showing a source, s, from which the beam of molecules (or atoms) issues and a device, D_1, that deflects molecules in different quantum states by different amounts so that only those in one state (or sometimes one group of states) passes through the slit, S_2 into the central region. In this region, the molecules are exposed to radiation, the frequency of which can be varied; usually in the same region there is a constant magnetic or electric field as well. The beam then enters a second state selector, D_2, that passes on to the detector, d, only those molecules still in the quantum state selected by D_1. In operation, the radiation frequency is varied until its energy, $h\nu$, is

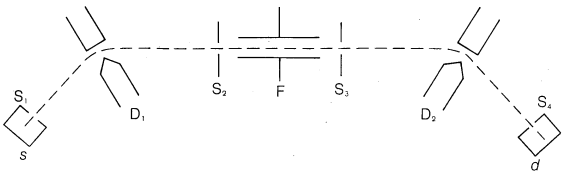


Figure 20: *Molecular beam spectrometer.*
Beam molecules (dashed line) originate in source s; S_1, S_2, S_3, S_4 are slits; D_1 and D_2 produce an inhomogeneous magnetic or electric field, which deflects molecules in different states by different amounts; F is the central region where molecules are subjected to oscillating electrical and magnetic fields; d is the molecule detector. Molecules that undergo a transition in F do not reach the detector.

equal to the energy transition (up or down) from the selected state of the molecules to some other state. If a transition to this other state is allowed, some of the molecules will then absorb or emit energy and thus be changed from their original state. Such molecules will not pass through the second state selector, D_2. The detector output will therefore show a minimum at this resonance frequency.

The heart of the device is clearly the state selectors D_1 and D_2. Several varieties are available. Small polar molecules, especially diatomic molecules, can often be separated according to their rotation–space orientation quantum numbers by an inhomogeneous (nonuniform) electric field. Molecules having sufficient rotational magnetic moments, even if they have no electrical dipole moment, may sometimes be state selected by an inhomogeneous magnetic field.

This method has been mostly used to observe transitions between different space-quantized states of the same rotational–vibrational level split apart in energy by an electrical or magnetic field, but transitions between different rotational states are also observed. The main application of the molecular-beam technique is to the study of hyperfine structures of transitions arising from tiny effects, such as interaction of the magnetic moments of nuclei with the magnetic moments generated by molecular rotation.

The molecular-beam method has been used to measure molecular electric dipole moments with great accuracy so that the small changes that occur from one vibrational state to another are easily measurable. These quantities are useful for testing theories of chemical binding. For example, a reasonably satisfactory and quite simple model for highly ionic molecules such as sodium chloride (in the vapour phase at high temperatures) attributes the chemical bond to the attraction of the oppositely charged ions balanced by the repulsion resulting from the interpenetration of the electron clouds of the ions. This picture can be made quantitative in a way that accounts for much molecular-beam data on ionic molecules.

Test of binding theories

Double and multiple photon effects. The interaction between matter and an electromagnetic field of radiation can usually be treated as a small effect that causes the matter to undergo occasional transitions with the absorption (or emission) of a single photon. With stronger radiation fields, however, transitions from one energy level to another can occur with the involvement of two (or even more) photons. Two photons can be absorbed simultaneously by the same molecule or one may be absorbed and one of a different frequency emitted (as in the Raman effect). Double photon absorption permits two different types of spectroscopy to be combined. For example, there can be simultaneous absorption of an infrared and a microwave photon. The development of lasers has provided sources of infrared and visible radiation strong enough to make these multiple-photon effects readily observable.

These various phenomena have extended the range of techniques available to the spectroscopist and have been most useful, therefore, for assisting in the analysis of complex spectra.

This brief discussion necessarily omits many topics, such as Mössbauer spectroscopy (see MOSSBAUER EFFECT), the effects of electric and magnetic fields, and

information obtainable from intensities and line widths. Molecular spectroscopy is a broad and growing field, continually developing new techniques and new capabilities for obtaining valuable information about molecules.

(E.B.W.)

IV. X-ray spectroscopy

Discovery of X-rays

A penetrating, electrically uncharged radiation was discovered in 1895 by a German physicist, Wilhelm Conrad Röntgen, and named X-radiation because its origin was unknown. This radiation was produced when electrons (cathode rays) struck glass or metal surfaces in high-voltage evacuated tubes, and was detected by the fluorescent glow of coated screens and by the exposure of photographic plates and films. The medical applications of this radiation that penetrated flesh more easily than bone were recognized immediately, and X-rays were in medical use in Vienna within three months of their discovery.

Initially, vain attempts were made to deflect the rays by means of electric and magnetic fields, on the assumption that they were composed of charged particles. Classical mechanics of the day showed that charged particles, such as cathode rays in an X-ray tube, when decelerated quickly from high to low speed in their collision with matter, should give off electromagnetic radiation (bremsstrahlung, or braking radiation), which in turn should show a wave character. Tests made to show the wave character of X-rays by diffraction were inconclusive, however, indicating that the wavelengths must be shorter than about 1×10^{-10} metre if X-rays were indeed waves. The first clear demonstrations of the wave nature of X-rays were provided in 1912 when X-rays were diffracted by the closely spaced atomic planes in a crystal of zinc blende (zinc sulfide). Because the details of the diffraction patterns depended upon the wavelength of radiation, these experiments set the stage for the spectroscopy of X-rays. The first spectrographs for this radiation were devised in 1912–13 by two British physicists—father and son—William Henry and Lawrence Bragg, who showed that not only did there exist continuum X-ray spectra, to be expected from processes involving the stopping of charged particles, but also discrete characteristic spectra, indicating that the details of atomic structure determine some X-ray properties (see Figure 2). The systematic increase of characteristic X-ray energies with atomic number was shown by Moseley in 1913 to be explainable on the basis of the Bohr theory of atomic structure.

Wavelengths for X-rays range from about 0.05 angstrom to 100 angstroms, with the range 10–100Å, known as soft X-rays, overlapping the far ultraviolet. A convenient unit of wavelength is the X-unit or X.U., defined so that approximately 1,000 X.U. = 1Å (the actual value is 1,000 X.U. = 1.00203 Å). The X-ray wavelength range is then approximately 50–100,000 X-units.

Relation to atomic structure. Short wavelength radiation can be produced by isolated atoms and ions by two related processes. If two or more electrons are removed from an atom, the remaining outer electrons are more tightly bound to the nucleus by its unbalanced charge, and transitions of these electrons from one level to another can result in photon wavelengths of 100 angstroms or less. The resulting spectra can be observed in spark sources. An alternate process occurs when in a neutral atom an electron is removed from an inner shell. This removal can be accomplished by bombarding the atom with electrons, protons, or other particles at sufficiently high energy and also by irradiation of the atom by sufficiently energetic X-rays. Almost immediately, the remaining electrons in the atom readjust, making transitions to fill the "hole" left by the removed electron, and X-rays are given off in these transitions. The latter process occurs in an ordinary X-ray tube, and the resultant X-ray lines, the characteristic spectrum, is superimposed on a spectrum of continuous radiation resulting from decelerated electrons.

The shells in an atom, labelled by $n = 1, 2, 3, 4, 5$, by the optical spectroscopist, are labelled *K, L, M, N, O,* . . . by the X-ray spectroscopist.

In many ways the vacancy, or "hole," remaining when an electron is removed from one of these shells, behaves as a positively charged electron; *e.g.*, if after an electron is removed from the *K* shell of an atom, an *L*-shell electron makes a transition to the *K* shell, the net result is the transfer of a single "hole" from the *K* to the *L* shell. The possible energies of such a vacancy thus can be labelled by the hydrogenic quantum numbers, *n, l, j, m*; because the production of a vacancy is equivalent to cancelling the negative charge at a region within the atom (*i.e.*, adding positive charge there), however, the energy of a vacancy is highest at the *K* shell, lowest in the outermost shell, and X-ray energy levels by convention are inverted compared to those of atoms having single outer electrons.

Inverted energy levels

Within a shell, there are some differences in vacancy energy according to whether or not an *s, p, d,* etc., electron has been removed and according to the total angular momentum of the electrons. The subshells are labelled by Roman numerals according to their energies. Figure 21 shows an X-ray energy-level diagram for uranium (*Z*, atomic number 92), including the allowed transitions.

From F.K. Richtmyer and E.H. Kennard, *Introduction to Modern Physics*, copyright (1947); used with permission of McGraw-Hill Book Company

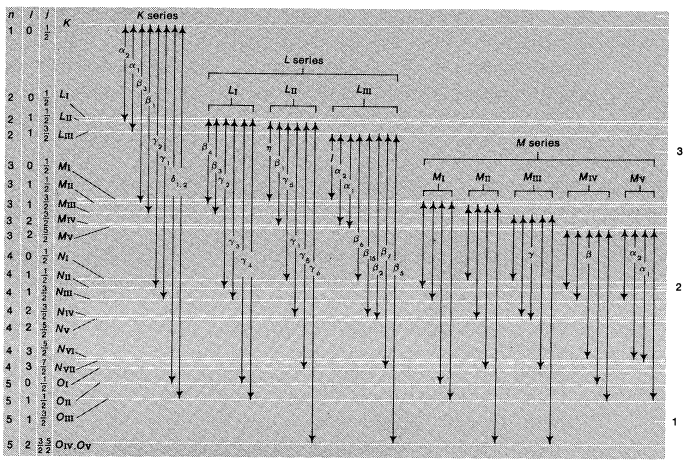


Figure 21: X-ray levels in uranium, showing transitions allowed by the selection rules $\Delta l = \pm 1$, and $\Delta j = \pm 1,0$. The ordinate scale is logarithmic. Since an X-ray level occurs when there is vacancy rather than an electron in a shell, the levels are inverted relative to hydrogenic levels.

If the continuous spectrum from an X-ray tube is passed through a pure absorbing material, it is found that the absorption coefficient changes suddenly at X-ray wavelengths corresponding to the *K, L, M,* etc., shell energies. The sudden increases of absorption coefficients (or absorption) as the wavelength is reduced past these shell energies are called absorption edges and serve as useful reference wavelengths that can be precisely determined (see Figure 2). Wavelengths of absorption edges range from 247.92 angstroms for the L_{II} edge in magnesium to below 0.1 angstrom for the *K* edges of elements heavier than plutonium ($Z \geq 94$).

The X-ray fluorescence radiation of materials is of considerable practical interest. Atoms irradiated by X-rays having sufficient energies, which may consist of either characteristic or continuous rays or both, lose electrons and as a result emit X-rays characteristic of their own structures. Such sources are weak; they have, however, an advantage over primary sources, such as X-ray tubes, in that there is little or no background caused by bremsstrahlung; the emitted spectrum is then nearly a pure-line spectrum. Such methods are used in the analyses of mixtures of unknown composition.

Production methods. An X-ray tube consists of a source of electrons (usually a heated filament) and a massive anode, mounted within an evacuated chamber or envelope (see Figure 3). A potential difference of 10–100 kilovolts is maintained between cathode (the negative electrode) and anode (the positive electrode), and typical power inputs range from 100 to several thousand watts. At the higher power levels, it is often necessary to cool the anode with circulating fluid. The anode materi-

al of an X-ray tube determines the range of wavelength for the discrete spectra, with the shortest discrete wavelengths produced by materials having highest atomic number. Bremsstrahlung, the continuous part of an X-ray spectrum, extends to a minimum wavelength (λ_0). It is found that this wavelength diminishes as the accelerating potential (V) is increased across the X-ray tube; in fact, the minimum wavelength and X-ray voltage are related by the Duane-Hunt law stating that electronic charge times voltage is equal to Planck's constant times speed of light divided by minimum wavelength—*i.e.*, $eV = hc/\lambda_0$.

If the X-ray fluorescence spectrum of some element is desired, a sample of the element is placed near the anode, usually outside the window of the tube where it is accessible. The potential voltage, of course, must be sufficiently high to release electrons from an inner shell (*e.g.*, at least 70 kilovolts to produce the characteristic K lines of tungsten). Baffles are placed in and around the tube to reduce the amount of primary radiation escaping from the tube other than through the exit window. Great care must always be taken in the operation of X-ray machines to avoid excessive exposure of any part of the body to the radiation.

Analysis. In the soft (10–100 angstroms) X-ray-wavelength region, it is possible to use grating spectrometers for analyzing spectra. Optical materials do not transmit radiation in this wavelength range, and they reflect this radiation inefficiently. It is necessary to use gratings at grazing incidence; that is, with the incident and reflected radiation moving nearly parallel to the grating.

Use of crystal spectrometers

For shorter wavelengths, crystal spectrometers are used. X-ray radiation falling on a set of geometrically aligned atoms is diffracted only in certain directions set by the requirement of constructive interference. A crystal spectrometer consists typically of an entrance slit, a bent crystal, and a detector, in a geometrical arrangement similar to that of a concave grating spectrometer. In such an instrument, the so-called bent crystal is a crystal of mica, quartz, germanium, or gold a few square centimetres in area and a millimetre or so thick, curved to a radius of a metre or more. The lattice spacing for the crystal is accurately known, and the observed angles of diffraction are used to measure and identify the wavelengths entering the spectrometer.

A source of X-rays of known wavelength can also be used to find the lattice spacing, crystal orientation, and crystal structure of an unknown crystalline material. The crystalline material is placed in a well-collimated beam of X-rays, and the angles of diffraction are recorded as a series of spots on photographic film. This method, known as the Laue method after a German physicist, Max Theodor Felix von Laue, is effective for studying crystals having dimensions of the order of a millimetre (0.04 inch) or larger. If microcrystalline materials (powdered crystals) are placed in the X-ray beam (Debye–Scherrer method, after Peter Debye and Paul Scherrer, chemists in Holland), the pattern observed is a series of concentric rings the diameters and intensities of which depend upon the lattice spacings of the crystal and its structure.

X-ray detectors. The most universally used detectors for X-ray spectroscopy are photographic plates and films and semiconductors (crystals that become conductive when irradiated).

For electrical recording of X-ray intensities, scintillators and ionization chambers are often used. A scintillator is a material that emits light when X-rays are absorbed. It is used in conjunction with a photomultiplier to provide an electrical signal, the intensity of which is proportional to the X-ray flux. An ionization chamber contains a positive and a negative electrode; when an X-ray photon enters the chamber through a thin window, it ionizes the gas inside and an ion current flows in the electrical circuit connected to the electrodes. The window is often beryllium, chosen for mechanical strength and relative transparency to X-rays. The gas is chosen to absorb strongly (*i.e.*, it gives a copious supply of ions) in the desired wavelength region. With increased voltage on the electrodes, the ionization chamber becomes a proportional counter, which gives out an amplified electrical pulse when an X-ray photon is absorbed within it. At still higher voltages, absorption of an X-ray photon with consequent ionization of many atoms in the gas initiates a discharge breakdown of the gas and causes a large electrical pulse output. This device is known as a Geiger–Müller tube (after two German physicists, Klaus Wilhelm Geiger, and Alex Müller) and forms the basis for radiation detectors known as Geiger counters.

Semiconductors

Semiconductor analyzer–detector combinations have replaced many of the other detectors. A semiconductor can serve as a detector because an X-ray photon passing through it creates electron-hole pairs, allowing charge to flow in the circuit attached to the semiconductor junction. The amount of charge can be proportional to the photon energy if the entire energy is absorbed in the junction; thus this device serves both as an energy analyzer and a detector of lines in an X-ray spectrum. Connected to a multichannel analyzer, the combination can record the number of photons per unit of energy and display them as a curve. In more sophisticated equipment, these data are automatically fed into a computer for analysis.

Applications. X-ray fluorescence and location of absorption edges can be used to identify quantitatively the elements present in animal tissues, paint, industrial chemicals, and many other substances. Emission of X-rays from high-temperature laboratory plasmas (a fourth state of matter, consisting of almost equal numbers of fast-moving particles of negative and positive charges) is used to probe the conditions within them. In the temperature range 10^{6}° K to 10^{7}° K, plasmas emit X-rays strongly; X-ray spectral measurements show both composition and temperature of a source. Similar measurements have been made on astronomical sources, such as the Sun. Because X-rays are absorbed by the Earth's atmosphere, such measurements are made above the atmosphere by apparatus carried by balloons, rockets, or satellites (see X-RAYS).

V. Radio-frequency spectra

The energy states of atoms, ions, molecules, and other particles are determined mostly by the mutual attraction of the electrons and the nucleus and by the mutual repulsion of the electrons. Electrons and nuclei have magnetic properties in addition to these electrostatic properties. The spin–orbit interaction has been discussed in the section on hydrogen-atom states. Other (usually weaker) magnetic interactions within the atom exist between the magnetic moments of different electrons and between the magnetic moment of each electron and the orbital motions of others. Separations between levels having different energies because of these magnetic interactions vary from less than 10^{8} hertz to greater than 10^{13} hertz, being generally greater for heavy atoms.

Origins. Nuclei of atoms often have intrinsic angular momentum, or nuclear spin, and magnetic moments because of the motions and intrinsic magnetic moments of their constituents; the interactions of nuclei with the magnetic fields of the circulating electrons change slightly the energies of the states in the atom. As a result, an atomic level that consists of several states having the same energy when the nucleus is nonmagnetic may be split into several closely spaced levels when the nucleus is one that has a magnetic moment. The levels will have different energies depending on the relative orientation of the nucleus and the magnetic field produced by the surrounding electrons. This additional structure of an atom's levels or of spectral lines caused by the magnetic properties of its nucleus is called magnetic hyperfine structure. The nuclear angular momentum is added to the orbital and electron-spin angular momenta to give the total angular momentum (F), the magnitude of which is $\sqrt{F(F+1)}\,(h/2\pi)$, in which h is Planck's constant, and the projection of which along the magnetic-field direction is $m_F h/2\pi$ in analogy with the quantum numbers J and m_J. Separations between levels differing only in the relative orientation of the magnetic field of the nucleus and electron range typically from 10^{6} hertz to 10^{10} hertz.

Hyperfine structure

Because of the motions and the intrinsic magnetic moments of its constituent electrons, an atom generally has a

total magnetic moment. The atom then has different energies when aligned in different ways in an applied magnetic field (Zeeman effect).

Atoms, ions, and molecules can make transitions from one state to another state that differ in energy because of one or more of these magnetic effects. Molecules also undergo transitions between rotational and vibrational states. Such transitions can be either spontaneous or can be induced by the application of appropriate external electric or magnetic fields at resonant frequencies. Transitions also can occur in atoms, molecules, and ions between high-lying electronic states near the ionization limit. The resulting spectra are known as radio-frequency (rf) spectra, or microwave spectra; they are observed typically in the frequency range from 10^6 to 10^{11} hertz.

The spontaneous-transition rate as an atom goes from an excited level to a lower one varies roughly as the cube of the frequency of the transition. Thus radio-frequency and microwave transitions occur spontaneously much less rapidly than do transitions at visible and ultraviolet frequencies. Because of this, most radio-frequency and microwave spectroscopy is done by forcing a sample of atoms to absorb radiation instead of waiting for it to emit radiation. These methods are facilitated by the availability of powerful electronic oscillators throughout this frequency range. The principal exception occurs in the field of radio astronomy; there emission spectra of astronomical sources, (*e.g.*, stars, galaxies, nebulae) are collected by large antennas and are amplified and detected by cooled, low-noise electronic devices.

Methods. The first measurements of the absorption spectra of molecules for the purpose of finding magnetic moments were made in 1940 by Rabi and his collaborators, using molecular and atomic beams (see above *Molecular-beam spectroscopy*). A beam focussed by magnets when no radio-frequency field was present was defocussed and lost when a frequency corresponding to one of the atomic absorption frequencies was applied to the coil, resulting in the atoms making transitions to other states. A spectrum then comprised a graph of the number of atoms coming through the magnets versus the frequency applied to the coil; the number of dips in this curve and the frequencies at which they occurred could be used to find the magnetic moment and angular momentum of the nucleus of the atom under investigation.

One of the most famous laboratory experiments with radio-frequency spectra was performed by two physicists, Willis E. Lamb, Jr. and Robert Curtis Retherford of the United States, in 1947. Their experiment involved a measurement of the energy difference between two almost-coincident levels in hydrogen, designated as $2^2S_{1/2}$ and $2^2P_{1/2}$. Although optical measurements had indicated that these levels might differ in energy, the measurements were difficult and were open to alternative interpretations. Atomic theory suggested that those levels should have exactly the same energy. Lamb and Retherford showed that hydrogen atoms in the $2^2S_{1/2}$ level would undergo a transition to the $2^2P_{1/2}$ level when subjected to radiation at about 1,058 megahertz; hence the levels did not have the same energy and the theory was incomplete. This energy separation in hydrogen (corresponding to a frequency of about 1,058 megahertz) is now known as the Lamb shift.

Radio-frequency measurements of intervals in ground levels and excited levels of atoms can be made by placing a sample of atoms (usually a vapour in a glass cell) within the coil of an oscillator and tuning the oscillator until a change is seen in the absorption of energy from the oscillator by the atoms. An alternative method is known as optical double resonance; here optical radiation corresponding to a transition in the atom of interest is passed through the cell and its intensity is monitored. If radio-frequency radiation is absorbed by the atoms in either of the levels involved, the intensity, polarization, or direction of reradiation of the absorbed light may be changed. In this way a sensitive optical measurement indicates whether or not a radio-frequency interval in the atom matches the frequency applied by the oscillator.

Optical double resonance

The operation of a maser (a type of microwave amplifier) depends upon the production of more atoms in the upper of two atomic levels separated by typically several hundred to several thousand megahertz. If such atoms are fed into a cavity that is tuned to the transition frequency, the atoms give off their excess energy to set up an oscillation in the cavity. Thus an oscillator is formed that resonates at the atomic frequency; this frequency can often be measured with extraordinary precision and can serve as a standard for time interval, called an atomic clock. The primary standard of time in the metric system is based on such an oscillator, which uses cesium vapour and oscillates at 9,192, 631, 770 hertz.

In those atoms in which the nucleus has a magnetic moment, the energies of the electrons depend slightly upon the orientation of the nucleus relative to the magnetic field produced by the electrons at the centre of the atom. The magnetic field at the nucleus depends somewhat upon the environment in which the atom is found, which in turn depends upon the atoms nearby. Thus the radio-frequency spectrum of a substance indicates both the constituents and the forms of chemical binding in the substance. Spectra resulting when the nucleus changes orientation are known as nuclear magnetic-resonance spectra, or NMR spectra, and are of considerable utility in identification of organic compounds (see MAGNETIC RESONANCE).

Radio-frequency transitions have been observed in astronomy in the frequency range from about 400 megahertz to about 9,000 megahertz (.75- to .033-metre wavelength). The earliest observation was of the 21-centimetre (1,420-megahertz) transition that occurs in the ground level of hydrogen as the proton spontaneously flips its spin direction relative to the electron from parallel ($F = 1$ states) to antiparallel ($F = 0$ states). Observations on this transition have provided much information about the temperature and density of hydrogen in the Earth's galaxy. Observation of the hydroxyl OH emission at 1,665 megahertz and other frequencies and of water emission at 2,220 megahertz indicates chemical composition and temperature for the regions in which they are observed. Transitions have been observed for which hydrogen atoms go from principal quantum numbers as high as n from 253 to 252, down to a transition as low as n from 91 to 90. Such atoms are presumably formed when protons capture electrons in large orbits; the spectra yield details of the local temperature and electric and magnetic fields. Lines caused by helium and possibly carbon have also been observed.

VI. Spectral analyses

The simplest method of analysis involves observation of a spectrum through a spectroscope. The advantages of this method are flexibility and speed; quantitative properties of a spectrum can be followed easily. Such observations, limited to the visible region, are useful for checking light sources for the presence of major constituents, because prominent, easily recognized lines or patterns can be found in the arc spectra of most elements. The transmitted spectrum of a solution can indicate the presence or absence of chemical substances.

ANALYSES OF PHOTOGRAPHIC DATA

Two principal classes of experiments in which photographic recording has found wide use are composition analysis and studies of the structure of atoms and ions.

In the simplest form of composition analysis, the spectrum of a source, consisting of a direct-current electric arc between two electrodes, is recorded and compared to spectra of pure elements by direct superposition. If the spectrograph possesses sufficiently high resolving power, the presence of an element in the source is indicated by coinciding lines on the two spectrograms. When quantitative determination is important, relative intensities may be compared. To do this, measured quantities of the sample the composition of which is being investigated may be mixed in several ratios with much larger quantities of graphite, copper, or other carrier material (called a matrix), and placed one at a time in the tip of the graphite or copper negative electrode of the arc. Several spectra of

Composition analysis

the sample material are thus obtained, in each of which one such electrode and a pure positive electrode (not of the element under observation) are used for the source. Comparison-arc spectra, each taken with the negative electrode of graphite or copper containing a small known amount of a pure element, are then compared with the spectra of the unknown material. If the operating characteristics of each of the arcs and of the spectrometer are maintained constant, and care is taken that all of the photographic films or plates are processed in the same way, it is possible to determine approximate amounts of many elements in concentrations ranging from greater than 50 percent down to a few parts per million.

As the admixture of any element in an arc is reduced, the weaker lines in its spectrum disappear first. The last lines to disappear are the most sensitive indicators of the presence of that element and are known as the persistent lines of the element.

If comparison spectra are not readily available for the element the presence of which is being investigated, it may be necessary for the investigator to measure wavelengths in the spectrum of the unknown material and compare these wavelengths with those tabulated for the suspected element. In order to facilitate measurement of wavelengths, a known reference spectrum (*e.g.*, that of iron) is usually exposed onto the same film without changing the spectrometer setting so that the known and unknown spectra are superimposed; wavelengths in the unknown spectrum are then calculated by interpolation between lines in the reference spectrum.

Study of the structure of a given atom or ion by means of its spectrum proceeds in a different manner. In such cases, it is necessary to make accurate measurements of wavelengths over as great a range as can be obtained. Reference wavelengths are again included on the photographic record of the spectrum of interest; if the greatest accuracy is desired, the reference wavelengths must have been measured in terms of the fundamental standard of length, the line of krypton-86 at about 6,057 angstroms, in terms of which the metre is defined. Determination of wavelengths in the spectrum of the atom of interest is again accomplished by interpolation between known reference lines, correcting where necessary for nonlinearity of the spectrograph.

After the wavelengths have been obtained, their wave numbers, proportional to the photon energies emitted by the atom, are calculated with correction for the index of refraction of air. Next, all sums and differences of wave numbers are calculated, and from recurring values of these quantities, relative positions of energy levels within the atom can be obtained. The levels are then classified according to their probable quantum numbers with the help of comparisons with isoelectronic spectra (atoms with the same number of electrons outside the last closed shell have similar spectra), application of selection rules, and study of the Zeeman effect in the spectrum.

For an element such as a rare earth or actinide, which may have a complex spectrum involving many thousands of lines, much of the analysis is automated. The photographic plate with its line spectrum is placed in a microdensitometer, which takes successive reading of the line positions across the plate and feeds these numbers directly into a digital computer. The numerical manipulations required for the analysis are then carried out electronically.

ANALYSES WITH SCANNED SPECTRA

The development of scanning spectrometers that use linear-response photoelectric detectors has revolutionized the study of details in the structures of spectral lines from laboratory and astrophysical sources. For example, changes amounting to only a few percent in the intensity of the Fraunhofer *D* lines, caused by the presence of sodium in the Earth's atmosphere, can be detected photoelectrically without difficulty but are difficult to observe photographically because of the nonlinear response of the photographic process. Changes of several percent in the widths (the measure of the spread in wavelength) of spectral lines from plasma sources can be followed if photoelectric detectors are used but not with photographic detectors; such changes may show changes in pressure or free-electron density within the discharge. In cases where two spectral lines overlap or blend, careful numerical analysis of data obtained by a photoelectric detector can often permit identification of positions and relative intensities of both lines; this analysis is far more difficult to do if the lines are recorded photographically and data are obtained by a microdensitometer.

Spectral widths

The quantum efficiency, the ratio of the number of incident photons to the number of recorded bits of information, can approach unity for photomultiplier detectors in the ultraviolet and visible regions, but the corresponding efficiency for photographic detectors is only about one part in 1,000; hence measurements with photoelectric detection are possible at lower light levels than with photographic detection. This sensitivity has made possible the recording of spectra from sources that send less than one photon per second into the spectrometer. Examples of studies in which low light levels are the rule include analyses of positions and profiles (*i.e.*, intensity changes across line widths) of spectral lines in weak discharges in a lamp at low temperatures to identify the details of processes leading to excitation, study of the spectra of fast ion beams excited by collisions with gases or thin foils, and high-resolution studies of absorption spectra of planetary atmospheres to identify their constituents.

In the production of a spark or in the initiation of a glow discharge, the effective temperature of the gas, the amount of ionization, and the relative excitations of different atomic levels change significantly within times of the order of microseconds. The details of such variations can be followed by photomultiplier detectors; in fact, photomultiplier detectors can be designed to follow changes occurring in less than a nanosecond (1×10^{-9} second). Time-resolved spectroscopy of sources is accomplished by arranging an electronic switch to operate the source repetitively, perhaps at line frequency of 50–60 times per second. A second switch, triggered by the first, sensitizes the detector photomultiplier for a brief interval at the desired time after the excitation pulses. The signals recorded by this photomultiplier depend on the spectrum only at that particular time after excitation. (J.O.S.Jr.)

BIBLIOGRAPHY

Atomic and X-ray spectroscopy: R.A. SAWYER, *Experimental Spectroscopy*, 3rd ed. (1963), treats in detail instruments and practical methods especially for photographic spectroscopy, and gives a survey of the historical development of spectroscopy, with many references. G. HERZBERG, *Atomspektren und Atomstruktur* (1936; Eng. trans., *Atomic Spectra and Atomic Structure*, 2nd ed., 1944), develops the elementary theory needed for the understanding of spectra; HARVEY ELLIOT WHITE, *Introduction to Atomic Spectra* (1934), is an older, but thorough treatment. Brief tabulations including the wavelengths of the brightest spectral lines emitted by most atoms may be found in R.C. WEAST (ed.), *Handbook of Chemistry and Physics*, 48th ed. (1967, revised periodically). Extensive tabulations of wavelengths may be found in G.R. HARRISON, *Wavelength Tables* (1939); A. ZAIDEL *et al.*, *Tables of Spectral Lines* (1970); and A.R. STRIGANOV and N.S. SVENTITSKII, *Tables of Spectral Lines of Neutral and Ionized Atoms* (1968). X-ray data, samples of optical spectra of some materials, tabulation of wavelengths, and details of methods of detection of radiation may be found in D.E. GRAY (ed.), *American Institute of Physics Handbook*, 2nd ed. (1963). A.E. SANDSTROM, "Experimental Methods of X-ray Spectroscopy," *Handbuch der Physik*, vol. 30 (1957), is a survey of X-ray spectroscopy, written in English, with many diagrams and tables.

Molecular spectroscopy: Elementary general books include: C.N. BANWELL, *Fundamentals of Molecular Spectroscopy* (1966); GORDON M. BARROW, *Introduction to Molecular Spectroscopy* (1962); and GERALD W. KING, *Spectroscopy and Molecular Sructure* (1964). Classic, authoritative works include: G. HERZBERG, *Molekülspektren und Molekülstruktur* (Eng. trans., *Molecular Spectra and Molecular Structure*, vol. 1, *Spectra of Diatomic Molecules*, 2nd ed., 1950; vol. 2, *Infrared and Raman Spectra of Polyatomic Molecules*, 1946; vol. 3, *Electronic Spectra and Electronic Structure of Polyatomic Molecules*, 1966); C.H. TOWNES and A.L. SCHAWLOW, *Microwave Spectroscopy* (1955); and N.F. RAMSEY, *Molecular Beams* (1963).

(J.O.S.Jr./E.B.W.)

Speech, Physiology of

Communication, though it may be defined in many ways, is, in essence, the transmission of information, questions, and intentions among creatures; in human communication the message includes thoughts and abstract ideas. Human communication by language emerged as man evolved from the other primates. Most animals communicate by acoustical, optical, and tactile means of transmission. Thus, most mammals have a laryngeal voice (from vocal cords); many lower vertebrates use simpler forms of respiratory generation of sound waves; and various insects possess some sound-generating apparatus on their legs, wings, or body to produce audible vibrations in the manner of crickets, cicadas, and grasshoppers.

Animal communication. Animal communication systems vary greatly in the complexity of the codes used in transmission and thus in the amount of information that can be processed at one time. Stridulating insects have but a small repertory of stereotyped sounds that vary only in rhythmical pattern. With evolution, the number of vocal patterns increases. The chicken produces a number of cries, each expressing a specific message: the cock's crow, the announcement that an egg has just been laid, the incessant clucking of the mother hen to keep the chicks together, the subdued warning that a preying bird has been sighted, the excited shrieks when the corn basket shows up, the contented cooing while resting in warm sand. All these vocal patterns are directed by chemical (hormonal) influences within the chicken's body; the rooster, for example, loses his crow after castration. The considerable variability of vocalization is well known among cats and dogs. Anger and joy, fear and fury, hunger and satiation, submission and aggression, rejection and welcome, and the love calls in particular are species specific, but they also remain stereotyped. Animals cannot learn calls beyond their inherited potentials. Each call is controlled by specific brain mechanisms under the influence of various hormones regulated by the pituitary gland and the brain. All of the systems of audible communication presuppose some kind of receiving organ; *e.g.*, the mammalian ear.

Human communication. While many animals possess voices of various types and inflections, man has learned to modulate his voice by articulating the laryngeal tone into audible oral speech. The terms speech and voice are by no means interchangeable: voice is the result of laryngeal sound production (phonation); speech is the superposition of oral and nasal resonances through movement of the oral articulators (see Figure 1). Articulation in itself is meaningless, until the conventional codes of a language have been learned, ordinarily first from the mother, then from other family members, and finally in school. "Sick," for example, means "ill" in English, while *sic* means "so" or "thus" in Latin. The difference between speech and language is one of application. Speech is the oral, audible expression of human language while talking. Writing and reading represent the use of human language through graphic symbols. The Morse code, naval flag signals, smoke signals of American Indians, African drum languages, and the almost extinct whistle language of the Pyrenees are all examples of special language usage without the need for oral speech. Each of these language systems has its own mode of transmission (electronic equipment, light, and air) and each is limited by darkness, distance, obstacles, etc.

Speech and language differentiated

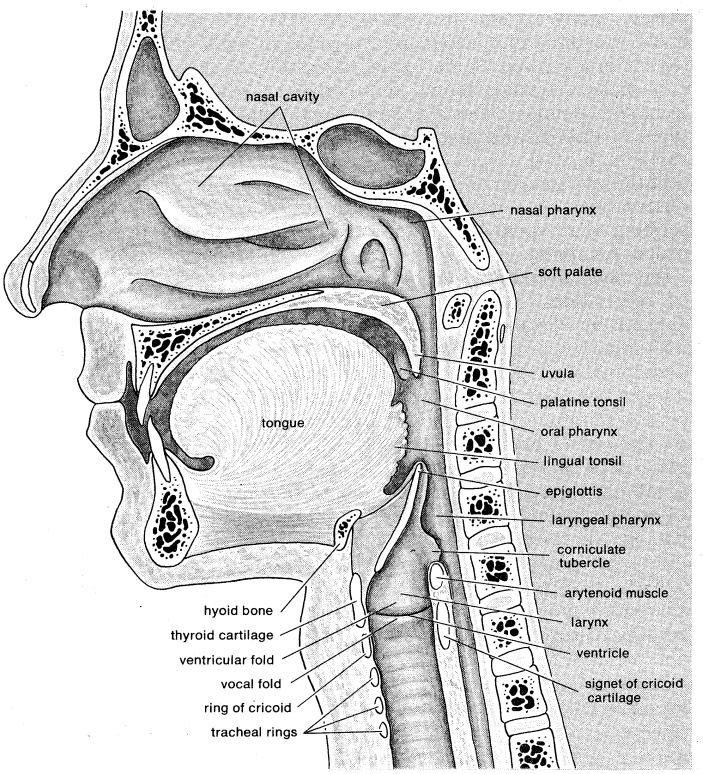


Figure 1: Sagittal section of head and neck.

Man became characteristically human when he learned to speak, perhaps 1,000,000 or more years ago, although the exact time will probably never be known. The ability to use symbolic and abstract spoken language is so far specifically human, though psychologists who have investigated the possibility of teaching spoken language to chimpanzees raised like human children in the family home have had results that, while limited, suggested that great affection and intensive training can teach a nonhuman some primitive beginnings of symbolic verbal behaviour (*e.g.*, sounds resembling "mama").

Man differs from the higher animals in several respects: he makes no mechanical use of his mouth in food gathering or tool making, and thus has freed his oral cavity for such activities as verbal articulation. He developed manual dexterity, the preferred use of one leading and dominant hand for skilled manipulation, drawing, and (later) writing. Preferred laterality (left or right) of hand, foot, eye, face muscles, tongue, and even ears is linked to the development of a distinctively human (known in no other animal) dominance of one brain hemisphere over the other in speech learning. Many nerve pathways from either side of the brain cross to the other side of the body. Thus, the human brain became ready for specialized language function in the dominant side of the brain, usually on the left side. It may be inferred that language had such precursors as tool making and cerebral dominance. Conversely, it would seem that prehistoric relics of fire, tools, weapons, and utensils indicate that the makers of such artifacts must have possessed some language. All human culture (and its propagation) appears to depend on some form of language to serve as the vehicle for teaching, persuasion, and tradition. Indeed, cultural impoverishment is seen in those with little or no language and speech: deaf-mutes, the mentally retarded, or the abnormally isolated. Without special instruction, learning is greatly impeded in deaf children, although the potential for great intelligence may be present.

Communication systems. Some social insects, such as the bees and ants, have a fairly complex system of communication. Biologists have described the stereotyped dancing movements used by bees to transmit specific information regarding food gathering, temperature control in the hive, dangers, etc. Observation of an anthill shows how the ants "trill" on each other's antennae to organize the various activities, albeit with seemingly senseless redundancy of effort.

Most creatures rely on the transmission of information through some coded system of communication. All of these systems are based on an elementary triad: an encoding sender (*e.g.*, the bodily structures that produce sound), a transmitting medium (*e.g.*, air), and a decoding receiver (*e.g.*, the ear and associated structures). At both ends of the chain, some kind of nervous system functions in the processes of encoding and decoding. The mechanical function of emitting and receiving the coded information is carried out by various types of impulse

Common features of communication systems

generators (*e.g.*, vocal cords) and stimulus receptors such as the ear. The connecting medium must simply be present, such as air or water for sound (vibration), open space for light, metal for electricity, etc.

The role of the nervous system. The perfection of human language thus depends on the learning potential of the nervous system, particularly the human brain. Some species have evolved highly specialized sensing systems. Bats produce laryngeal impulses in the ultrasonic range near 200,000 cycles per second; they are called ultrasonic to indicate that man cannot hear them. The reflections of these high-frequency waves from the irregular walls in dark caves, or from the body of a flying insect, are received by the bat's ears during flight and serve as a precise means of orientation by echo localization. Some mammals, such as mice and dogs, are known to hear in the ultrasonic range and surpass humans in this respect. Despite their frequently remarkable sensory talents, most animals are typically unable to learn human spoken language; their brains are not sufficiently specialized; for example, signs of cerebral dominance have not been found. Animals cannot learn to speak because their sound generators (parts of the mouth and throat) are not sufficiently flexible mechanically or sufficiently free from concentrating on the task of food getting.

Primary and secondary function. Bodily organs all function primarily to maintain life. The human larynx, for example, evolved as a valve that prevents the aspiration of fluids or solids into the lungs. Various reflexes (coughing, gagging, vomiting, sneezing, etc.) are inborn and thus neurologically predetermined as a further protective device. A normal baby at birth exhibits such basic reflexes; these are governed by specific centres in the brain. Sucking, swallowing, breathing, and other built-in (nonlearned) vital functions proceed automatically and usually beyond voluntary control.

Secondary, superimposed, learned functions also depend on brain mechanisms and the bodily organs they control. The articulated movements of speech have evolved from inborn reflexes serving food intake and breathing. Producing voice by phonation is a modification of exhaling. Voice without language is a reflex, possessed by many animals including the normal human baby, which cries out with the first breath of life. But speech is a superimposed, secondary function that each infant must learn from someone else.

The distinction between primary and secondary functions is helpful in understanding the frequent occurrence of disorders of voice and speech despite the presence of seemingly normal organs for their production. Inborn reflexes seem to have emerged early in animal evolution; such primary functions are thus said to be phylogenetically ancient, having appeared in lower animals many millions of years ago. The superimposed secondary functions, which have been learned by man and must be learned anew by each infant, are phylogenetically very recent. Many examples in medicine suggest that the old primary functions are more resistant to disease and derangement than are the secondary. In cases of human emotional disorder, speech may be severely distorted or totally absent, while the reflex functions of the person's throat and larynx remain perfectly intact.

Production of voice sounds and articulated speech

Respiratory mechanisms

Human speech is served by a bellows-like respiratory activator, which furnishes the driving energy; a phonating sound generator in the larynx (low in the throat) to transform the energy; a sound-molding resonator in the pharynx (higher in the throat), where the individual voice pattern is shaped; and a speech-forming articulator in the oral cavity (mouth). Normally, but not necessarily, the four structures function in close coordination. Audible speech without any voice is possible during toneless whisper; there can be phonation without oral articulation as in some aspects of yodeling that depend on pharyngeal and laryngeal changes. Silent articulation without breath and voice may be used for lipreading. While the anatomy and physiology of respiration are discussed in the articles RESPIRATION, HUMAN; RESPIRATORY SYSTEM, HUMAN, a few comments on phonic breathing are in order here.

Differences between quiet breathing and phonic respiration

An early achievement in experimental phonetics at about the end of the 19th century was a description of the differences between quiet breathing and phonic (speaking) respiration. One typically breathes approximately 18 to 20 times per minute during rest and much more frequently under strenuous effort. Quiet respiration at rest as well as deep respiration in physical effort are characterized by symmetry and synchrony of inhalation (inspiration) and exhalation (expiration). Inspiration and expiration are equally long, equally deep, and transport the same amount of air during the same period of time, approximately one pint of air per breath at rest. Recordings (with a device called a pneumograph) of respiratory movements during rest depict a curve in which peaks are followed by valleys in fairly regular alternation.

Phonic respiration is different; inhalation is much deeper than it is during rest and much more rapid. After one takes this deep breath (one or two quarts of air), phonic exhalation proceeds slowly and fairly regularly for as long as the spoken utterance lasts. Trained speakers and singers are able to phonate on one breath for at least 30 seconds, often for as much as 45 seconds, and exceptionally up to one minute. The period during which one can hold a tone on one breath with moderate effort is called his maximum phonation time; it depends on such factors as body physiology, state of health, age, body size, physical training, and the competence of the laryngeal voice generator—that is, the ability of the glottis (the vocal cords and the opening between them) to convert the moving energy of the breath stream into audible sound. Phonation time is markedly reduced in all the laryngeal diseases and disorders that weaken the precision of glottal closure, in which the cords (vocal folds) come close together, for phonation.

Repiratory movements when one is awake and asleep, at rest and at work, silent and speaking are under constant regulation by the nervous system. Specific respiratory centres within the brain stem regulate the details of respiratory mechanics according to the body needs of the moment. Conversely, the impact of emotions is heard immediately in the manner in which respiration drives the phonic generator; the timid voice of fear, the barking voice of fury, the feeble monotony of melancholy, or the raucous vehemence during agitation are examples. Conversely, many organic diseases of the nervous system or of the breathing mechanism are projected in the sound of the sufferer's voice. Some forms of nervous system disease make the voice sound tremulous; the voice of the asthmatic sounds laboured and short winded; certain types of disease affecting a part of the brain called the cerebellum cause respiration to be forced and strained so that the voice becomes extremely low and grunting. Such observations have led to the traditional practice of prescribing that vocal education begin with exercises in proper breathing.

Types of breathing

The mechanism of phonic breathing involves three types of respiration: (1) predominantly pectoral breathing (chiefly by elevation of the chest), (2) predominantly abdominal breathing (through marked movements of the abdominal wall), (3) optimal combination of both (with widening of the lower chest). The female uses upper chest respiration predominantly, the male relies primarily on abdominal breathing. Many voice coaches stress the ideal of a mixture of pectoral (chest) and abdominal breathing for economy of movement. Any exaggeration of one particular breathing habit is impractical and may damage the voice.

LARYNGEAL ANATOMY

The morphology (structure) of the larynx is studied according to the cartilages, muscles, nerves, blood vessels, and membranes of which it is composed.

Cartilages of the larynx. The frame or skeleton of the larynx is composed of several cartilages, three single and three pairs (see RESPIRATION, HUMAN). Single cartilages are the shield-shaped thyroid in front, whose prominence forms the "Adam's apple" in the male; the cricoid carti-

lage below, which resembles a signet ring and connects the thyroid to the trachea or windpipe; and the leaf-shaped epiglottis, or laryngeal lid, on top. Among the paired cartilages are the two arytenoids, which ride on the cricoid plate and move the vocal cords sideways; the two corniculate cartilages of Santorini on top of the arytenoids; and the two cuneiform cartilages of Wrisberg. The cartilages are held together by ligaments and membranes, particularly around their joints. The larynx is connected below to the uppermost ring of the trachea (see Figure 1 again), while above it is connected by the thyrohyoid ligaments to the hyoid bone beneath the tongue. Most of the laryngeal cartilages ossify (turn to bone) to variable degrees with age under the influence of masculinizing hormones. This fact is an important sign in the X-ray diagnosis of certain vocal disorders. If a man shows less ossification than is normal for his age, he may be deficient in male hormones; this may also account for an effeminate sound in his voice. Conversely, when a woman shows increased laryngeal ossification, she may suffer from virilizing hormones, which might also explain any lowering and roughening in her voice.

Laryngeal muscles. There are two types of laryngeal muscles, the external (extrinsic) ones, which move the larynx as a whole, and the internal (intrinsic) ones, which move the vocal folds to shape the glottis. It is helpful to remember that the anatomical names of most such muscles are derived from their origin on one structure to their insertion on another.

The extrinsic muscles comprise the thyropharyngeus, extending from the posterior border of the thyroid cartilage to the pharyngeal constrictor muscle, and the cricopharyngeus, from the cricoid cartilage to the lower portion of the pharynx and the opening of the esophagus (the food tube connecting the mouth and stomach). This cricopharyngeus muscle aids in the closing of the esophagus whenever it is not open for swallowing. Under the influence of emotional tension, the cricopharyngeus muscle may go into a spasm, which leads to a painful sensation of tightening in the throat, usually described as a "lump in the throat." A disorder of this sort (previously referred to as globus hystericus) is now believed to be a sensation of cricopharyngeus spasm from emotional tension or imbalance via excessive activity of the autonomic (involuntary) nervous system.

"Lump in the throat"

Extrinsic muscles. Although it is situated outside the laryngeal cartilages, the short cricothyroid muscle, a triangular muscle between the respective two cartilages, is traditionally discussed among the intrinsic (internal) muscles. Whenever this muscle contracts, the cricoid and thyroid cartilages are brought together anteriorly. This moves the anterior (forward) insertion of the vocal cords inside the thyroid wing forward, while their posterior (backward) insertion on the arytenoid cartilages is shifted backward. From this rotation results a marked elongation of the vocal folds clearly visible on X-ray films. This stretching action is the chief mechanism for raising the pitch of the sound generated and thus for the differentiation of vocal registers (*e.g.*, chest voice, falsetto). For embryologic reasons, the cricothyroid is the only laryngeal muscle that has its own nerve supply from the superior laryngeal nerve, a high branch of the vagus nerve (which issues from the brainstem). All other laryngeal muscles are innervated by the recurrent or inferior (low) laryngeal nerve, a low branch of the vagus nerve. This fact is important in the diagnosis of laryngeal paralysis because the resulting immobilization of the vocal cord and the remaining vocal function depend on the type of paralysis; *i.e.*, whether only the high or the low nerve or both of the laryngeal nerves are paralyzed on one side.

Intrinsic muscles. The intrinsic muscles include all of the following. The thyroarytenoid muscle extends from the inside of the anterior edge of the thyroid cartilage to the anterior vocal process of the arytenoid cartilage. This muscle may be separated into two portions, an internal part within the vocal cord and an external part between the vocal cord and the wing of the thyroid cartilage. For the most part, the fibres run parallel with the vocal cord. When they contract, they shorten the cord, make it thick, and round its edge. The external portion assists in bringing the vocal cords together, thus making glottal closure more tight.

The cricoarytenoid muscles

The cricoarytenoids are two muscle pairs: one lateral pair (to the side) and one posterior pair (backward). These two pairs of muscles have an antagonistic (opposing) action. The posterior cricoarytenoids are the muscles of inspiration that open the glottis. They arise from the posterior surface of the cricoid plate and are attached, in an upward, forward, and outward direction, to the lateral muscular process of the arytenoid cartilage. When these muscles contract, they rotate the arytenoid outward, thus opening the glottis. The lateral cricoarytenoids belong among the muscles of expiration, the adductor group. They arise from the lateral ring of the cricoid cartilage and insert into the muscular process of the arytenoid in an upward and backward direction. Contraction of the lateral cricoarytenoids rotates the arytenoid cartilages inward so that the vocal folds are brought together.

The two sides of the interarytenoid muscle are blended into one single mass, which extends from the muscular process of one arytenoid to that of the other. The action of this muscle is to pull together the posterior aspect of the arytenoid cartilages, thus closing the posterior portion of the cartilaginous glottis between the vocal processes of the arytenoids.

A fold from the top of the arytenoid to the lateral margin of the epiglottis on each side is supported by a bilateral band of muscle, the aryepiglotticus muscle. This semicircular structure aids in narrowing the laryngeal vestibule by pulling the arytenoids together and the epiglottis down. This is another example of the sphincter action ("valve" function) of all adducting laryngeal muscles that bring the vocal cords together. This sphincter action, by tightening of its closure, is the basis for all laryngeal protection. When this primitive sphincter mechanism intrudes into the refined coordination of phonation, it constricts the voice and causes the throaty quality of retracted resonance. This primitive, protective mechanism is at the root of many functional voice disorders. Moreover, the constricting sphincter action by many muscles is very strong because it is opposed by only one muscle, the abducting posterior cricoarytenoid.

Vocal cords. The two true vocal cords (or folds) represent the chief mechanism of the larynx in its function as a valve for opening the airway for breathing and to close it during swallowing (see Figure 2). The vocal cords are

From Lederer and Hollender, Basic Otolaryngology (1956); F. A. Davis Company

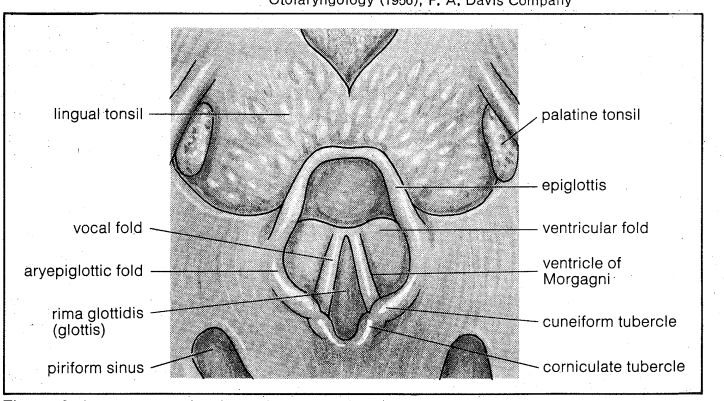


Figure 2: Laryngoscopic view of glottis and vocal cords (folds).

supported by the thyroarytenoid ligaments, which extend from the vocal process of the arytenoid cartilages forward to the inside angle of the thyroid wings. This anterior insertion occurs on two closely adjacent points, the anterior commissure. The thyroarytenoid ligament is composed of elastic fibres that support the medial or free margin of the vocal cords.

The inner cavity of the larynx is covered by a continuous mucous membrane, which closely follows the outlines of all structures. Immediately above and slightly

lateral to the vocal cords, the membrane expands into lateral excavations, one ventricle of Morgagni on each side. This recess opens anteriorly into a still smaller cavity, the laryngeal saccule or appendix. As the mucous membrane emerges again from the upper surface of each ventricle, it creates a second fold on each side—the ventricular fold, or false cord. These two ventricular folds are parallel to the vocal cords but slightly lateral to them so that the vocal cords remain uncovered when inspected with a mirror. The false cords close tightly during each sphincter action for swallowing; when this primitive mechanism is used for phonation, it causes the severe hoarseness of false-cord voice (ventricular dysphonia).

False vocal cords

The mucous membrane ascends on each side from the margins of the ventricular folds of the upper border of the laryngeal vestibule, forming the aryepiglottic folds. These folds extend from the apex of the arytenoids to the lateral margin of the epiglottis. Laterally from this ring enclosing the laryngeal vestibule, the mucous membrane descends downward to cover the upper-outer aspects of the larynx where the mucous membrane blends with the mucous lining of the piriform sinus of each side. These pear-shaped recesses mark the beginning of the entrance of the pharyngeal foodway into the esophagus.

The mucous membrane of the larynx consists of respiratory epithelium made up of ciliated columnar cells. Ciliated cells are so named because they bear hairlike projections that continuously undulate upward toward the oral cavity, moving mucus and polluting substances out of the airways. The true vocal cords, however, are exceptional in that they are covered by stratified squamous epithelium (squamous cells are flat or scalelike) as found in the alimentary tract. The arrangement is functional, since the vocal cords have to bear considerable mechanical strain during their rapid vibration for phonation, which occurs during many hours of the day. The transition from the respiratory to the stratified epithelium above and below the vocal cords is marked by superior and inferior arcuate (arched) lines. Unfortunately, such transitional epithelium also has the drawback of being easily disturbed by chronic irritation, which is one reason why the large majority of laryngeal cancers begin on the vocal cords. The mucous membrane of the larynx contains numerous mucous glands in all areas covered by respiratory epithelium, excepting again the vocal cords. These glands are especially numerous over the epiglottis and in the ventricles of Morgagni. The mucus secreted by these glands serves as a lubricant for the mucous membrane and prevents its drying in the constant airstream.

The vocal cords also mark the division of the larynx into an upper and lower compartment. These divisions reflect the development of the larynx from several embryonal components called branchial arches. The supraglottic portion differs from the one beneath the vocal cords in that the upper portion is innervated sensorially by the superior laryngeal nerve and the lower (infraglottic) portion by the recurrent (or inferior) laryngeal nerve. The lymphatics (*i.e.*, the vessels for the lymph flow) from the upper portion drain in an upward lateral direction, while the lower lymphatics drain in a lateral downward direction.

The space between the vocal cords is called rima glottidis, glottal chink, or simply glottis (Greek for tongue). When the vocal cords are separated (abducted) for respiration, the glottis assumes a triangular shape with the apex at the anterior commissure. During phonation, the vocal cords are brought together (adducted or approximated), so that they lie more or less parallel to each other. The glottis is the origin of voice, although not in the form of a "fluttering tongue" as the Greeks believed.

Dimensions of the vocal cords

The vocal cords vary greatly in dimension, the variance depending on the size of the entire larynx, which in turn depends on age, sex, body size, and body type. Before puberty, the larynx of boys and girls is about equally small. During puberty, the male larynx grows considerably under the influence of the male hormones so that eventually it is approximately one-third larger than the female larynx. The larynx and the vocal cords thus reflect body size. In tall, heavy males the vocal cords may be as long as 25 millimetres (one inch), representing the low-pitched instrument of a bass voice. A high-pitched tenor voice is produced by vocal cords of the same length as in a low-voiced female contralto. The highest female voices are produced by the shortest vocal cords (14 millimetres), which are not much longer than the infantile vocal cords before puberty (10–12 millimetres). The larynx is, among other things, a musical instrument that follows the physical laws of acoustics fairly closely.

LARYNGEAL FUNCTION

The physiology of the larynx has been investigated for many years. The first laryngeal mirror was invented in 1829, but its value was not recognized. Almost 30 years later, a Spanish singing teacher used a dental mirror to observe his own larynx during his study of artistic singing. Soon after, the art of laryngoscopy was introduced into clinical practice; the medical specialty of laryngology was born. During the remainder of the 19th century and early in the 20th century, many important discoveries were made by numerous German, French, and Italian laryngologists, physiologists, and phoneticians.

Theory of voice production. The physical production of voice has been explained for a long time by the myoelastic or aerodynamic theory, as follows: when the vocal cords are brought into the closed position of phonation by the adducting muscles, a coordinated expiratory effort sets in. Air in the lungs, compressed by the expiratory effort, is driven upward through the trachea against the undersurface of the vocal cords. As soon as the subglottic pressure has risen sufficiently to overcome the closing effort of the vocal cords, the glottis is burst open, a puff of air escapes, the subglottic pressure is reduced, and the elasticity of the glottis together with the effect of the moving air causes the adducted cords to snap shut. The subglottic pressure rises again and the entire cycle is repeated. These cycles of exploding air puffs occur as frequently as the physical interaction of the subglottic pressure with the glottic resistance permits. The latter is determined by the tension of the vocal cords and their closing force. The number of these cycles per second is small for tones of low pitch and much greater for high tones, as will be explained later. The resulting laryngeal fundamental tone thus varies greatly in audible pitch.

According to the myoelastic theory, the production of laryngeal voice is a mechanical phenomenon directed by aerodynamic principles and muscular coordination. The vocal cords vibrate purely passively in the blowing airstream and are merely maintained in their position of phonation by the adducting muscles as these are activated by the laryngeal nerves. This vibration is not an active phenomenon like the whirring of the wings of a flying insect. Evidence for the myoelastic theory can be demonstrated in various ways. High-speed motion pictures of the vocal cords have been made, photographing their vibration at the rate of 4,000 or more frames per second. When such a picture is then projected at regular film speeds of 16 or 24 frames per second, the available film length is greatly extended in duration so that each of the hundreds of vocal-cord vibrations per second can be seen in ultraslow motion. A tone of 250 cycles per second (cps or Hz), for example, filmed at 4,000 frames and played back at 16 frames per second will permit each of the 250 vibrations to be seen for one second. Other evidence supporting the myoelastic theory is found in observations such as the fact that a nearly normal voice can be produced despite bilateral (on both sides) vocal-cord paralysis.

Vocal registers. *The basic registers.* For many centuries the so-called vocal registers were well known to the classical masters of the bel canto style of singing, the basic registers being called chest voice, midvoice, and head voice. These terms are derived from observations, for example, that in the low-chest register the resonances are felt chiefly over the chest. When sitting on a wooden bench with a large male, one can feel the vibrations of his low voice being transmitted through the back of the bench. In the high head voice, the vibrations are felt chiefly over the skull. The practice of singing is based on

Table 1: Laryngeal Activity in Chest and Falsetto Register of the Male

	chest register	falsetto register
Position of larynx	low	elevated
Shape of cords	rounded, full	sharp-edged, thin
Length of cords	short	elongated
Tension of cords	low isotonic inner tension	great isometric outer tension
Ventricular folds	adducted	abducted
Ventricles	narrow	wide
Glottal closure	complete in each cycle	open oval slit
Length of glottis	full length of cords	short, anterior
Vibrations	broad, entire surface	narrow, marginal
Amplitude of vibration	wide	small
Frequency of vibration	low	high
Vocal pitch	low	high
Epiglottis	lowered	elevated
Supraglottic space	narrow	widened
Thyroarytenoid	very active	inactive
Cricothyroid	inactive below 150 Hz	very active
Harmonics	prominent higher harmonics	fundamental almost sinusoidal
Vocal timbre	full, resonant, loud	thin, weak, soft
Perceptual characteristic	masculine	juvenile, feminine
Function	speaking voice	special artistic effects

several artistic subdivisions in both sexes, depending on factors as discussed below. Other vocal phenomena may be heard below and above normal register limits, such as extra low tones, the "vocal fry."

Transition between vocal registers

The natural transition between two adjacent registers may be compared to the gearshift of a car. The same absolute vehicle speed can be maintained by driving either with the engine turning fast while in low gear or with fewer engine revolutions in the next higher gear. The register mechanism of the human voice is quite similar in this respect. Where the registers overlap, a series of transitional tones may be sung with either of the adjacent registers. These tones of the same fundamental frequency, sound level, and basic sound category in different vocal registers have recently been defined as isoparametric tones. In the untrained male voice, the transition between the midvoice and the high falsetto sounds abrupt; this so-called register break is similar to the noisy gearshift in a run-down truck. One aim of vocal education is to teach smoothly equalized register transitions.

Loud phonation of any given tone shifts its register mechanism toward the next lower register; for example, a crescendo falsetto tone grows into loud head voice. Conversely, soft intonation raises the mechanism to the next higher type, as when a loud head tone fades into soft falsetto. This phenomenon is the physiologic basis of *messa di voce*, the technique of swelling tones. Thus, the characteristic mechanism of each register represents a continuum of intralaryngeal adjustments. In the male voice, the gradual and overlapping transitions of phonic function may be aligned as follows: low chest tones, loud–soft; transition; middle register, loud–soft; transition; loud head voice—soft artistic falsetto–thin natural falsetto. X-ray studies can show the difference between the loud male head voice and the soft male falsetto. The former employs the midvoice mechanism, the latter the falsetto mechanism. In the female voice, the two lower registers behave similarly, while head voice can be only loud or soft and may be followed by a fourth register, the flageolet or whistle register of the highest coloratura sopranos. The Italian term falsetto simply means false soprano, as in a castrato (castrated) singer. Hence, the normal female cannot have a falsetto voice.

Studies of register differences. Studies devoted to the problem of voice register may be divided into two groups: observations of the visible laryngeal mechanism and studies of the audible register differences.

Studies of the visible laryngeal mechanism for the production of different registers began with the laryngoscope. Modern laryngostroboscopes employ the oscillating light of a high-power fluorescent light source that is monitored by the laryngeal vibrations through a throat microphone. Such devices, when they flash on and off at just the right rate, make the vocal cord movements appear much slower than they actually are, so that the observer perceives a slow-motion pattern. High-speed cinematography (moviemaking) has elucidated many details of vocal cord function for the various registers. Radioscopic (X-ray) methods were introduced only a few years after the discovery of X-rays in 1895. Among these, lateral (from the side) radioscopy of the larynx reveals the mechanism of vocal cord tension; frontal X-ray films demonstrate the typical configuration of the vocal cords for each register. Mechanical recordings of the respiratory movements of the chest, originally with rubber belts and lately with electronic strain gauges, disclose the breathing patterns for the various registers. Breath support (*appoggio*) of singing instruction can be demonstrated through such recordings, as well as by radiography of the chest. Aerodynamic measurements of pressure, flow rate, and volume of the air exhaled during specific phonic tasks have produced additional details. Electromyography (study of muscle currents) involving the insertion of needle electrodes into certain laryngeal muscles permits the isolated recording of finely coordinated muscular effort during the singing in various registers.

Differences in sound-wave patterns

A second group of investigations concerns audible register differences as an acoustic phenomenon. Electroacoustic analysis demonstrates the specific sound-wave patterns (harmonic spectra) of each register. In general, the full chest voice is rich in higher harmonics, whereas the thin falsetto voice is composed chiefly of sound-wave energy distribution near the vocal fundamental (the relatively narrow band of wave frequencies that characterizes any particular voice). The subjective impressions of singers during the production of an ascending scale reflect the voluntary techniques of vocal breath control, such as with respiratory support (*appoggio*). Positioning of the larynx, suitable shaping of the pharyngo-oral resonator (vocal tract), proper placement of the tongue, and the specific tension of the soft palate belong among the learned techniques of register equalization. Definite vibrations may be felt in the thorax, in the area of the hard palate, or above the nose. These subjectively felt resonances depend on bone conduction of the laryngeal sound. Very little has as yet been done regarding the subjective evaluation of voice registers by listening judges. These perceptual factors are still little understood, but it appears that multiple acoustic perceptions operate in voice-register judgment.

It is clear that the vocal registers represent a continuum of laryngeal adjustments in response to different respiratory-mechanical requirements necessary for the production of the individual frequency range. The poles of these adjustments at the opposites of chest voice and male falsetto voice illustrate the chief differences, as shown in Table 1; the midvoice occupies an intermediate position.

Vocal attributes. *Vocal frequency.* The voice has various attributes; these are chiefly frequency, harmonic structure, and intensity. The immediate result of vocal cord vibration is the fundamental tone of the voice, which determines its pitch. In physical terms, the frequency of vibration as the foremost vocal attribute corresponds to the number of air puffs per second, counted as cycles per

second (cps or Hz). This frequency is determined by both stable and variable factors. The stable determinants of the individual voice range depend on the laryngeal dimensions as related to sex, age, and body type. The smaller a larynx, the higher its pitch range. Within this individually fixed range, variables that influence the pitch of a given phonation include: tension of the cord, force of glottal closure indicated by the glottal resistance, and expiratory air pressure. Growing tension of the cricothyroid muscle (as the external vocal cord tensor) increases the vocal pitch, and vice-versa. Increased glottal closure and expiratory effort add to this tensing effect under certain circumstances. For example, 100 vibrations per second produce a low chest tone of a low male voice, while 1,000 are close to the "high C" of a female soprano. An average vocal range normally encompasses two musical octaves (*e.g.*, 100 to 400 vibrations per second); trained singers may reach three or more octaves.

Voice types. Musical practice for centuries has recognized six basic voice types: bass, baritone, and tenor in the male, in contrast to contralto, mezzo-soprano, and soprano in the female. Sex, therefore, is one of the first determinants of voice type in the two categories. Body type and general physical constitution represent the second determinant of the individual voice type because the laryngeal dimensions vary in fairly strict conformity to whether the body type is large or husky or frail or small. A tall, athletic male usually has a large, spacious larynx. Repeated observations show that short, dainty females tend to have a small and delicately built larynx. The intermediate voice types of the male baritone and the female mezzo-soprano usually represent the corresponding intermediate body types. The art of singing recognizes additional subdivisions. The voice of a basso profundo is extremely low and heavy. The lyric tenor possesses a high, light, and flexible voice. Still higher and lighter is the counter tenor (as used in singing oratorios) who is the male counterpart of the highest female voice found in the extra high and light coloratura soprano. The dramatic voices employed in the Wagnerian operas represent intermediate forms between a male tenor (or high baritone) and a heroically masculine body type. The female dramatic soprano is usually heavily built; her strong mezzo-soprano voice can produce the high soprano tones.

The relationship between registers and voice types

The registers are related to voice types. As a general rule, the low voices possess a large range of chest voice with a much smaller range of head voice. The reverse holds for the high voice types, while baritone and mezzo-soprano assume an intermediate position. In the normal individual and the well-trained singer in particular, the midvoice encompasses one musical octave. As a further rule of thumb, the traditional and optimal transition tones follow a fairly stable and general pattern. As summarized in Figure 3, the three female voice types usually show the first transition from chest to midvoice at the tones d^1, e^1, and f^1, above middle c^1, respectively. The second transition between midvoice and head register in the three female voice types is almost precisely one octave higher. An extra-low contralto voice may prefer to shift the two transitions at slightly lower frequencies, whereas a very high coloratura soprano may prefer the two shifts a semitone (halftone) higher. The two transition tones of the three male voice types are situated almost precisely one octave lower than the respective six female transition tones. It should not be overlooked that the specific features in male voices sound approximately one octave lower than in the female voices of corresponding type. This octave phenomenon stems from the larger dimension of the adult male larynx. (The musical custom of writing the tenor part on the soprano stave in contrast to the correct notation of bass and baritone in the bass clef is a misleading tradition that derives from an old custom of four-part writing, for the tenor always sounds one octave lower than the soprano.)

Vocal ranges. As shown in Figure 3, the individual ranges of the singing voice extend from about 80 cycles per second in the low bass to about 1,050 cycles per second in the "high C" of the soprano (all values are approximated). The lowest note of serious musical literature is a low D with 73 cycles per second at the end of Osmin's aria in Mozart's *Abduction from the Seraglio*, and the highest is a high f^3 with almost 1,400 cycles per second sung by the Queen of the Night in Mozart's *Magic Flute*. Exceptionally high soprano tones are no longer sung with vocal cord vibration but are produced in the flageolet (or whistle) register simply by whistling through the narrow eliptical slit between the overtensed and motionless vocal cords. When citing the exceptional vocalistic feats of singers from the classical bel canto era, it should not be overlooked that musical pitch has been rising markedly since those days. Concert pitch is presently standardized at 440 cycles per second for the international tuning tone a^1. In the last half of the 18th century, the reference tone was at least one semitone lower.

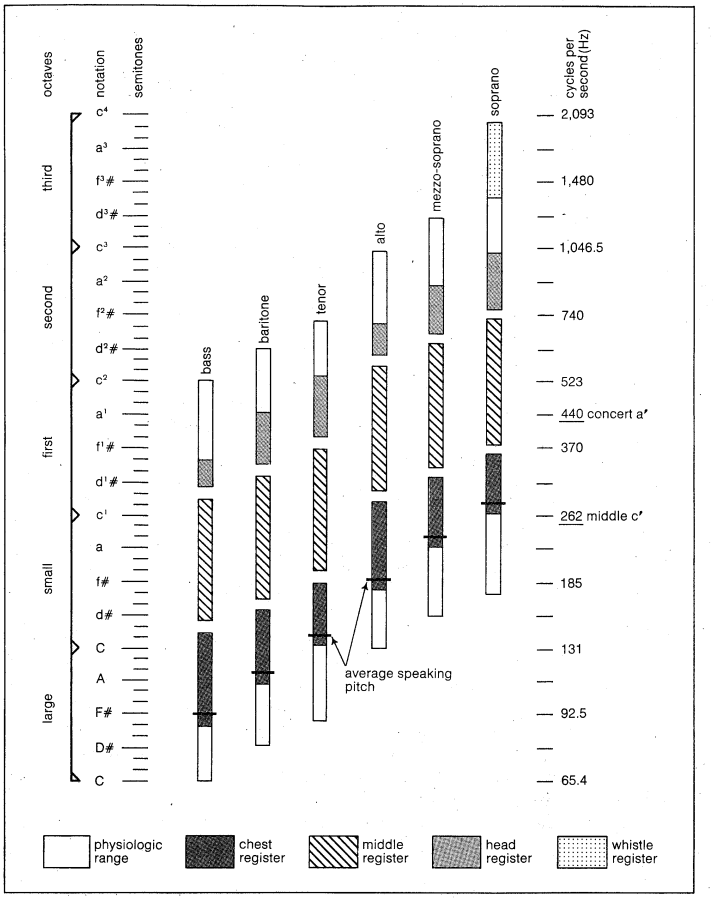


Figure 3: Relationships of registers, vocal range, average speaking pitch, and extreme tones of human voice. Octaves and notation according to von Helmholtz; frequencies based on standard concert a' = 440 Hz.

From *Archives of Otolaryngology*, vol. 62, p. 597; copyright 1955, American Medical Association

Harmonic structure. A second attribute of vocal sound, harmonic structure, depends on the wave form produced by the vibrating vocal cords. Like any musical instrument, the human voice is not a pure tone (as produced by a tuning fork); rather, it is composed of a fundamental tone (or frequency of vibration) and a series of higher frequencies called upper harmonics, usually corresponding to a simple mathematical ratio of harmonics, which is 1:2:3:4:5, etc. Thus, if a vocal fundamental has a frequency of 100 cycles per second, the second harmonic will be at 200, the third at 300, and so on. As long as the harmonics are precise multiples of the fundamental, the voice will sound clear and pleasant. If nonharmonic components are added (giving an irregular ratio), increasing degrees of roughness, harshness, or hoarseness will be perceived in relation to the intensity of the noise components in the frequency spectrum.

The primary laryngeal tone composed of its fundamental and harmonics is radiated into the supraglottic vocal tract (above the glottis). The cavities formed by the

Modulation of vocal tone

pharynx, nasopharynx, nose, and oral cavity represent resonators. Since they are variable in size and shape through the movements of the pharyngeal musculature, the palatal valve, and the tongue in particular, the individual sizes of the supraglottic resonating chambers can be varied in countless degrees. The shaping of the vocal tract thus determines the modulation of the voice through resonance and damping. As a general rule, a long and wide vocal tract enhances the lower harmonics, producing a full, dark, and resonant voice. Conversely, shortening and narrowing of the vocal tract leads to higher resonances with lightening of the voice and the perceptual attributes ranging from shrill and strident to constricted and guttural.

Vocal styles. These types of vocal resonance may be illustrated with a continual series of vocal practices that have been studied through physiologic and electroacoustic analysis. This perceptual series begins with the full, loud, and sonorous sound during the natural vocalizations for laughing, yawning, and yodelling. The rich higher harmonics responsible for the perceptual qualities of these vocalizations are produced by a maximally lowered larynx and greatly widened resonator. At the next step is the sonorous and full sound of so-called covered singing in the German opera style. Rich in higher harmonics (or overtones), this vocal style is performed with lowered larynx, elevated epiglottis, and widened throat cavity. A large group of open or uncovered singing styles lying in the centre of the series extends from the extremely uncovered, flat, and "white" openness of, for example, Spanish flamenco singing, over the flat style of popular singing, to the brightness of Italian bel canto. Approaching the other pole of the series, the large group of functional voice disorders results from constricted resonance of the vocal tract. It is typical of these hyperkinetic (overactive) vocal disorders that the voice is produced with marked laryngeal elevation, constriction of the laryngeal vestibule, and often with pronounced elementary sphincter action of the larynx. The extreme end of this functional series is characterized by the use of the larynx as a primitive sphincter organ as employed in ventriloquism. The maximally elevated and constricted larynx within a very narrow throat cavity produces the high-pitched, thin, muffled, and weak quality of ventriloquism, which is characterized by great reduction of the higher harmonics.

Use of the larynx in ventriloquism

Individual voice quality. Apart from the variable influences of the vocal tract on the momentary vocal resonance according to training and intention, the supraglottic resonator exerts a constant influence on the vocal quality by shaping its individual characteristics. Just as human faces differ in almost endless variations, the configuration of the supraglottic structures is also highly characteristic, having, in fact, been called the "inner face." The anatomical shape and the physiologic flexibility of the vocal tract serve to mold the individual vocal personality in at least two ways: by its inborn shape and by the learned behaviour of using it for communication. Any individual's mother tongue shapes his articulatory behaviour into certain patterns, which remain audible in all languages that he learns after puberty and constitute one aspect of the so-called foreign accent. It often is easy to recognize a speaker over the telephone after having listened to his voice a few times without necessarily having met him in person. The ability to recognize a given speaker solely by the quality and inflection of his voice is the basis of efforts to produce "voice prints" that should be as unmistakably identifying as fingerprints are.

Intensity. Vocal intensity, the third major vocal attribute, depends primarily on the amplitude of vocal cord vibrations and thus on the pressure of the subglottic airstream. The greater the expiratory effort, the greater the vocal volume. Another component of vocal intensity is the radiating efficiency of the sound generator and its superimposed resonator. The larynx has been compared to the physical shape of a horn. This construction is most efficient in acoustical practice, as seen in the shape of wind instruments, car horns, sirens, loudspeakers, etc. A well-shaped, wide, and flexible vocal tract enhances the projective potential of the voice. Conversely, a morphologically narrow, pathologically constricted, or emotionally tightened throat produces a muffled, constricted sound with poor carrying power.

Laughing and yawning

The inborn automatic reflexes of laughing and yawning illustrate the resonator action of the vocal organ. Together with a widely opened mouth, flat tongue, elevated palate, and maximally widened pharynx, the larynx assumes a lowered position with maximally elevated epiglottis. This configuration is ideal for the unimpeded radiation of the vocal cord vibrations so that the resulting sound is loud and bright, with a gaily ringing quality; it is the sound of happy laughter. The opposite is present with the painfully tight-throated, choked sobbing of someone crying in despair.

PRODUCTION OF SPEECH

As shown by physiologic, culturally learned, and pathologic examples, the vocal functions represent highly coordinated actions of all organs participating in phonation so that it is difficult to view each factor, attribute, or quality in isolation. This restriction is even more apparent in the study of speaking.

Singing and speaking. A major difference between singing and speaking is psychological in nature. Singing as a physiological performance is exhibited by the majority of human beings who have what seems to be an inborn musical sense that depends on appropriate development of their highest cortical (brain) centres for audition. Although the art of singing in a particular artistic style typically demands formal study, the untrained use of the voice for self-expression through singing develops spontaneously in late childhood and during the period following vocal maturation. Singing involves the use of inherited neural mechanisms that are regulated in part by deeper, subcortical (below the cortex) brain centres, particularly those related to emotional activity. Singing serves many as a way of emotional relief and is related to the social activities of human play. Although song among humans is not as intimately related to sexual propagation as it is in certain animals (*e.g.*, birds), people are still influenced by such sensual stimuli as love songs and madrigals, as well as ceremonial and religious performances.

The practice of spontaneous singing and of artistic song satisfies emotional needs, but it may not always communicate in a clear ideational sense. When a brain stroke causes aphasia (loss of language for communication), for example, the singing voice often remains normal or at least better preserved, so that some aphasics who cannot say a word can sing with good articulation. This observation has been used to explain that disorders causing aphasia may damage other brain areas than those used for singing. Another example is the severe stutterer who can sing or whisper with fluency. The same dichotomy of communicative speech and declamatory singing is often seen in cases of spastic dysphonia (a peculiar, grave voice disorder without demonstrable brain damage that causes a painfully choked and halting manner of speaking, while singing usually remains undisturbed).

Perceptual differences between speaking and singing

In the perceptual category, the principal differences between speaking and singing concern the rhythmic patterns. Speaking uses gliding vocal inflections with rapid pitch variations as well as frequent and abrupt intensity modifications for syllabic accentuation. The rhythmical pattern of stresses, unstressed syllables, and breathing pauses is dictated by the meaning of the sentence. The so-called prosodic features of speech (*i.e.*, its melodic inflections) follow the general, regional, and dialectal rules of a given language. In this sense, the essence of speaking is its continual flexibility, variability, and adaptability.

Singing differs from speaking in the following respects. The melody is followed in precise and discrete steps over customary musical intervals (*e.g.*, from the note f to the note "cee" on the piano), which commonly are not smaller than semitones in Western music, though quarter and eighth tones are frequently used in Oriental or African music. The vowels are prolonged because they carry the melody. The rhythm of the fixed tonal steps follows the

pattern prescribed by the composer and long notes may be sustained for special effects.

Exceptions to these general rules are found in the portamento, a gliding change between two pitch levels, of Western song, used sparingly as an embellishment. Parlando singing is a speaking type of song, used in the recitativo of Italian opera style. In these intentionally communicative preludes to formal arias—because they tell most of the story—the rhythm of the spoken word is incorporated into the melody, which, in turn, to a certain degree, follows the prosodic vocal inflection.

The melodic inflection of speech communicates considerable meaning in certain languages, such as in Africa and China. This problem of linguistic tonality or word melody requires the appropriate selection of various rising, sustained, or falling intervals to express the full meaning of a word. Chinese words are monosyllabic, and their multiple meanings cannot be understood without the appropriate prosodic inflection by the "tones" of the particular dialect. If Chinese is spoken without vocal inflection, such as when whispering, intelligibility is reduced by at least one-third.

Articulation and speech sounds. Voice alone has little communicative function, except as elementary exclamation. The chief function of voice in speech is as a carrier wave that is modulated into speech sounds by such articulating organs as the lips and tongue.

Phonetic classification. For many centuries, speech scientists have classified speech (phonetic) sounds into two categories: vowels and consonants. Vowels represent resonances of the suitably shaped vocal tract, while consonants are produced by total or partial occlusions within the various zones of articulation. Almost equally old is a phonetic classification of voiced sounds with simultaneous laryngeal tone and unvoiced sounds without laryngeal participation. Vowels are voiced in almost all languages, while the consonants are likely to occur in pairs, one voiced and one unvoiced (B–P, V–F, Z–S, etc.). In acoustic terms, the voiced consonants represent sound-noise mixtures, the unvoiced partners being complex noises. Additional distinguishing features of the consonants are their mechanisms of articulation. Explosion of a preceding occlusion (an obstruction of the breath by the tongue, for example) produces the plosives or stop consonants such as the sound for P. A friction noise within a narrow channel gives rise to the fricative sounds such as that for F.

Articulation zones

Moreover, the various types of speech sounds are produced in different articulation zones marking the place of articulation that is typical for each language. The labial zone (between both lips) and the labiodental zone (between lower lip and upper teeth) are relatively stable in most languages. The linguodental zone (between tongue tip and teeth) varies greatly in different languages. In the Latin and Slavic languages, dental articulation with the tongue is often low against the teeth without aspiration. When this low linguodental articulation is used in English, it is called dentalization, a dialectal error because English typically requires higher linguoalveolar articulation (touching the tongue to the front part of the palate) with marked aspiration. North German assumes an intermediate position. The palatal articulation zone is subdivided into several segments along the midline between tongue and palate. Pharyngeal articulation does not ordinarily exist in English but is normal in such languages as German (as in the word *ach*) and Arabic. Laryngeal articulation represents two phonemes (elements of speech sounds): the glottal stop produced by bursting firm glottal closure for the hard attack of vowels, such as in North German, and the production of a faint hissing noise between the incompletely closed and motionless vocal cords.

Finally, speech sounds are also grouped as oral sounds without nasal resonance and as nasal sounds with various degrees of nasalization. Oral sounds are produced by fairly tight occlusion through contraction and elevation of the soft palate against the constricted posterior pharyngeal wall. The nasal sounds are produced with relaxed and lowered velum (soft palate) so that part of the sounding air column can enter the nasopharynx and nose, thus eliciting the resonance of these spaces. Three nasal sounds ordinarily occur in English (those for M, N, and Ng). French and certain other Latin languages make much freer use of nasalized vowels, such as in the final syllable "tion."

Phonetic science still faces the problem of providing a clear definition of the differences between nasal resonance, nasalization, and nasality. In essence, these are differences of degree and linguistic context. It may be said that nasal resonance produces nasal sounds, nasalization is the addition of nasal resonance to vowels, while nasality is a pathological phenomenon meaning the admixture of nasal resonances and noises during the articulation of those speech sounds that should not be nasal. (More will be said later in connection with cleft palate.)

Phonemes and phonetic transcription

Phonetic science describes the numerous speech sounds of a given language as its phonemes, which correspond to the smaller number of written letters or graphemes of conventional printed or written language only in a loose fashion, especially in English. (Close transcription is feasible with a system called the International Phonetic Alphabet.) The next larger unit of speech is the syllable, which contains at least one vowel or one vowel and an adjacent consonant. The vowel is that part of a word with high acoustic energy (loudness), while the consonants vary greatly in intensity, depending on articulation. The description of words (morphemes) and sentences (grammemes) is a task of linguistics. These linguistic units often overlap. "I" is a phoneme, a diphthong (combination of two vowels), a syllable, a word, and can even be a sentence, such as in: Who is it?—I. Words consist of one or more syllables, and sentences consist of one or more words.

Acoustic analysis. Perceptual recognition of speech sounds depends on the fact that the vocal tract is shaped during their articulation in such a manner that the resulting resonances produce characteristic bands (of sound-wave frequency) within the total energy spectrum, the so-called formants. These frequency bands are demonstrated by electroacoustic sound analysis, with such devices as the sound spectrograph. A suitable arrangement of electric sound-wave filters with variable band widths permits the graphic display of speech sounds in a three-dimensional system. Time is depicted on the horizontal axis, frequency (often from 100 to 8,000 cycles per second) on the vertical axis, and intensity by graded blackening of the heat-sensitive recording paper, increasing darkness indicating greater intensity. During phonetic analysis, the recorded formants are labelled as F1, F2, F3, etc. Their position on the frequency scale reflects the perceptual impressions they give to the listener. On this formant scale, the vowel u contains the lowest formant, followed by o, a, and e, while i has the highest formant with the perceptual correlate of thin reediness. Vowels represent complex sounds with relatively pure harmonic structure. Nonharmonic noise admixtures detract from their clarity, giving the perceptual impression of hoarseness.

The physical correlates of nasality are difficult to explain. Modern research has shown that the essence of nasality is represented by several features: weakened lower formants through damping, the occurrence of abnormal resonances and antiresonances, and the appearance of a new formant in higher frequency regions.

Central regulation. The question of what the brain does to make the mouth speak or the hand write is still incompletely understood despite a rapidly growing number of studies by specialists in many sciences, including neurology, psychology, psycholinguistics, neurophysiology, aphasiology, speech pathology, cybernetics, and others. A basic understanding, however, has emerged from such study. In evolution, one of the oldest structures in the brain is the so-called limbic system, which evolved as part of the olfactory (smell) sense. It traverses both hemispheres in a front to back direction, connecting many vitally important brain centres as if it were a basic mainline for the distribution of energy and information. The limbic system involves the so-called reticular activating system (structures in the brain stem), which repre-

The limbic system

sents the chief brain mechanism of arousal, such as from sleep or from rest to activity. In man, all activities of thinking and moving (as expressed by speaking or writing) require the guidance of the brain cortex.

In contrast to animals, man possesses several language centres in the dominant brain hemisphere (on the left side in a clearly right-handed person). It was previously believed that left-handers had their dominant hemisphere on the right side, but recent findings tend to show that many left-handed persons have the language centres more equally developed in both hemispheres or that the left side of the brain is indeed dominant. The foot of the third frontal convolution of the brain cortex, called Broca's area, is involved with motor elaboration of all movements for expressive language. Its destruction through disease or injury causes expressive aphasia, the inability to speak or write (see Figure 4). The posterior third of

From N. Geschwind, *Science*, vol. 170, no. 3961, p. 941 (1970); Copyright 1970 by the American Association for the Advancement of Science

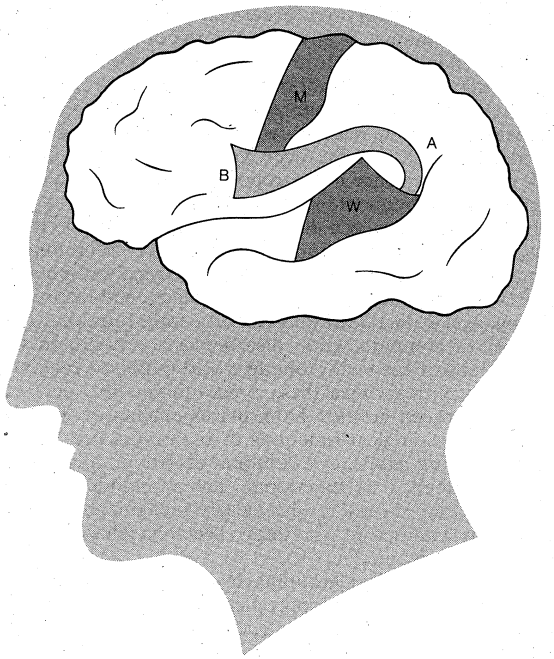


Figure 4: Lateral surface of left hemisphere of brain. B, Broca's area; M, motor cortex; W, Wernicke's area; A, arcuate fasciculus.

the upper temporal convolution represents Wernicke's area of receptive speech comprehension. Damage to this area produces receptive aphasia, the inability to understand what is spoken or written as if the patient had never known that language.

Broca's area surrounds and serves to regulate the function of other brain parts that initiate the complex patterns of bodily movement (somatomotor function) necessary for the performance of a given motor act. Swallowing is an inborn reflex (present at birth) in the somatomotor area for mouth, throat, and larynx. From these cells in the motor cortex of the brain emerge fibres that connect eventually with the cranial and spinal nerves that control the muscles of oral speech.

In the opposite direction, fibres from the inner ear have a first relay station in the so-called acoustic nuclei of the brain stem. From here the impulses from the ear ascend, via various regulating relay stations for the acoustic reflexes and directional hearing, to the cortical projection of the auditory fibres on the upper surface of the superior temporal convolution (on each side of the brain cortex). This is the cortical hearing centre where the effects of sound stimuli seem to become conscious and understandable. Surrounding this audito-sensory area of initial crude recognition, the inner and outer auditopsychic regions spread over the remainder of the temporal lobe of the brain, where sound signals of all kinds appear to be remembered, comprehended, and fully appreciated. Wernicke's area (the posterior part of the outer auditopsychic region) appears to be uniquely important for the comprehension of speech sounds.

The integrity of these language areas in the cortex seems insufficient for the smooth production and reception of language. The cortical centres are interconnected with various subcortical areas (deeper within the brain) such as those for emotional integration in the thalamus and for the coordination of movements in the cerebellum (hindbrain).

Feedback mechanisms. All creatures regulate their performance instantaneously comparing it with what it was intended to be through so-called feedback mechanisms involving the nervous system. Auditory feedback through the ear, for example, informs the speaker about the pitch, volume, and inflection of his voice, the accuracy of articulation, the selection of the appropriate words, and other audible features of his utterance. Another feedback system through the proprioceptive sense (represented by sensory structures within muscles, tendons, joints, and other moving parts) provides continual information on the position of these parts. Limitations of these systems curtail the quality of speech as observed in pathologic examples (deafness, paralysis, underdevelopment).

Substitutes for the larynx. A growing number of middle-aged or older patients have had their larynx removed (laryngectomy) because of cancer. Laryngectomy requires the suturing of the remaining trachea into a hole above the sternum (breastbone), creating a permanent tracheal stoma (or aperture) through which the air enters and leaves the lungs. The oral cavity is reconnected directly to the esophagus. Having lost his pulmonary activator (air from the lungs) and laryngeal sound generator, such an alaryngeal patient is without a voice (aphonic) and becomes effectively speechless; the faint smacking noises made by the remaining oral structures for articulation are practically unintelligible. This type of pseudo-whispering through buccal (mouth) speech is discouraged to help the patient later relearn useful speech on his own. A frequently successful method of rehabilitation for such alaryngeal aphonia is the development of what is called esophageal or belching voice.

Alaryngeal aphonia

Esophageal voice. Some European birds and other animals can produce a voice in which air is actively aspirated into the esophagus and then eructated (belched), as many people can do without practice. The sound generator is formed by the upper esophageal sphincter (the cricopharyngeus muscle in man). As a replacement for vocal cord function, the substitute esophageal voice is very low in pitch, usually about 60 cycles per second in humans. Training usually elevates this grunting pitch to about 80 or 100 cycles.

Esophageal voice in man has been reported in the literature since at least 1841 when such a case was presented before the Academy of Sciences in Paris. After the perfection of the laryngectomy procedure at the end of the 19th century, systematic instruction in esophageal (belching) phonation was elaborated, and the principles of this vicarious phonation were explored. Laryngectomized persons in many countries often congregate socially in "Lost Cord Clubs" and exchange solutions of problems stemming from the alaryngeal condition.

Artificial larynx. Approximately one-third of all laryngectomized persons are unable to learn esophageal phonation for various reasons, such as age, general health, hearing loss, illiteracy, linguistic barriers, rural residence, or other social reasons. These persons, however, can use an artificial larynx to substitute for the vocal carrier wave of articulation. Numerous mechanical and pneumatic models have been invented, but the modern electric larynx is most serviceable. It consists of a plastic case about the size of a flashlight, containing ordinary batteries, a buzzing sound source, and a vibrating head that is held against the throat to let the sound enter the pharynx through the skin. Ordinary articulation thus becomes easily audible and intelligible. Other models lead the sound waves through a tube into the mouth or are encased in a special upper dental plate. More recent

efforts aim at surgically inserting an electric sound source directly into the neck tissues to produce a more natural sound resembling that of normal speech.

Synthetic production of speech sounds. The essence of speech and its artificial re-creation has fascinated scientists for several centuries. Although some of the earlier speaking machines represented simple circus tricks or plain fraud, an Austrian amateur phonetician, in 1791, published a book describing a pneumomechanical device for the production of artificial speech sounds.

Speech synthesizers. Modern electronic speech synthesizers have been constructed in various phonetic laboratories for the past several decades. Some of these are named the "Coder," "Voder," and "Vocoder," abbreviations for longer names (*e.g.*, "Voder" standing for Voice Operation Demonstrator). In essence, they are electrical analogues of the human vocal tract. Appropriately arranged electric circuits produce a voicelike tone, a modulator of the harmonic components of this fundamental tone, and a hissing-noise generator to produce the sibilant and other unvoiced consonant sounds. Resonating circuits furnish the energy concentrations within certain frequency areas to simulate the characteristic formants of each speech sound. The resulting speechlike sounds are highly controllable and amazingly natural as long as they are produced as continuants. For example, it is possible to imitate the various subtypes of the hard U.S. sound for R (as in "car") by moving a few levers or knobs. Difficulties become greater when many other attributes of fluent speech are to be imitated, such as coarticulation of adjacent sounds, fluctuating nasalization, and other segment features and transients of connected articulation. Speech synthesizers have, nevertheless, made a contribution to the study of the various physical characteristics that contribute to the perception and recognition of speech sounds.

Speech recognizers

The counterpart of speech synthesizers is the speech recognizer, a device that receives speech signals through a microphone or phono-optical device, analyzes the acoustic components, and transforms the signals into graphic symbols by typing them on paper. Modern models may incorporate computers to store some of the information that permits the device directly to type from dictation. Early models had great difficulties with the correct spelling of homophonous words (those that sound alike but differ in spelling and meaning) such as "to, too, two" or "threw, through, thru." Human transcribers usually have no difficulty with these distinctions because they listen to major parts of the sentence to recognize each word from the context and general situation. The computerized machines developed in the 1970s, however, had to be programmed for each detailed aspect of speech recognition that people normally learn through many years of general schooling and specialized training.

Disorders of speech

THE STUDY OF SPEECH PATHOLOGY

Prevalence of speech disorders. A comprehensive evaluation of knowledge and research in speech and hearing, published in 1969 by a section of the U.S. Public Health Service, under the title *Human Communication and Its Disorders—An Overview*, reveals that approximately 9,000,000 Americans are reported to have hearing impairments of handicapping magnitude; another 2,000,000 are said to suffer from central processing disorders from neurologic disease; and 10,000,000 to have speech disorders. The cost of coping with these problems is about $500,000,000 per year, while the annual deficit in earning power among such impaired persons approximates nearly $2,000,000,000. The same report estimated the prevalence of various speech disorders in the United States as shown in Table 2. The disorders of articulation among young children continue to be numerically most frequent.

No data in similar detail are available from other countries; however, from experience in Germany, Austria, and other Central European countries, it may be said that the incidence and prevalence of speech disorders follow similar patterns among other Western nations. Certain deviations from these trends are equally typical. The incidence of cleft palate is very high among American Indians, such as in Montana, while it is much lower among Negroes than in Caucasians, such as in Jamaica. Stuttering is more frequent in northern countries, such as in Scandinavia, than it is in southern regions—*e.g.*, Italy. Similar trends have been noted in the Far East. Nonorganic, functional, or emotional disorders of voice and speech reflect psychological tensions exerted by the environment. For such reasons, vocal nodules (growths on the vocal cords), contact ulcer, vocal fatigue, etc., occur with greater incidence in metropolitan areas and are conspicuously less frequent in rural regions.

Table 2: Estimated Prevalence of Speech Defects in the United States

type of speech problem	ages 5–21 years		all ages	
	percent	number	percent	number
Functional articulatory	3.0	1,500,000	3.0	6,000,000
Stuttering	.7	350,000	.7	1,400,000
Voice	.2	100,000	.2	400,000
Cleft palate speech	.1	50,000	.1	200,000
Cerebral palsy speech	.2	100,000	.2	400,000
Retarded speech development	.3	150,000	.3	600,000
Impaired hearing (with speech defect)	.5	250,000	.5	1,000,000
Total	5.0	2,500,000	5.0	10,000,000

Classification of speech disorders. In accordance with physiological considerations, disorders of communication are first classified into disorders of voice and phonic respiration, disorders of articulated speech, and disorders of language. It has been known for a long time that the majority of communication disorders are not caused by local lesions of the teeth, tongue, vocal cords, or regulating brain centres. Since these predominant disorders of voice and speech develop from derangements of the underlying physiological functions of breathing, use of the voice, speaking habits, or emotional disorders, this group has been labelled as functional. The remainder of the communication disorders with clearly recognizable structural abnormalities in the total speech mechanism has been termed organic.

Interaction of functional and organic disorders

While this empirical grouping has certain implications for the selection of the appropriate treatment, it is not satisfactory because organic structure and living function can never be separated. Certain functional disorders of the voice caused by its habitual abuse may very well lead to secondary structural changes, such as the growths (polyps and nodules) of the vocal cords, which develop as a result of vocal abuse. On the other hand, all of the obviously organic and structural lesions, such as loss of the tongue from accident or surgery, almost inevitably will be followed by emotional and other psychological reactions. In this case, the functional components are of secondary nature but to a great extent will influence the total picture of disturbance, including the patient's ability to adjust to his limitation, to relearn a new mode of appropriate function, and to make the best of his condition.

Within these major groups, the various types of communication disorders have for a long time, and in most parts of the world, been described by the listener's perceptual impression. Most languages employ specific words for the various types of abnormal speech, such as stuttering, stammering, cluttering, mumbling, lisping, whispering, and many others. The problem with such subjective and symptomatic labels is the fact that they try to define the final, audible result, the recognizable phenomenon, and not by any means the underlying basis. This general human tendency to describe disorders of communication by what the listener hears is analogous to the attempts of early medicine to classify diseases by the patient's symptoms that the diagnosing physician could see or hear or feel or perhaps smell. Before the great discoveries of the 19th century had erected a logical basis for medical pathology, the various diseases were classified as numerous types of fevers, congestions, dyscrasias, etc. Thus, malaria was originally thought to be caused by the evil emanations (miasma) of the bad air (*mal aria*) near swamps

until it was recognized to be caused by a blood parasite transmitted by the mosquito.

The various approaches of medical, psychiatric, psychological, educational, behavioral, and other schools of speech pathology have made great advances in the recent past and better systems of classification continue to be proposed. They aim at grouping the observable symptoms of speech disorders according to the underlying origins instead of the listener's subjective impressions. While this is relatively easy in the case of language loss from, for example, a brain stroke because the destroyed brain areas can be identified at autopsy, it is more difficult in the case of the large group of so-called functional speech disorders for two reasons: first, they are definitely not caused by gross, easily visible organic lesions, and, second, many functional disorders are outgrown through maturation or appropriate learning (laboratory study of the involved tissues in such cases would reveal no detectable lesions). It is hoped that refined methods of study in the areas of both "functional" psychology and "organic" neurophysiology will eventually reveal the structural bases for the prevalent disorders of voice and speech.

Treatment and rehabilitation. The selection of methods in the medical treatment or educational rehabilitation of communication disorders depends primarily on the underlying basis for the disturbance. Any case of chronic hoarseness should be evaluated first by a laryngologist to establish a precise diagnosis. This is particularly important in the older age groups in which an incipient laryngeal cancer is often overlooked because the patient does not pay attention to his deteriorating voice. The prognosis of all cancers becomes rapidly poorer the longer the disease remains unrecognized. As soon as disease of the larynx is excluded as a cause of the vocal complaint, vocal rehabilitation by a competent speech pathologist should be considered.

Treatment by appropriate specialists

Malformations, diseases, or injuries of the peripheral speech mechanism are treated by appropriate specialists; the plastic surgeon repairs a cleft of the palate; the neurologist and internist treat the stroke patient until he has recovered sufficiently to be referred for re-education of his language abilities. The pediatrician treats the child with mental retardation, while the geneticist counsels the family regarding the possible inheritance of the disorder and its future avoidance. Deafness or severe hearing loss in early childhood is a typical cause for severe delay of language development and should be promptly recognized through appropriate examination by the ear surgeon (otologist) and hearing specialist (audiologist). Cases of childhood autism (withdrawal, severe eccentricities) or early schizophrenia are now being recognized with increasing frequency by speech pathologists, child psychiatrists, pediatricians, and clinical psychologists. This multitude of various professional interests in the recognition and rehabilitation of such exceptional children is well served by the coordination of these efforts in the modern team approach. But again, the largest group of disorders of voice and speech has causes other than these grossly organic lesions. They belong within the province of speech rehabilitation by experts in speech pathology and other functional practitioners.

Development of speech correction. That mankind has been troubled by speech afflictions since the beginning of recorded history can be gleaned from numerous remarks in the books of the Bible. Further, many scientific and medical writers from the time of antiquity to the Middle Ages reported observations of speech and voice disorders. The recommended remedies merely reflected the inadequacies of the philosophical or empirical notions of their times. Scientifically oriented speech pathology originated in Germany during the latter part of the 19th century, following closely the development of otolaryngology. Three names stand out in this respect: C.L. Merkel (*Anthropophonik*; 1857), Adolph Kussmaul (*The Disorders of Speech*; 1877), and Hermann Gutzmann, Sr., who became the first professor of speech pathology at the University of Berlin Medical School around 1900.

During the same time, the new science of experimental phonetics was developed by Jean-Pierre Rousselot in Paris, who promptly recognized the great contributions that experimental phonetics could make to the study of normal and disturbed speech. This close collaboration of medical speech pathology with experimental phonetics has remained typical for the European continent where speech correction is customarily carried out under the direction of physicians in the ear, nose, and throat departments of the university hospitals. The designation of speech and voice pathology as logopedics and phoniatrics with its medical orientation subsequently reached many other civilized nations, notably in Japan and on the South American continent. The national organizations in most of these areas are now represented in the International Association of Logopedics and Phoniatrics, which was founded in Vienna in 1924.

Experimental phonetics

The evolution of speech correction in the Anglo-Saxon countries followed a different trend. Although Great Britain has had a long tradition in general and experimental phonetics, its College of Speech Therapists was organized as an examining and supervisory body in 1945. Similar organizations followed in other areas of the British Commonwealth.

American speech pathology elected a different way. The American Speech and Hearing Association (ASHA), founded in 1925 in New York City, became the organizing, examining, and supervisory body for a rapidly growing membership, which surpassed 12,000 by 1970. More than 200 colleges and universities in the United States, many accredited by ASHA, offer degrees in speech pathology and audiology, some including work at the doctoral level. The large majority of ASHA members hold the master's degree and work as speech clinicians in the public school systems. A smaller number with master's degrees and a still smaller number with doctoral degrees staff the more than 300 clinics that deal with communication disorders and that are usually affiliated with hospitals, colleges, universities, and occasionally with civic organizations.

Soviet Russian speech correction originally followed the developments of European logopedics and phoniatrics and is now directed by the Department of Logopedics in the Government Institute of Defectology of the Academy of Pedagogical Sciences in Moscow. One facet of speech pathology in the U.S.S.R. is its emphasis on Pavlovian theory (conditioning and retraining) and intensive use of neuropsychiatric methods, including pharmacology, sleep therapy, and other intensive treatment programs during hospitalization. Similar trends operate in the eastern European countries under Russian influence, such as in Czechoslovakia, where the first independent medical department of logopedics and phoniatrics was organized at the Faculty of Medicine of Charles University in Prague.

MAJOR TYPES OF SPEECH DISORDERS

Voice disorders. In international terminology, disorders of the voice are described as dysphonia. Depending on the underlying cause, the various types of dysphonia are subdivided by the specifying adjective. Thus, a vocal disorder stemming from paralysis of the larynx is a paralytic dysphonia; injury (trauma) of the larynx may produce traumatic dysphonia; endocrine dysphonia reflects the voice changes resulting from disease of the various endocrine glands such as the pituitary. The various dysphonias of clearly organic origin from systemic disease (*e.g.*, muscular, nervous, or degenerative disease afflicting the entire body) or from local laryngeal changes differ in their visible symptomatology, as well as in the perceptual impression produced by the abnormal voice. Nevertheless, it has not yet been possible to define the acoustical alterations in the vocal spectrum that would allow a clear and objective differentiation among the subjective graduations of an abnormal voice as hoarse, harsh, husky, breathy, grating, gravelly, gritty, etc.

Because a large group of dysphonias have no visible laryngeal causes, they are grouped as nonorganic. Two main types of these so-called functional voice disorders may be distinguished: the habitual dysphonias that arise

Emotional causes

from faulty speaking habits and the psychogenic dysphonias that stem from emotional causes. Both of these types of dysphonia again occur in two basic subtypes, the hyperkinetic (overactive) and the hypokinetic (underactive) since emotional disorders interfere with voluntary vocal function by either causing excessive or depressed physiological activity. In the hyperkinetic disorders, the highly coordinated patterns of phonation regress to the primitive, forceful, and exaggerated sphincter action of the larynx as seen during gagging. The result is hyperkinetic dysphonia, the gratingly harsh vocal disorder due to excessive muscular action in a constricted larynx. In the second subtype, the movements for phonation regress even more deeply to the original function of respiration; the sluggish larynx remains more or less open, and the glottis is incompletely closed for phonation, leading to hypokinetic dysphonia with subdued, breathy huskiness.

Disorders of language development. The most frequent speech disorders are those that disturb the child's acquisition or learning of language. While those concerned with modern terminology are striving for an improved classification according to the etiologic (causative) factors, it is still customary to classify these disorders on the basis of the complaint: absence of speech, baby talk, poorly intelligible articulation, lisping, etc. Recent studies of large numbers of children with such developmental language disorders have shown that at least two chief classes of these disorders may be distinguished: general language disability from genetic factors with a familial (inherited) pattern chiefly from the paternal side, and acquired language disorders due to damage sustained before, during, or shortly after birth (*i.e.*, perinatally).

These latter perinatal damages encompass the gamut of toxic, infectious, traumatic, nutritional, hormonal, and other damages that may hurt the growing fetus or young infant. Major and minor birth injury is not an infrequent factor. Hereditary factors also encompass a great variety of genetically predetermined influences, including familial tendency to exhibit slow language development, lesser endowment in the brain area for language, inferior function in the highest brain areas of auditory performance without organic damage to the ears, slow maturation of motor function (including clumsiness and deviation from normal cerebral dominance), and other signs of delayed cerebral growth. Additional environmental causes include poor language patterns used by the family, parental neglect, emotional maladjustment, general weakness from prolonged disease, as well as various socioeconomic, cultural, and other psychological influences.

While some otherwise perfectly normal children, particularly boys, may not elect to begin talking until age three, making good progress in every respect from then on, the absence of speech after age two may be caused by any of the conditions mentioned thus far and would appear to merit prompt investigation. If an organic cause can be detected, the symptomatic description of delayed language development then yields to a specific etiologic (causal) diagnosis. Although it is best to describe the absence of speech in early childhood as simply delayed language development, some American investigators tend to refer to this condition as congenital (present at birth) aphasia, a term rejected by most European scholars who argue that there cannot be an inborn or early acquired aphasia before a language has been learned.

Many children encounter unusual difficulties in mastering the patterns of articulation of their mother tongue and are said to manifest articulatory immaturity (infantile dyslalia). If no organic cause can be found, the probable cause may be delayed maturation of psychomotor skills.

Marked delays of language development are often followed by a period of inability to learn the rules of grammar and syntax at the usual age (dysgrammatism). Though this is often a sign of inherited language disability, it may reflect mental retardation or other types of brain damage.

Some children who have suffered such laboured language development may then go through a period of retarded reading and writing disability, a condition often defined as dyslexia. Again, there are two chief varieties: the primary or developmental reading and writing disability due to constitutional (organic) and hereditary factors, and a large secondary group of symptomatic reading disorders acquired through any of the influences that retard language development in general, including troubles with vision. Practically all investigators agree that primary or developmental dyslexia shows a marked hereditary tendency (is familial) and is typically associated with other disorders of psychomotor development, deviation from the prevalent right handedness, and poor function in the auditory area in the brain, often associated with lack of musical talent. Primary dyslexia is significantly associated with other developmental speech disorders.

Reading and writing disability

Cluttering. Those who have followed large numbers of children with language disabilities for a long time are impressed by the fact that these various manifestations of an underlying developmental disorder often terminate in a peculiar impediment of speech that has been well known to European speech pathologists for many years but is not yet fully recognized in American speech literature: cluttering (or tachyphemia), the hasty, slóppy, erratic, stumbling, jerky, and poorly intelligible speech that may somewhat resemble stuttering (see below) but differs from it markedly in that the clutterer is usually unaware of it, remains unconcerned, and does not seem to fear speaking situations. Its association with other past or persistent signs of subnormal language development also differentiates congenital cluttering from emotional stuttering. Experts are strikingly unanimous in stressing the hereditary nature of cluttering.

Lisping. Although lisping belongs among the articulatory disorders and usually has the same causes as articulatory disorders (dyslalia) in general, it differs from other disorders of articulation in several respects. For one, lisping occurs in various varieties: with the tongue tip protruding between the front teeth, with a slurping noise in the cheek pouch, with the tongue too far back along the palatal midline, with excessive tongue pressure against the teeth; in several snorting and nasal subtypes, as well as through substitute hissing deep in the throat or even within the larynx, such as in cleft-palate speech. Moreover, the causes of lisping include a diverse group of particular conditions: abnormal number or position of teeth; imitation of other lispers; deficiency of palatal closure; slight hearing loss in the high frequencies; as well as several psychologic causes, such as effeminate affectation, infantile mannerisms, or mental disturbance. Lisping is less easily outgrown than the other infantile dyslalias and may persist into adult life if not corrected.

Causes of lisping

Stuttering. Academically known as dysphemia, what is called stuttering in the U.S. is usually named stammering in Great Britain. While everyone seems to know what stuttering sounds like, experts do not agree about what really causes it. In the age groups after puberty, stuttering is the most frequent and conspicuous type of disturbed speech encountered. This is one reason why among the studies dealing with speech pathology in the world literature those devoted to stuttering are the largest single group. Despite numerous and intensive studies of the problem, findings and conclusions are far from unanimous. A great number of theories have been proposed to explain the origin and nature of stuttering, which range from the premise that subtle physical disturbances in the nervous system (so-called neurogenic asynchronies) are responsible to the opinion that psychological maladjustment alone is to blame. The report issued by the U.S. Public Health Service (see above *Prevalence of Speech Pathology*) concludes that "stuttering remains an enigma while illustrating the type of disorder which does not have either a clean-cut organic cause or a clearly habitual basis."

Research findings indicate (as is the case with many developmental speech disorders, particularly language disability, articulatory disorders, reading disability, and cluttering) that trouble with stuttering affects the male sex at least four times more frequently than the female. Hereditary predisposition has been noted in many studies of large groups of stutterers, with evidence for an inherited tendency found among as many as 40 percent of the

stutterers studied. Some experts insist that stuttering is not a single disease entity but that it comprises several types of the disorder with different causes. According to such views, the familial occurrence of stuttering represents a combination of the stuttering symptom with a cluttering tendency that is inherited. Although imitation of another stutterer may form the basis for acquiring the habit, purely psychological explanations that stress parental attitudes in training their children fail to reveal why many stutterers have siblings (brothers or sisters) with perfectly normal speech.

Treatment of stuttering

The treatment of stuttering is difficult and often demands much skill and responsibility on the part of the therapist. The possibility of some specific medical cure seems remote at the present time. Even the most advanced methods of modern psychiatry have failed to produce superior results in treatment. For a time it was hoped that new psychopharmacological drugs (*e.g.*, tranquillizers) might facilitate and accelerate recovery from stuttering, but these efforts have been disappointing thus far. The typical program of management in this disorder is a strict program of psychotherapy (talking freely with a psychiatrist or psychologist so as to reduce emotional problems) supported by various applications of learning theory or behavioral theory (in retraining the stutterer) and other techniques depending on the therapist's position. It is widely agreed that the patient must acquire a better adjustment to the problems of his life and that he needs to develop a technique for controlling his symptoms and fears. Prognosis (predicted outcome of treatment) thus is held to depend greatly on the patient's motivation and perseverance. It is interesting to note that experienced investigators no longer aspire to a "cure" of stuttering through an etiologic (causal) approach. Instead of focussing on underlying causes, they aim at making the patient "symptom-free" via symptomatic therapy.

Prevention of stuttering may be aided through parent counselling. The normal, immature speech of many children is characterized by various nonfluencies; these include hesitations, syllable repetition, groping for the right word, and vocalizations between words such as "ah-ah." Some misguided parents castigate these normal signs of developing speech with various admonitions and, even worse, try to forbid the nonfluencies by mislabelling them as stuttering. In some children, this parental interference associates normal nonfluency with feelings of insecurity and fear, tending to make the child become a real stutterer. Much research has been devoted to this probable etiology for one type of stuttering; its elimination through parental guidance indeed has been reported to help in reducing the number of stutterers.

Dysphasia and aphasia. Dysphasia means the partial or total loss of language as a result of lesions in those parts of the brain that are directly related to language function. Stroke in elderly patients and head injury in younger ones are typical causes. Aphasia is seen most frequently when the left side of the brain is afflicted, as evidenced by paralysis of the right arm and leg. Evidence indicates that the left hemisphere is dominant in all right-handed individuals and in some left-handers as well. Some experts even believe that the left brain hemisphere is dominant for language in most individuals regardless of handedness and that dominance of the right brain is exceptional in some left-handers. According to other opinions, dominance for language is more evenly distributed in both hemispheres in left-handed persons. As explained before, two major brain areas are recognized as intimately associated with language function: Broca's area in the third frontal convolution and Wernicke's area in the posterior third of the upper temporal convolution. The angular gyrus at the junction of the temporal (side), parietal (top), and occipital (back) lobes of the brain is believed to be related to graphic language as used for reading and writing.

Aphasiology, the science of aphasic language loss, is studied by neurologists, neurosurgeons, some phoniatrists, certain speech pathologists, as well as some psychologists and linguists. This diversity of research background accounts, in part, for the great diversity in theoretical approaches to aphasia. Numerous classification schemes that have been proposed vary from simple groupings into a few main types of aphasia to complicated systems with many forms and subtypes of aphasic disturbances. Similar to research in stuttering, the literature on aphasia is exceptionally large and growing.

The phenomenon of aphasia

The essence of aphasia is the loss of memory for the meaning of language and its production. Thus, in the predominantly expressive or motor forms, the patient can no longer remember the intricate patterns for articulation; he can no longer form a word in speaking or writing, even though he may know what he wants to express. In the predominantly receptive or sensory forms, the patient can talk freely, sometimes excessively and incessantly (logorrhea), although with numerous errors and meaningless clichés, but he no longer comprehends what is said to him or what he tries to read. Those who recover from receptive forms of the disorder are likely to explain that during their aphasia spoken language sounded like an unintelligible, alien tongue. The degree to which there is combination of expressive and receptive symptoms varies greatly with the type and extent of brain lesion. There may be total loss of all language functions (global aphasia) to slight residual errors or misunderstandings when the brain damage is only slight or temporary. A major complication of aphasia is the frequent association with right hemiplegia, in which the paralyzed hand is no longer serviceable for writing. Retraining of the left hand for writing may then become necessary.

Management of aphasia has two goals: one, the physical recovery of the patient through treatment by the internist, neurologist, and possibly brain surgeon and two, the re-education of the brain functions that remain, the aim being to help the patient relearn some use of language under the guidance of the speech pathologist. The better the recovery from the brain lesion, the more chances there are of prompt and complete return of language. When the brain losses are permanent, the patient must relearn each word, sentence, and phrase like the young child, albeit in a more cumbersome manner, apparently using brain parts that still function. He is repeatedly shown the picture of an object along with its printed name; these words are spoken to him by his teacher over and over in the hope that he will learn to repeat the word or phrase until he can say it spontaneously. Words that are relevant to the patient's everyday life are emphasized first with due consideration for his interests and past occupation; later he may advance to the use of abstract concepts and of higher levels of language. Various types of automated training programs are available, including a simple play-back device that shows an inserted card with a picture on it, the name of the pictured object printed next to it, and the audible word recorded on a strip of magnetic tape. Devices of this type enable the patient to practice at his own pace.

Symptomatic speech disorders. Under this heading may be summarized various types of communication disorders that develop on the basis of known structural lesions or metabolic disturbances. Etiologic classifications group these impediments according to the types of organic diseases, as well as in respect to the afflicted effector organs (such as the tongue). Disturbed speech from lesions in the various parts of the nervous system is known as dysarthria. Mental retardation usually limits the development of linguistic ability to the same extent as it does intellectual capacity; this language disorder has been described as dyslogia. Mental disturbances can also manifest themselves in linguistic symptoms, such as in the peculiar (dysphrenic) mode of speech among sufferers of schizophrenia. Hearing loss dating from early childhood leads to a typical distortion of the speech pattern for which various names have been coined, such as audiogenic dyslalia. Visible defects in such oral articulators as the lips and teeth limit the mechanics of articulation and thus reduce the quality and intelligibility of speech; such speech problems are known collectively as dysglossia.

Dysarthria. Damage to those parts of the nervous system that regulate the actions of voice and speech cause distinctive alterations of the speech pattern. The most im-

portant disorder of this type is cerebral palsy from brain injury before, during, or soon after birth. The majority of cerebral palsy victims retain normal intelligence, but are handicapped by distortions of voluntary movements, including those for speaking. Just as walking may be stilted and jerky and arm movements crude and uncontrolled, the patterns of voice and speech will reflect the same distortions. Great advances in rehabilitation have been achieved in the recent past, such as with the well-known Bobath method, which is based on learned suppression of primitive reflexes.

Effects of cerebral palsy

Shaking palsy. Another type of dysarthria is observed in cases of Parkinson's disease. This affliction of the nervous system makes body movements either excessively rigid or tremulously repetitive, such as with the so-called "pill-rolling" sign of hand tremor. Voice and speech reflect the same patterns, sometimes appearing as the first obvious symptoms of the disease. Brain surgery is used in an attempt to improve these abnormal body movements, often with good general results, although speech may be little improved or even made worse. Neurologists also are collecting observations on promising drug treatments for Parkinson's disease, such as with L-Dopa.

When a neurologic disease damages the bulbar nuclei (in the brain stem) of the nerves innervating the organs for speech, the effect of bulbar dysarthria is heard. In severe cases, voice may be reduced to a faint grunting, while articulation deteriorates into mumbling or total muteness. Chorea, another neural disorder manifested in sudden jerky movements of the entire body, may be associated with bizarre, explosive disruptions of the speech flow. Cerebellar dysarthria results from disease of the brain part called the cerebellum (which regulates fine motor coordination), leading to various disorganizations of speech including a "drunken" (ataxic) quality.

Mental retardation. Major advances in the science of inheritance (genetics) and in biochemistry have clarified a number of inborn metabolic causes of many types of mental retardation (see HUMAN BEHAVIOUR, INNATE FACTORS IN). Some of these forms of retardation can be avoided through genetic counselling of prospective parents or prevented by prompt treatment following early diagnosis. (In some cases a change in diet is effective.) Improved prenatal care may reduce the size of another group of mentally retarded individuals whose problem stems from brain damage sustained during fetal life before birth. The number of unknown causes in other cases of retardation is still considerable, however.

Management of the mental retardate

The total management of the mentally retarded child concerns a large group of specialists: pediatrician, geneticist, clinical psychologist, audiologist and speech pathologist, and special educators in institutions for such children. Following classification on the basis of a level of educability, the child undergoes a composite training program geared to his capabilities. Intensive training in language usage not only tends to improve his communication skills but also forms a bridge to increased general learning. There has been recent stress on the not infrequent combination of mental retardation with hearing loss; in such cases, the audiologist prescribes a hearing aid and auditory training to improve the receptive aspects of communication.

Language and mental disorder. Expert analysis of the bizarre speech patterns associated with certain psychiatric disturbances is of primary diagnostic significance. If a mute child persists in stereotyped rituals and strange behaviour, a diagnosis of childhood autism is likely to be made. Some experts distinguish this from a similar disorder called childhood schizophrenia, in which previously good general and linguistic development falls apart in association with similarly bizarre behaviour. In adolescence, a sudden change of voice to a shrill falsetto or weird chanting may herald the outbreak of juvenile schizophrenic disease. Infantile lisping, strange distortions of articulation, and various eccentricities in verbal expression are other signs of schizophrasic speech in the adult schizophrenic. Linguistic study may help the psychiatrist in analyzing the patient's ways of thinking and to provide a measure of the sufferer's progress under therapy. Great therapeutic stress is placed on establishing contact with the autistic child by eliciting from him some sort of communication. Language is felt to be one of the best bridges to break open the closed inner world of such children.

Speech of the hard of hearing. Hearing loss that dates from childhood hinders the normal development of language because the most important sensory portal for speech learning remains deficient. Such children learn to say the sounds of speech as they hear them; that is, in a muffled, distorted, or even inaudible fashion. The articulatory disorder (audiogenic dyslalia) usually reflects the measured (audiometric) pattern of hearing loss. If sound waves at high frequencies cannot be heard, speech sounds with formants in that high-frequency region will be affected. The hissing sibilants contain the highest formants and are therefore most typically disturbed by high-frequency hearing loss. The lower frequencies that can still be heard limit the audible formants to this residual range, which transposes the normal formant patterns into the abnormally lowered frequency band. As a result, a sharp hissing S is spoken as a muted Sh, and the light vowels are transformed into their darker counterparts (for example, the word "set" may be pronounced as "shot" or "shöd" as if it were German).

The voice reflects analogous changes. In the case of conductive hearing loss (in which neural structures for hearing are intact), the patient hears himself well through the bones of his skull but cannot hear others. Because his bone conduction remains good, his voice will sound loud to him and he tends to keep it subdued. The opposite occurs with neural (or perceptive) hearing loss in which there is nerve dysfunction. In this case, the patient hears his own voice as poorly as that of others, and he tends to talk in an overloud and screeching voice.

Correction of audiogenic dyslalia may be possible through early fitting of a suitable hearing aid, intensive auditory training, and speech exercises aided by audiovisual playback devices (auditory trainers, tape recorders, visible speech devices, etc.).

Deaf-mutism

Profound or total deafness going back to early childhood without special training inevitably leads to deaf-mutism, the absence of language development as a direct result of deafness. Deaf-mute children have traditionally been educated in special schools for the deaf, where the oral method (showing how to shape the oral structures for each speech sound) of teaching speech has competed with the older manual method of allowing the deaf to communicate through their own gestural or finger spelling. Recent advances in training include the liberal use of amplification devices (*e.g.*, group hearing aids) in all school situations as well as the earliest possible fitting of hearing aids following definite diagnosis. This can be accomplished even when the patient cannot communicate that he hears (*e.g.*, in very young babies) through such methods as recording brain-wave patterns in response to measured sound stimulation (EEG response audiometry).

Speech impediments from defective articulators. Dysglossic (from defective oral structures) disorders in articulation have interested mankind for the longest time; the biblical and poetical interchange of the words for "tongue" and "speech" in many languages has kept alive the assumption that speech originates within the fleshy tongue. For the same reason, the popular term tongue-tied still persists to refer to an abnormally immobile tongue and some assumed resulting limitation of linguistic ability. Even the academic designation of the science of languages as "linguistic" is traditionally based on the Latin word *lingua* for tongue.

Tongue-tie. In practice, the condition of a true tongue-tie (ankyloglossia) occurs only rarely and is quite easily corrected through a simple operation. Even when the shortened band beneath the tongue tip is permitted to persist, very little speech disturbance, if any, is audible in such persons. The only limitation of articulation to be expected is the decreased ability to protrude the tongue tip between the teeth for the English Th and the trilling lingual R in southern German, Latin, and Slavic languages. In any event, a true tongue-tie never causes stut-

tering, lisping, or any other of the major speech disorders.

Loss of tongue. Major defects of the tongue from paralysis, injury, or surgery reduce the articulation of the lingual sounds to the same extent that the tongue's mobility is visibly limited. Spontaneous compensation is usually quite prompt, depending on the patient's general linguistic talent. One exception is complete bilateral (both sides) paralysis of the tongue, which causes a very severe disorder of chewing and swallowing as well as severe limitation of speech intelligibility. The total loss of the tongue (true aglossia) from injury or surgery is often amazingly well compensated. Patients can learn to use residual portions of a tongue stump as well as other oral structures to substitute for the missing tongue; indeed, some persons without a tongue have relearned to speak so well that the listener would not suspect its absence.

Nasal speech. Several types of nasal speech are not easily diagnosed; even specialized physicians are often not fully aware of the differences.

Increased nasal resonance leads to open nasality (hypernasal speech), affecting all oral speech sounds that should not be nasal. Organic causes impair the accuracy of palatal occlusion during emission of the nonnasal sounds. Among these are paralysis, congenital malformation, injury, or defects of the palate. The functional causes of palatal sluggishness include imitation, faulty speech habits, dialectal influences, hearing loss, mental retardation, or psychiatric disorders.

Decreased nasal resonance

Decreased nasal resonance produces closed nasality (hyponasal speech), which muffles the three nasal resonants (M, N, and Ng). The best known organic causes are an acute cold, hay fever, large adenoids, as well as all other nasal diseases that obstruct the airway. Functional causes are less frequent, in the form of a rare, faulty speech habit; occasionally the problem comes from mental retardation or from severe language disability.

Mixed nasality poses a serious problem; it stems from the combination of one cause of open nasality with another of closed nasality (one may be of organic type and the other functional, or both may be organic). A typical combination is the open nasality from paralysis (paresis) of the palate (or its congenital deficiency) combined with closed nasality from obstruction of the nasopharynx by adenoids. The resulting mixed nasality causes the nasal resonants to sound muffled and subdued, while careful testing reveals slight open nasality on all oral sounds.

Treatment of nasal speech is unlikely to be successful without prudent balancing of all factors involved. This responsibility rests primarily with the diagnosing otolaryngologist or phoniatrist. The complaint of "talking through the nose" should never warrant a tonsil and adenoid operation without complete evaluation. If this were done in the example cited just above, removal of the obstructing adenoids would alleviate only the closed nasality, while making the open component more severe with marked deterioration of the patient's speech. This dilemma is often present when adenoids cause chronic middle-ear disease with hearing loss, while the palate is incompetent, as from a slight congenital malformation or paresis. In such case, the patient's general health and his hearing must be weighed against the possibility of making his speech worse.

Cleft palate speech. This type of organic dysglossia has also been named rhinoglossia (Greek *rhin, rhis*: "nose") because it is an organic cause of excessively nasal speech. Clefts of the lip, upper jaw, and hard and soft palate occur in various types and combinations. Cleft palate is a congenital (present at birth) malformation that develops for various reasons during the early weeks of embryonic life. The causes may be grouped as follows: inheritance in some cases; embryonal damage from various environmental causes in others; while still other cases of cleft palate are part of a syndrome of general or multiple malformation, which may again be hereditary, the result of chromosomal aberrations, or of environmental origin. Genetic influences are inferred from evidence that North American Indians (Montana) show a high incidence of cleft palate, but Negroes (as on the island of Jamaica) show an unusually low incidence.

Hereditary influences

The speech disorder in cleft palate cases is complex. As a direct, mechanical result of the absent velopharyngeal (soft palate-pharynx) closure, voiced sounds assume an abnormal nasal quality from the unusual resonance of the nasal chambers. The influence of nasality on the perceptual and acoustic characteristics of the speech sounds has been discussed earlier. A second component is a mechanical limitation in articulating plosive sounds (*e.g.*, P) because the necessary intra-oral pressure cannot be achieved. For the same reason, the sibilant sounds (*e.g.*, S) are greatly distorted because the articulating air escapes through the nose before it can produce the characteristic hissing noise at one of the places of constriction. The child with an unrepaired cleft palate seems unconsciously to attempt to overcome these mechanical limitations in several ways. He tries to constrict the nostrils through facial grimaces, and he attempts to increase the expiratory air pressure. He also succumbs to the "law of centripetal regression of articulation"; that is, he shifts the articulation of the oral sounds below the point of deficient palatal occlusion, down into the pharynx or even the larynx. It is easy to demonstrate by X-ray fluoroscopy how the larynx becomes constricted during each plosive or sibilant sound in such cases.

Mechanism of cleft palate disorder

These mechanical limitations of articulation in cleft palate lead to secondary modifications of behaviour. Linguistic ability is retarded through general limitation of speech learning; emotional reactions to the handicap may complicate the clinical picture; a vicious circle is thus established, which in turn impedes the sufferer's general intellectual and special linguistic development.

Other coincidental disorders complicate the pattern of cleft-palate speech. Many cleft-palate patients also suffer from middle-ear infection, and the resulting hearing loss may add the component of audiogenic dyslalia. In other cases, the structural malformation may be complicated by mental retardation. Coincidental genetic factors may aggravate the speech problem when the syndrome of familial (neurally inherited) language disability is superimposed. In such cases, the mechanical speech disorder of rhinoglossia becomes complicated by the additional signs of delayed speech development, articulatory dyspraxia, dysgrammatism, reading and writing disability, etc. This combination is evident in some cases following the successful surgical correction of the palatal deficiency. Although the nasality is promptly alleviated in such cases, the various speech disorders from language disability may persist, since they have nothing directly to do with the previous cleft palate.

Treatment of cleft palate

Management of cleft palate is the task of a well-coordinated team. The plastic surgeon closes the palatal deficiency, the psychologist analyzes the intellectual endowment, the speech pathologist corrects the various components of disordered speech, the audiologist determines coincidental hearing loss, and the otolaryngologist cares for the health of nose and ears. Tonsils and adenoids may be diseased, leading to the frequent complication of middle-ear infection with hearing loss. This condition may require removal of tonsils and adenoids to preserve hearing. A dilemma is then created; removal of tonsils and adenoids is generally contra-indicated in cases of deficient palatal closure because this operation increases the degree of open nasality. The child's welfare then should be carefully considered in order to avoid any harm.

When surgical repair is not feasible, the palatal defect may be covered by a special prosthetic plate (obturator) similar to false dental appliances. This technique has been known for many centuries, and various models of obturators have been constructed in the course of time. Cleft-palate care therefore includes the services of a prosthodontist (who makes false teeth) for the optimal construction of such appliances. If worn by children, obturators should be regularly altered to maintain continuous fit as the patient grows.

Well-organized centres for the management of cleft palate are able to offer excellent care so that the speech impediment can be ideally or almost completely corrected in many cases. The problem of cleft palate thus demonstrates once more that human speech requires complex

detailed study of its normal functions and manifold approaches for the correction of its imperfections.

BIBLIOGRAPHY. R. LUCHSINGER and G.E. ARNOLD, *Handbuch der Stimm- und Sprachheilkunde*, 3rd ed. (1970; Eng. trans., *Voice, Speech, Language*, 2nd ed., 1972), the largest text on disorders of communication with extensive references and international orientation; R. CARHART (ed.), *Human Communication and its Disorders: An Overview* (1969), an important report of the U.S. Department of Health, Education and Welfare on statistics, established facts, and research needs; L.E. TRAVIS (ed.), *Handbook of Speech Pathology* (1957), a fundamental American publication by many leading contributors; C. VAN RIPER, *Speech Correction: Principles and Methods*, 4th ed. (1963), a widely used text for study and references; R.W. WEST and M. ANSBERRY, *The Rehabilitation of Speech*, 4th ed. (1968), another classic in the field of speech pathology; V.E. NEGUS, *The Comparative Anatomy and Physiology of the Larynx* (1950), an excellent introduction to the understanding of function and evolution of the voice organ in man and animals; G. FANT, *Acoustic Theory of Speech Production* (1960), a modern, detailed exposition of acoustic phonetics, the science of the acoustic structure of speech sounds and their production; H. FLETCHER, *Speech and Hearing in Communication*, 2nd ed. (1953), a classic work on the physiology of speech and hearing; P.B. DENES and E.N. PINSON, *The Speech Chain* (1963), highly recommended for high school and college level students.

See also recent issues of these journals: *American Speech and Hearing Association* (ASHA) *Journal* (monthly); *Journal of Speech and Hearing Disorders* (quarterly); *Journal of Speech and Hearing Research* (quarterly); *Cleft Palate Journal* (quarterly), all U.S.; *Folia Phoniatrica* (bimonthly, Basel); and the *British Journal of Disorders of Communication* (semi-annual, London).

(G.E.A.)

Spencer, Herbert

Herbert Spencer, an early upholder of the theory of evolution, maintained that every aspect of reality must be viewed in terms of a continuing development from lower to higher stages. He thus regarded human progress as "not an accident but a necessity." His naturalistic philosophy, a broad synthesis of knowledge buttressed by a detailed, scientific examination of a vast range of biological and social phenomena, significantly influenced the development of biology, psychology, anthropology, and sociology.

Life and works. Spencer was born in Derby on April 27, 1820. His father, William George Spencer, was a schoolmaster; and his parents' dissenting religious convictions inspired in him a nonconformity that continued active even after he had abandoned the Christian faith. Spencer declined an offer from his uncle, the Rev. Thomas Spencer, to send him to Cambridge, and in consequence his higher education was largely the result of his own reading, which was chiefly in the natural sciences. He was, for a few months, a schoolteacher and from 1837 to 1841 a railway civil engineer.

In 1842 he contributed some letters (republished later as a pamphlet, *The Proper Sphere of Government*, 1843) to *The Nonconformist*, in which he argued that it is the business of governments to uphold natural rights and that they do more harm than good when they go beyond this. After some association with progressive journalism through such papers as *The Zoist* (devoted to mesmerism and phrenology) and *The Pilot* (the organ of the Complete Suffrage Union), Spencer became in 1848 a subeditor of *The Economist*. In 1851 he published *Social Statics* (reissued in 1955), which contained in embryo most of his later views, including his argument in favour of an extreme form of economic and social laissez-faire. About 1850 Spencer became acquainted with Marian Evans (the novelist George Eliot), and his philosophical conversations with her led some of their friends to expect that they would marry; but in his *Autobiography* (1904) Spencer denies any such desire, much as he admired Miss Evans' intellectual powers. Other friends were G.H. Lewes, T.H. Huxley, and J.S. Mill. In 1853 Spencer, having received a legacy from his uncle, resigned his position with *The Economist*.

Having published the first part of *The Principles of Psychology* in 1855, Spencer in 1860 issued a prospectus



Spencer.
The Gernsheim Collection, the University of Texas at Austin

and accepted subscriptions for a comprehensive work, *The Synthetic Philosophy*, which was to include, besides the already published *Principles of Psychology*, volumes on first principles and on biology, sociology, and morality. *First Principles* was published in 1862, and between then and 1896, when the third volume of *The Principles of Sociology* appeared, the task was completed. In order to prepare the ground for *The Principles of Sociology*, Spencer started in 1873 a series of works called *Descriptive Sociology*, in which information was provided about the social institutions of various societies, both primitive and civilized. The series was interrupted in 1881 because of lack of public support. Spencer was a friend and adviser of Beatrice Potter, later Beatrice Webb, the social reformer, who frequently visited Spencer during his last illness and left a sympathetic and sad record of his last years in *My Apprenticeship* (1926). Spencer died on Dec. 8, 1903, at Brighton, leaving a will by which trustees were set up to complete the publication of the *Descriptive Sociology*. The series comprised 19 parts (1873–1934).

Writing of *The Synthetic Philosophy*

Spencer was one of the most argumentative and most discussed English thinkers of the Victorian period. His strongly scientific orientation led him to urge the importance of examining social phenomena in a scientific way. He believed that all aspects of his thought formed a coherent and closely ordered system. Science and philosophy, he held, gave support to and enhanced individualism and progress. Though it is natural to cite him as the great exponent of Victorian optimism, it is notable that he was by no means unaffected by the pessimism that from time to time clouded the Victorian confidence. Evolution, he taught, would be followed by dissolution, and individualism would come into its own only after an era of socialism and war.

The synthetic philosophy in outline. Spencer saw philosophy as a synthesis of the fundamental principles of the special sciences, a sort of scientific summa to replace the theological systems of the Middle Ages. He thought of unification in terms of development, and his whole scheme was in fact suggested to him by the evolution of biological species. In an article, "The Development Hypothesis" (1852), Spencer rejected the notion of special creation; and in "Progress: its Law and Cause" (1857), he applied the idea, borrowed from the Estonian naturalist, K.E. von Baer, that biological development in the individual is from the homogeneous to the heterogeneous to the evolution of the solar system; of animal species; of human society; and of industry, art, language, and science. It should be noticed that Spencer published his idea of the evolution of biological species before the views of Charles Darwin and the British naturalist Alfred Russel

From the homogeneous to the heterogeneous

Wallace were known; but Spencer at that time thought that evolution was caused by the inheritance of acquired characteristics, whereas Darwin and Wallace attributed it to natural selection. Spencer later accepted the theory that natural selection was one of the causes of biological evolution, and himself coined the phrase "survival of the fittest" (*Principles of Biology* [1864], vol. 1, p. 444). He did not, however, posit any of these causes as ultimate principles of evolution. In his view the principle of change from homogeneity to heterogeneity was more important and was itself a consequence of principles still more fundamental. In *First Principles* he argued that there is a fundamental law of matter, which he called the law of the persistence of force, from which it follows that nothing homogeneous can remain as such if it is acted upon, because any external force must affect some part of it differently from other parts and cause difference and variety to arise. From this, he continued, it would follow that any force that continues to act on what is homogeneous must bring about an increasing variety. This "law of the multiplication of effects," due to an unknown and unknowable absolute force, is in Spencer's view the clue to the understanding of all development, cosmic as well as biological.

Sociology and social philosophy. That Spencer first derived his general evolutionary scheme from reflection on human society is seen in *Social Statics,* in which social evolution is held to be a process of increasing "individuation." He saw human societies as evolving by means of increasing division of labour from undifferentiated hordes into complex civilizations. Spencer believed that the fundamental sociological classification was between military societies, in which cooperation was secured by force, and industrial societies, in which cooperation was voluntary and spontaneous.

Military versus industrial societies

Evolution is not the only biological conception that Spencer applied in his sociological theories. He made a detailed comparison between animal organisms and human societies. In both he found a regulative system (the central nervous system in the one, government in the other), a sustaining system (alimentation in the one case, industry in the other), and a distribution system, (veins and arteries in the first; roads, telegraphs, etc., in the second). The great difference between an animal and a social organism, he said, is that whereas in the former there is one consciousness relating to the whole, in the latter consciousness exists in each member only; society exists for the benefit of its members and not they for its benefit.

Liberal principle of variety

This individualism is the key to all of Spencer's work. His contrast between military and industrial societies is drawn between despotism, which is primitive and bad, and individualism, which is civilized and good. He believed that in industrial society the order achieved, though planned by no one, is delicately adjusted to the needs of all parties. In *The Man versus the State* (1884) he wrote that England's Tories generally favour a military and Liberals an industrial social order but that the Liberals of the latter half of the 19th century, with their legislation on hours of work, liquor licensing, sanitation, education, etc., were developing a "New Toryism" and preparing the way for a "coming slavery." "The function of liberalism in the past was that of putting a limit to the powers of kings. The function of true liberalism in the future will be that of putting a limit to the powers of parliaments."

Metaphysics. In his emphasis on variety and differentiation, Spencer was unwittingly repeating, in a 19th-century idiom, the metaphysics of liberalism that Spinoza and Leibniz had adumbrated in the 17th century. Spinoza had maintained that "God or Nature" has an infinity of attributes in which every possibility is actualized, and Leibniz had argued that the perfection of God is exhibited in the infinite variety of the universe. Though neither of them believed that time is an ultimate feature of reality, Spencer combined a belief in the reality of time with a belief in the eventual actualization of every possible variety of being. He thus gave metaphysical support to the liberal principle of variety, according to which a differentiated and developing society is preferable to a monotonous and static one.

Evaluation. Spencer's attempt to synthesize the sciences showed a sublime audacity that has not been repeated because the intellectual specialization he welcomed and predicted increased even beyond his expectations. His sociology, although it gave an impetus to the study of society, was superseded as a result of the development of social anthropology since his day and was much more concerned with providing a rationale for his social ideals than he himself appreciated. Primitive men, for example, are not the childlike emotional creatures that he thought them to be, nor is religion to be explained only in terms of the souls of ancestors. When T.H. Huxley said that Spencer's idea of a tragedy was "a deduction killed by a fact," he called attention to the system-building feature of Spencer's work that led him to look for what confirmed his theories and to ignore or to reinterpret what conflicted with them.

MAJOR WORKS

PHILOSOPHY AND RELIGION: *The Nature and Reality of Religion,* 1885 (withdrawn from publication). Spencer's series on *Synthetic Philosophy* comprises: *First Principles,* 1862; *The Principles of Biology,* 2 vol., 1864–67; *The Principles of Psychology,* 1855; *The Principles of Sociology,* 3 vol., 1876–96; *The Principles of Ethics,* 2 vol., 1892–93.

POLITICAL AND SOCIAL: *The Proper Sphere of Government,* 1843; *Social Statics,* 1851; *Education: Intellectual, Moral, Physical,* 1861; *The Study of Sociology,* 1872; *The Man versus the State,* 1884; *Facts and Comments,* 1902.

OTHER WORKS: *Essays: Scientific, Political, and Speculative,* 3 vol., 1891; *Autobiography,* 1904, an intellectual rather than a personal autobiography.

BIBLIOGRAPHY. A so-called official biography is that by DAVID DUNCAN, *Life and Letters of Herbert Spencer,* 2 vol. (1908). T.H. HUXLEY, in "Evolution and Ethics," in *Collected Essays,* vol. 9 (1894), accuses Spencer of "fanatical individualism" and "reasoned savagery." H. SIDGWICK, *Lectures on the Ethics of T.H. Green, Mr. Herbert Spencer and J. Martineau,* pp. 135–312 (1902); and G.E. MOORE, *Principia Ethica,* sect. 31–35 (1903, reprinted 1954), contain detailed criticisms of Spencer's ethical theory. See also J. RUMNEY, *Herbert Spencer's Sociology: A Study in the History of Social Theory* (1934); S. ANDRESKI (ed.), *Principles of Sociology,* abr. ed. (1969), which includes an analysis and assessment of Spencer's sociological theory; D. MACRAE (ed.), *The Man Versus the State: With Four Essays on Politics and Society* (1969), in which the introduction (pp. 7–54) deals with Spencer's views on society and politics; and J.Y.D. PEEL, *Herbert Spencer: The Evolution of a Sociologist* (1971).

(H.B.A.)

Spenser, Edmund

A contemporary of William Shakespeare, Edmund Spenser was considered in his day to be the greatest of English poets, who had glorified England and her language by his long allegorical poem *The Faerie Queene,* just as Virgil had glorified Rome and the Latin tongue by his epic poem the *Aeneid.* Spenser had a strong influence upon his immediate successors, and the sensuous features of his style, as well as his nine-line stanza-form, were later admired and imitated by such poets as Byron and Shelley in the Romantic Period of the late 18th and early 19th centuries. He is widely studied today as one of the chief begetters of the English literary Renaissance, a golden age for poetry in English, and as a master who embodied in poetic myth a view of the virtuous life in a Christian universe.

Youth and education. He did all this in a life-span of about 46 years. Very little is certainly known about the man. He was born in London, most probably in 1552 or 1553. He was related to a noble Midlands family of Spencer, whose fortunes had been made through sheep raising, and to three of whom—Ladies Carey, Compton, and Strange—he dedicated poems. His own immediate family was not wealthy. He was entered as a "poor boy" in the Merchant Taylors' grammar school, headed by Richard Mulcaster, a teacher with advanced ideas and respect for English as a literary medium. Still, Spenser, in 45 hours of class per week, would have studied mainly Latin, with some Hebrew, Greek, and music.

In 1569, when he was about 16, his English versions of



Spenser, oil painting by an unknown artist. In the collection of Pembroke College, Cambridge.
By courtesy of the Master and Fellows, Pembroke College, Cambridge

poems by the 16th-century French poet Joachim du Bellay and his translation of a French version of a poem by the Italian poet Petrarch appeared at the beginning of an anti-Catholic prose tract, "The Theatre for Worldlings"; they were no doubt commissioned by its chief author, the wealthy Flemish expatriate Jan van der Noodt. (Some of these poems he later revised for his *Complaints* volume.)

Sizar at Cambridge

From May 1569 Spenser was a student in Pembroke Hall (now Pembroke College) of the University of Cambridge, where, along with perhaps a quarter of the students, he was classed as a sizar—a student who, out of financial necessity, performed various menial or semimenial duties. According to the Statutes of 1570 (which, like many such regulations, may have been an attempt to enforce what in part was unenforceable), his day would have run from 5 AM to 8 or 9 PM, with compulsory attendance at worship and lectures. For his bachelor of arts degree, which he received in 1573, he would have studied, in Latin, the ancient authorities on rhetoric, logic, and philosophy. For his master's degree he would have studied Greek language and literature, arithmetic, geometry, and the astronomy of the Ptolemaic system. Because of an epidemic he in fact left Cambridge in 1574 but received his master of arts degree in 1576.

His best known friend at Cambridge was the slightly older Gabriel Harvey, a fellow of Pembroke, who was learned, witty, and enthusiastic for ancient and modern literature but also pedantic, devious, and ambitious for the life of action and power at the same time that he was unaware of his own shortcomings for this role. There is no reason to believe that Spenser shared the most distasteful of these qualities, but, in the atmosphere of social mobility and among the new aristocracy of Tudor England, it is not surprising that Spenser hoped for preferment to higher position. He shows, at some points in his career, symptoms of frustration over the unresponsiveness of the powerful to his needs and wishes, but the evidence does not suggest that this frustration was obsessive.

His period at Cambridge was undoubtedly important for the acquisition of his very wide knowledge not only of the Latin and some of the Greek classics but also of Italian, French, and English literature of his own and earlier times. His knowledge of the traditional forms and themes of lyrical and narrative poetry provided foundations for him to build his own highly original compositions. Without the Roman epic poet Virgil's *Aeneid*, the 15th-century Italian Ludovico Ariosto's *Orlando Furioso*, and, later, Torquato Tasso's *Gerusalemme liberata* (1581), Spenser could not have written his heroic, or epic, poem *The Faerie Queene*; without Virgil's *Bucolics* and the later tradition of pastoral poetry in Italy and France, he could not have written *The Shepheardes Calender*; and without the Latin, Italian, and French examples of the highly traditional marriage ode and the sonnet and canzone forms of Petrarch and succeeding sonneteers, he could not have written his greatest lyric, *Epithalamion*, and its accompanying sonnets, *Amoretti*. The patterns of meaning in his poetry are often woven out of the traditional interpretations—developed through Classical times and his own—of pagan myth, divinities, and philosophies and out of an equally strong experience of the faith and doctrines of Christianity; these patterns he further enriched by the use of medieval and contemporary story, legend, and folklore.

Spenser's religious training was a most important part of his education. He could not have avoided some involvement in the bitter struggles that took place in his university over the path the new Church of England was to tread between Roman Catholicism and extreme Puritanism. Contrary to a former view, there is little reason to believe that he inclined toward the Puritanical side. His first known appointment (after a blank of several years, when he may have been in the north of England) was in 1578 as secretary to Bishop John Young of Rochester, formerly master of Spenser's college at Cambridge. His first important publication, *The Shepheardes Calender* (1579 or 1580), is more concerned with the bishops and affairs of the English Church than is any of his later work.

Early works. *The Shepheardes Calender* can be called the first work of the English literary Renaissance. Following the example of Virgil and of many later poets, Spenser was beginning his career with a series of eclogues (literally "selections," usually short poems in the form of pastoral dialogues), in which various characters, in the guise of innocent and simple shepherds, converse about life and love in a variety of elegantly managed verse forms, formulating weighty—often satirical—opinions on questions of the day. The paradoxical combination in pastoral poetry of the simple, isolated life of shepherds with the sophisticated social ambitions of the figures symbolized or discussed by these shepherds (and of their probable readership) has been of much interest in modern literary criticism. The *Calender* consists of 12 eclogues, one named after each month of the year. One of the shepherds, Colin Clout, who excels in poetry but is ruined by his hopeless love for one Rosalind, is Spenser himself. The eclogue "April" is in praise of the shepherdess Elisa, really the Queen (Elizabeth I) herself; "October" examines the various kinds of verse composition and suggests how discouraging it is for a modern poet to try for success in any of them. "November," a lament for a lady called Dido, may covertly express regret for what seemed in 1579 the Queen's intended French marriage, which was strongly opposed at home by Robert Dudley, earl of Leicester, Sir Philip Sidney, and others of Spenser's acquaintance. Most of the eclogues, however, concern good or bad shepherds—that is to say, pastors—of Christian congregations, including his own beloved "Roffy," Bishop Young. The *Calender* was well received in its day, and it is still a revelation of what could be done poetically in English after a long period of much mediocrity and provinciality. The archaic quality of its language, sometimes deplored, was partly motivated by a desire to continue older English poetic traditions, like that of Chaucer. Archaic vocabulary is not so marked a feature of his later work.

Initiator of the English literary Renaissance

The two years 1578–80 probably produced more changes in Spenser's life than did any other corresponding period. He appears by 1580 to have been serving the fascinating, highly placed, and unscrupulous Earl of Leicester and to have become a member of the literary circle led by Sir Philip Sidney, Leicester's nephew, to whom the *Calender* was dedicated and who praised it in his important critical work *The Defence of Poesie* (1595). Spenser remained permanently devoted to this brilliant writer and good nobleman, embodied him variously in his own poetry (for example, as the Sir Calidore of *The Faerie Queene*, Book VI), and mourned his early death in an elegy. By 1580 Spenser had also started work on *The Faerie Queene* and in the previous year had apparently married one Machabyas Chylde. It is an interesting sidelight on his personal character, of which next to

Relations with Sidney and Harvey in 1578–80

nothing is known, that at a meeting with Gabriel Harvey in London in 1578, Spenser gave his friend four English books of humorous tales with the proviso that if Harvey did not read them all by January 1, he should surrender to Spenser his four volumes in Greek of the amusing writer Lucian. From another of Harvey's notations it is known that Spenser gave him a little book recommending foreign travel, with practical hints and a description of Naples (though nothing is known of Spenser's ever having travelled abroad). Later notations suggest that by 1580 Spenser may have been living as a modish young man about town, with Italianate and French leanings. Latin verses by Spenser sent to Harvey in 1579 show him eager to undertake a contemplated (but apparently never completed) trip to France on Leicester's business but reluctant to leave behind someone in England with whom he was head over heels in love. He asserts that Harvey would never thus have allowed love to stand in the way of ambition. With the same letter went another one, also published, pointing out the danger to himself of dedicating a work to Sidney if Sidney might happen to consider it insufficiently meritorious. (Unless Harvey published these letters without Spenser's prior knowledge, it is difficult to understand how this latter observation was allowed to become public property.) The ironies in Harvey's and Spenser's exchange of letters are so intricate as to make it difficult to draw further conclusions from them about Spenser, except that he was young, ambitious, accomplished, and sincerely interested in the theory and practice of poetry. Finally, in 1580, at the end of this two-year period, he was made secretary to the new lord deputy of Ireland, Arthur Lord Grey, who was a friend of the Sidney family. Spenser's life was to lie mostly in Ireland for his remaining 19 years.

Career in Ireland. Sixteenth-century Ireland and the Irish were looked on by the English as a colony, although the supposed threat of an invasion by Spain, the Catholic enemy, and the conflict between an imposed English Church and the at first decadent Catholicism of the Irish, were further complicating factors. Irish chieftains and the Anglo-Irish nobility, dating from as far back as the Norman conquest, encouraged native resistance to newly arrived English officials and landowners. As Grey's secretary, Spenser accompanied the Lord Deputy on risky military campaigns as well as on more routine journeys.

Service in Ireland

But Ireland was certainly no exile for Spenser. He became deeply attached to the country, its people and its landscapes. He several times gives poetic form to romantic or idyllic reactions to the Irish scene (as in "Mutabilitie Cantos," VII, vi, 37–39; *Colin Clouts Come home againe*, 88–155), along with an equally strong dislike for its unproductive wildness and lawlessness (as in "Mutabilitie Cantos," VII, vi, 54–55). The conflict between Grey's direct, drastic governmental measures and the Queen's characteristic procrastinating and temporizing style soon led to his frustration and recall. But Spenser, like many others, admired Grey's methods; his later tract, *A View of the Present State of Ireland* (written 1595–96, published 1633), argues lucidly for a typically 16th-century theory of rule: firm measures, consistently applied, with gentleness only for completely submissive subject populations. At the point where Spenser, treating the virtue of Justice in *Faerie Queene*, Book V, embodies the injustice of another national state against his own, his hero Sir Artegall, in conquering Grantorto, becomes temporarily a projection of Grey in his victory over a small Spanish expeditionary force in Ireland.

For four or five years from about 1584, Spenser carried out the duties of a second important official position in Ireland, deputizing for his friend Lodowick Bryskett as clerk of the lords president (governors) of Munster, the southernmost Irish province. The fruits of his service in Ireland are plain. He was given a sinecure post and other favours, including the right to dispose of certain forfeited parcels of land (he no doubt indulged in profitable land speculation). For a time he leased the small property of New Abbey, County Kildare, and on this basis was first designated "gentleman." Finally, he obtained a much larger estate in Munster. One of the chief preoccupations of the presidents of this province, scarred as it was by war and starvation, was to repopulate it. To this end, large "plantations" were awarded to English "undertakers," who undertook to make them self-sustaining by occupying them with Englishmen of various trades. In 1588 or 1589 Spenser took over the 3,000-acre plantation of Kilcolman, about 25 miles to the north and a little to the west of Cork. No doubt he took there his son and daughter and his wife, if she was still alive (she is known to have died by 1594). By acquiring this estate, Spenser made his choice for the future: to rise into the privileged class of what was, to all intents, a colonial land of opportunity rather than to seek power and position on the more crowded ground of the homeland where he had made his poetic reputation (and where he was to increase it still further). In his new situation he, like other undertakers, had much conflict with the local representative of the older Anglo-Irish aristocracy, Maurice Lord Roche; further, they had very limited success in filling the plantations with English families. Even after the defeat of the Armada (1588), the danger of Spanish invasion was taken seriously, and the threat of an Irish uprising was to prove all too real ten years later. Nevertheless, no matter what may be thought about the justice or the viability of the English hegemony in Ireland, it was under these conditions that Spenser brought his greatest poetry to completion.

Spenser's congenial literary friend in Ireland was the older official Lodowick Bryskett, a man of cosmopolitan culture. The son of an Italian merchant who had settled in England, he had attended Cambridge and accompanied Sir Philip Sidney on extensive Continental travels. In a prose dialogue he shows us Spenser (no doubt partly fictionalized), proclaiming that his *Faerie Queene* is to be composed in several books, each devoted to a single virtue. The date pointed to seems to be 1582, but Bryskett's book was published long after Spenser had made the same point in his introductory letter to Raleigh, in the first edition of his great poem, which purported to explain what was to be the general scheme of the work.

"The Faerie Queene." In its present form, *The Faerie Queene* consists of six books and the fragment (known as the "Mutabilitie Cantos") of a seventh. According to the introductory letter, it was to contain 12 books each telling the adventure of one of Gloriana's knights (the origin of these adventures was to be explained in the last book, not extant). Like other poets, Spenser must have modified his general plan many times, yet this letter, inconsistent though it is with various plot details in the books which are extant, is probably a faithful mirror of his thinking at one stage. The stories actually published were those of Holiness (the Red Cross Knight), Temperance (Sir Guyon), Chastity (Britomart, a female knight), Friendship (ostensibly concerning Triamond and Cambello, although these play a very small part), Justice (Artegall), and Courtesy (Calidore). What is most characteristic of Spenser in this work is his serious view of the capacity of the romance form to act as a paradigm of human experience: the moral life as quest, pilgrimage, aspiration; as eternal war with an enemy, still to be known; as encounter, crisis, the moment of illumination—ethics with the added dimensions of mystery, terror, love, and victory and with all the generous virtues exalted. Modern impatience with the obscure allusions in the poem, with its political and ecclesiastical topicalities, is a failure to share the great conflict of Spenser's time between Protestant England and Catholic Spain—to him, the war between good and evil was here and now. As a setting, Spenser invented the land of Faerie and its queen, Gloriana. To express himself he invented a nine-line stanza, the first eight of five stresses and the last of six, whose rhyme pattern is ababbcbcc. Spenser proves himself a master: picture, music, metre, story—all are at one with the deeper significance of his poem, providing a moral heraldry of colours, emblems, legends, folklore, mythical allusion, all prompting deep, instinctive responses. Enjoyment of the poem's sensuous surface is itself to undergo an experience, an ascent in vision with the protagonists.

Presentation of *The Faerie Queene* to Elizabeth I

The poem was published with the help of Sir Walter Raleigh, who owned large lands to the east of Spenser's estate. He and the poet came together at Kilcolman in 1589 and became well acquainted with one another's poetry. Spenser implies that Raleigh persuaded Spenser to accompany him back to England to present the completed portion of *The Faerie Queene* to Queen Elizabeth herself. The history of this episode is charmingly evoked in *Colin Clouts Come home againe*, which is also one of Spenser's most effective pastoral embodiments of a provincial innocent up against the sophistications of a centre of power, with subsequent reflections on false, superficial love and the true love that finally animates a concordant universe.

Arriving thus in London with the support of the Queen's favourite, Spenser was well received—not least by Elizabeth herself. The first three books of *The Faerie Queene* were duly published in 1590, together with a dedication to her and commendatory sonnets to notables of the court. These included, as an afterthought, Lord Burghley, who was perhaps the most powerful figure there. Spenser saw the book through the press, made a hurried visit to Ireland, and returned speedily to London—presumably in the hope of preferment. At this time he supervised the printing of certain other of his poems in a collection called *Complaints*, many of which had probably been written earlier in his career and were now being published so as to profit from the great success of his new heroic poem. It is difficult to believe that the many titles of poems which have not survived but were mentioned earlier in his career (including "Dreames," "Dying Pelicane," and many more) were not published in revised form and under other titles in his known work; for *Complaints* suggests by its miscellaneous and uneven character that Spenser was hastily bringing to the light of day nearly every last shred that he had to offer: early translations, an elegy, and the delightful mock-heroic poem "Muiopotmos" are contained in it; another item, the beast fable "Prosopopoia, or Mother Hubberd's Tale," apparently caused the authorities to withdraw unsold copies of the volume (perhaps in 1592) because it contained a covert attack on Lord Burghley. Spenser seems to have believed that this lord's powerful influence was blocking his own and other worthy men's progress. Nevertheless, in 1591 the Queen gave Spenser a small pension for life.

Last years. Back in Ireland, Spenser pressed on with his writing, in spite of the burdens of his estate. In early 1595 he published *Amoretti* and *Epithalamion*, a sonnet sequence and a marriage ode celebrating Spenser's marriage to his second wife (Elizabeth Boyle) after what appears to have been an impassioned courtship in 1594. This group of poems is unique among Renaissance sonnet sequences in that it celebrated a successful love affair culminating in marriage. As in passages in *The Faerie Queene* (III, vi; VII, vii), Spenser develops in *Epithalamion* an aspect of Platonism (descending from the Greek philosopher's *Timaeus* but also much cultivated during the Middle Ages) according to which the changes wrought by time in our imperfect universe imitate the unchanging excellence of eternity through their cyclical character, continually restoring what they continually remove. Such are day and night, the circling year, the rebirth of plant and animal life, and the reproduction of man under the stimulus of sexual pleasure in marriage. Spenser conveys the unity of these cycles in his marriage ode, where he symbolizes the daily and annual cycles of the sun by giving the poem 24 stanzas (representing the hours) and a total of 365 pentameter and hexameter lines (representing the days).

Second marriage

Other kinds of Platonism, sometimes interwoven with other philosophical strands, appear in the *Amoretti* and in many other of Spenser's works. Platonic doctrines, so popular in Renaissance poetry (that true beauty is not material but is heaven-sent, bespeaking in principle an inner spiritual excellence; and that Love is, in one way, a god who leads men to such beauty and reduces warring chaos to harmony) appear in a number of Spenser's poems, especially in the *Fowre Hymnes* (1596), where Platonism is combined with Christianity. In that year Spenser may have again been busy in London supervising the publication of a number of works, including Books IV, V, and VI of *The Faerie Queene*.

This burst of publication was the last of his lifetime. His early death may have been precipitated by the penetration into Munster of the Irish uprising of 1598. The undertakers and other loyalists failed to make head against this. Kilcolman was burned, and Spenser, probably in despair despite the Privy Council's having just recommended his appointment to the important post of sheriff of Cork, carried official letters about the desperate state of affairs from the president to London. He died there on January 13, 1599, but the long-held belief that it was "for lack of bread" (as the poet Ben Jonson told William Drummond of Hawthornden in 1618) is incredible, considering his official status and patronage. He was, in fact, buried with ceremony in Westminster Abbey close by the grave of Geoffrey Chaucer.

MAJOR WORKS

POEMS: *The Shepheardes Calender* (1579); *The Faerie Queene*, Books I-III (1590); *Complaints* (1591); *Daphnaïda* (1591); *Amoretti* and *Epithalamion* (1595); *The Faerie Queene*, Books I-VI, 2 vol. (1596); *Fowre Hymnes* (1596); *Prothalamion* (1596); *The Faerie Queene*, Books I-VI with the fragment of Book VII known as "Mutabilitie Cantos" (1609).

TRANSLATIONS: "A Theatre, Wherein Be Represented the Miseries and Calamities That Follow the Voluptuous Worldlings," ed. by Jan van der Noodt, containing anonymous translations in verse by Spenser of poems from the French (1569).

PROSE: *Three Proper, and Wittie, Familiar Letters*, by Gabriel Harvey and Spenser (1580); *A View of the Present State of Ireland* (1633).

BIBLIOGRAPHY. WALDO F. MCNEIR and FOSTER PROVOST, *Annotated Bibliography of Edmund Spenser, 1937–1960* (1962).

Editions: E.A. GREENLAW *et al.* (eds.), *The Works of Edmund Spenser: A Variorum Edition*, 10 vol. (1932–49), complete, detailed information; R.E. NEIL DODGE (ed.), *The Complete Poetical Works of Edmund Spenser* (1908), and J.C. SMITH and E. DE SELINCOURT (eds.), *The Poetical Works of Edmund Spenser* (1912), two convenient one-volume editions; C.G. OSGOOD, *A Concordance to the Poems of Edmund Spenser* (1915).

Biography: ALEXANDER C. JUDSON, *The Life of Edmund Spenser* (1945), the authoritative biography.

Critical studies: WILLIAM NELSON, *The Poetry of Edmund Spenser* (1963); W.L. RENWICK, *Edmund Spenser: An Essay on Renaissance Poetry* (1925); ROBERT ELLRODT, *Neoplatonism in the Poetry of Spenser* (1960); VIRGIL K. WHITAKER, *The Religious Basis of Spenser's Thought* (1950); W.R. MUELLER (ed.), *Spenser's Critics: Changing Currents in Literary Taste* (1959); PAUL J. ALPERS, *The Poetry of the Faerie Queene* (1967); JANE APTEKAR, *Icons of Justice: Iconography and Thematic Imagery in Book V of the Faerie Queene* (1969); HARRY BERGER, *The Allegorical Temper: Vision and Reality in Book II of Spenser's Faerie Queene* (1957); DONALD CHENEY, *Spenser's Image of Nature: Wild Man and Shepherd in the Faerie Queene* (1966) on Book VI; MAURICE EVANS, *Spenser's Anatomy of Heroism: A Commentary on the Faerie Queene* (1970); ALASTAIR FOWLER, *Spenser and the Numbers of Time* (1964), on numerical structure of the *Faerie Queene*; A.C. HAMILTON, *The Structure of Allegory in the Faerie Queene* (1961); GRAHAM G. HOUGH, *A Preface to the Faerie Queene* (1962); C.S. LEWIS, *Spenser's Images of Life*, ed. by ALASTAIR FOWLER (1967); THOMAS P. ROCHE, *The Kindly Flame: A Study of the Third and Fourth Books of Spenser's Faerie Queene* (1964); KATHLEEN WILLIAMS, *Spenser's "Faerie Queene"* (U.S. title, *Spenser's World of Glass*, 1966); PAUL E. MCLANE, *Spenser's Shepheardes Calender: A Study in Elizabethan Allegory* (1961); A. KENT HIEATT, *Short Time's Endless Monument: The Symbolism of the Numbers in Edmund Spenser's Epithalamion* (1960).

(A.K.H.)

Speransky, Mikhail Mikhaylovich

Mikhail Mikhaylovich Speransky, a leading Russian statesman under the emperors Alexander I and Nicholas I, compiled the first complete collection of Russian law. He was born the son of the village priest at Cherkutino in central Russia on January 12 (January 1, old style),

1772. Known in his childhood simply as Misha, son of Mikhail the priest, he was sent at the age of 12 to the ecclesiastical seminary in Vladimir, the provincial capital. The deficiency of lacking a surname was overcome by an imaginative uncle, who dubbed him Speransky, a Russified form of the Latin word for hope. The boy soon distinguished himself by his ability to analyze problems and to express his thoughts with grace and clarity, but he already displayed an aloofness that emphasized his consciousness of his intellectual superiority yet cloaked his very real desire to feel the affection of those whom he respected, a quality that was to be a handicap in his later official career.

Novosti Press Agency



Speransky, portrait by an unknown artist.

As a priest's son, he was sent at government expense to the Main Seminary newly founded in St. Petersburg. On completion of the course, he should have returned to his native diocese as a teacher. But a practice sermon so pleased the Metropolitan of St. Petersburg that the Synod granted permission to retain Speransky as a teacher of mathematics in the Main Seminary. Speransky resisted the urging of the Metropolitan that he take monastic vows, a step that would have opened to him the possibility of rising to the highest offices in the Church. Despite his refusal, he was, in 1795, appointed instructor of philosophy and prefect of the seminary.

Secretary to Prince Kurakin

At this point, Speransky's future prospects were radically changed. Prince A.B. Kurakin took him into his household as secretary. Here he deepened his knowledge of the thought of the French Enlightenment and was introduced to the Idealist philosophy of Immanuel Kant. On the accession of Emperor Paul I (1796), Kurakin was appointed procurator general of the Senate, a post as close as possible in the Russian system of that time to that of prime minister. He was, thus, powerful enough to secure Speransky's release from his priestly status and thus enable him to enter government service. Speransky was pushed rapidly upward through the lower bureaucratic grades; by the end of 1798, still not 27 years old, he had already risen high enough in the Table of Ranks to be entitled to enjoy, on a hereditary basis, all the privileges of "the most ancient nobility."

In the same year, Speransky met an English girl whose widowed mother had come to Russia as a governess. He became so enamored of her that, although she knew no Russian and he understood no English, a courtship in fractured French led to their marriage. In the following year a daughter was born, but the mother, suffering from tuberculosis, died a few months later. Speransky, completely shattered, disappeared for a time. He never remarried but immersed himself wholly in his work. When Kurakin suddenly fell from favour, Speransky's tact, his evident ability, and his industry enabled him to continue his career. Under Paul's successor, Alexander I, he was assigned to ever more responsible positions, at first in the new Ministry of the Interior, where he gained invaluable experience in drafting legislation and was the prime mover in founding *Severnaya pochta* or *Novaya Sankt-Petersburgskaya gazeta*, Russia's first official newspaper. In 1807 he became intimately associated with the Emperor himself, as his administrative secretary and assistant. In 1808 he accompanied Alexander to his meeting with Napoleon, who described him as "the only clear head in Russia." Though he proved to be not yet able to cope successfully with the task of codifying the country's laws, he reorganized the seminaries and secured the establishment of the first Russian lycée (state secondary school).

Secretary to the Emperor

In 1809 he laid the basis for his own downfall by two measures that outraged the bureaucratic nobility: one required that holders of court titles perform actual service to the state; the other required that all officials must pass examinations in order to be promoted at various stages of their careers. The angry nobles began to refer contemptuously to him as the *popovich* ("priest's son"). It was in this year, also, that Speransky proposed his new "constitution" (the Plan of 1809). Well aware that Alexander wanted no tampering with the essence of the autocracy or with its basis in serfdom, Speransky prepared complicated plans for dividing the population into three classes with varying degrees of political and civil rights and for creation of elective assemblies, the dumas, and an appointive State Council. The latter was set up on January 1, 1810, but the dumas, innocuous though they would have been, remained on paper. In these years (1807–12) when he had the Emperor's confidence, Speransky was responsible for a number of financial and administrative reforms designed not to change the essence of the state structure but to improve its functioning. His obvious pro-French leanings, however, added to the hostility of the nobles, whose pocketbooks had suffered by Russian participation in the Continental System, the systematic economic warfare employed by Napoleon against England.

Speransky's aloof personality and his continued association with persons inferior to him in social standing had prevented him from making friends among men with political prestige. He was thus left defenseless against his high-placed enemies at court, including the emperor's sister, Catharine of Oldenburg. In 1811 the renowned historian N.M. Karamzin attacked him in his well-known memoir, *Of Old and New Russia*. In March 1812, Speransky was summarily dismissed. Returning to his home at midnight, he found a police carriage waiting at his door. Without even taking leave of his daughter, the fallen minister started on the long journey to exile in Nizhny-Novgorod, whence he was soon transferred to the even more distant Perm, in the Urals.

Exile

Two years later he was permitted to return to his estate near Novgorod, but it was not until 1816 and only after he had stooped to appeal to his successor in Alexander's favour, Count A.A. Arakcheyev (whom the poet Pushkin contrasted with Speransky as Alexander's "evil genius"), that he was permitted to re-enter state service—though only as provincial governor in remote Penza. In 1819, however, he was promoted to be governor general of Siberia, where he effected significant administrative reforms. In 1821 he was summoned to St. Petersburg and appointed a member of the State Council, in which he was too prudent to advocate further reforms, lest he again irritate his master.

Under Alexander's successor, Nicholas I, Speransky's great talents were again utilized, first as a member of the special tribunal that tried and sentenced the Decembrists, a group of officers who staged a liberal revolt on Nicholas' accession in December 1825. Here he again demonstrated his ability to read an emperor's mind; it was he who drafted the letter to the court that secured a significant reduction of the sentences the tribunal had imposed. In the same year he became, in effect, the head of the Second Division of the Emperor's personal chancellery. Still an efficient workhorse, he took part in the labours

Work for Nicholas I

of Nicholas' secret committees for study of the peasant problem. His major achievement, however, was the publication, in 1830, of the first *Complete Collection of the Laws of the Russian Empire* (*Polnoye sobraniye zakonov Rossiyskoy imperii*), 45 volumes. On the basis of this compilation, which began with the *Code* (*Sobornoye ulozheniye*) of 1649, he supervised preparation of a *Digest of the Laws* (*Svod zakonov Rossiyskoy imperii*), 15 volumes (1832–39). In 1837 he was awarded the high Russian decoration, the highest grade of the Order of Andrew the First-Called, and, in January 1839, was accorded the title of count. He died a few weeks later, on February 23 (February 11, O.S.), in St. Petersburg.

BIBLIOGRAPHY. The best biography of Speransky is the excellent work of MARC RAEFF, *Michael Speransky: Statesman of Imperial Russia, 1772–1839* (1957). See also his *Siberia and the Reforms of 1822* (1956). For Speransky's own early writings (to 1809), see *M.M. Speranskii: Proiekty i zapiski*, ed. by S.N. VALKA (1961).

(J.D.C.)

Sphenisciformes

The penguins are a group of flightless seabirds constituting a particularly distinct and homogeneous order, Sphenisciformes; the one family, Spheniscidae, is limited in distribution to the Southern Hemisphere. Although a few penguins inhabit temperate regions and one, the Galapagos penguin (*Spheniscus mendiculus*), lives at the Equator, the majority of the 18 species breed on islands in subantarctic waters.

Drawing by Christian D. Olsen

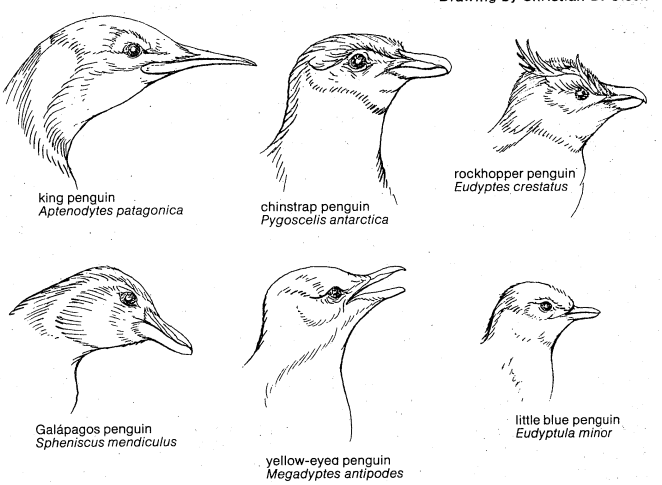


Figure 1: Heads of representative sphenisciformes.

The stocky, short-legged appearance of penguins has endeared them to the popular mind. They range from about 35 centimetres (14 inches) in height and approximately one kilogram (about two pounds) in weight, in the little blue, or fairy, penguin (*Eudyptula minor*), to 115 centimetres (45 inches) and 25 to 40 kilograms (55 to 90 pounds) in the emperor penguin (*Aptenodytes forsteri*). Most are black on the back and white below, often with lines of black across the upper breast or spots of white on the head. Colour is rare, being limited to red or yellow irises of the eye in some species; red beaks or feet in a few; yellow brow tufts in the three species of *Eudyptes*; and orange and yellow on the head, neck, and breast in the two species of *Aptenodytes*.

The total populations of some species, such as the emperor penguin, are estimated in hundreds of thousands, but those of most species of smaller penguins certainly run to several millions. The largest breeding colonies are on the islands between 50° south latitude and Antarctica. These immense colonies, some of which contain hundreds of thousands of individuals, represent a large potential food resource, but the economic importance of penguins is negligible. Nineteenth-century whalers and seal hunters visited some colonies for meat and eggs, and a penguin oil industry once took large numbers of birds; by the early 20th century, however, this exploitation was no longer profitable, and most colonies were left alone or actively protected. Some species are now increasing in numbers, apparently as a result of the drastic reduction of Antarctic whales, which compete with penguins for the krill (minute crustaceans) on which both feed.

Natural history. *Reproduction.* Many features of the life cycle vary with the body size of the species and its geographical distribution; the chronology of breeding may also vary within a species in relation to latitude. Certain species, such as the Cape or jackass penguin (*Spheniscus demersus*), probably other members of this genus, and the little penguin, breed twice a year, but the majority of species breed only once each year and the king penguin (*Aptenodytes patagonica*) twice in three years. One egg is laid by the emperor and king penguins, and two, or occasionally three, by all others. Most penguins begin breeding in the austral spring or summer; king penguins are on a 14- to 18-month cycle, and the timing of an individual pair depends on the success or failure of the previous breeding attempt. Some populations of the gentoo penguin (*Pygoscelis papua*) also breed in winter. The breeding of the emperor penguin begins in autumn, apparently timed so that the long developmental period will produce the young in midsummer, when their chances of survival are greatest.

The gentoo, which has a circumpolar distribution, is notable for the lack of synchrony among populations, but otherwise its breeding schedule is essentially typical of most penguins. In the Crozet Islands, off southern Africa, for example, egg laying takes place in July. The two eggs are incubated for 35 or 36 days, and the rearing of the chick takes two months. The last immature birds go to sea in January, leaving the adults to molt in February.

Many types of visual and vocal displays are employed between the arrival of the birds at the colony and their departure. Courtship calls are used during pairing and to a lesser degree during the succeeding phases of breeding. There is marked vocal dimorphism between sexes in the emperor penguin and the king penguin, and less marked dimorphism in some other species. Upon arrival at the colony each bird returns to the nest that it left the previous year and generally rejoins its mate of the previous year, unless the death of the latter forces it to choose another partner. This applies even to the emperor penguin, which is capable of finding its mate despite the absence of a nest and the large size of the colony.

Types of display

The displays that occur with the reassembly of the colony and the finding of mates, as well as those preceding copulation, are quite similar among the majority of species, but the accompanying vocalizations are more diverse. Various species have been described as braying, trumpeting, croaking, cackling, and cooing. The behaviour of experienced older birds is more elaborate and more effective than that of younger individuals. Adélie penguins (*Pygoscelis adeliae*) may return to the reproductive colony from their third year onward but do not breed successfully until their fifth or sixth year.

Incubation of eggs is performed by both sexes in all species except the emperor penguin, in which it is done by the male alone. It begins immediately following the laying of the single egg in the emperor and king penguins and after the laying of the second egg in other penguins. With the advent of incubation, the bustle and myriad cries that characterized mating give way to quiet and inactivity. Faulty incubation behaviour by inexperienced birds frequently results in the abandonment or breakage of eggs. The mortality (eggs and chicks) is very important at the egg stage, varying from year to year depending on climatic conditions, the percentage of young birds in the reproductive population, and the pressure of predation. In general, the mortality (eggs and chicks) is from 40 to 80 percent of the eggs laid. In coastal colonies predators include, in order of importance: skuas (*Catharacta*); sheathbills (*Chionis*); and the giant petrel (*Macronectes giganteus*). On the Australian, African, and South American continents, the nocturnal habits of certain penguins and the fact that they nest in burrows substantially limit predation, which is mostly by gulls and man.

Following egg laying, the female usually departs for the sea to feed, returning to relieve her mate after about ten to 20 days, depending on the species; thereafter, they alternate in periods of a week or two. The female emperor penguin, however, must often walk 80 to 160 kilometres (50 to 100 miles) from colony to the sea and does not return until the end of the incubation period. During the 64-day incubation period, which extends through the height of the Antarctic winter, the male emperor penguin incubates the egg, holding it on his feet and living on stored fat reserves. During violent winter storms, members of the colony gather for mutual protection from wind and cold in tightly packed crowds called huddles.

Emergence from the shell takes 24 to 48 hours, during which the brooding parent is particularly irritable. The chick shows feeding behaviour immediately on hatching, receiving a regurgitated "soup" of crustaceans or fish by inserting its bill into the open mouth of the parent. During its early days the young bird is sheltered under the body of one of its parents, who take turns foraging and brooding. Growing larger, the young bird remains at a parent's side, although able to maintain its body heat and move about alone. In surface-nesting species, following the individual rearing phase in the nest or (in the emperor and king penguins) on the feet of the parent, the chick joins 100 or more of its contemporaries in a nursery

Jean Prevost

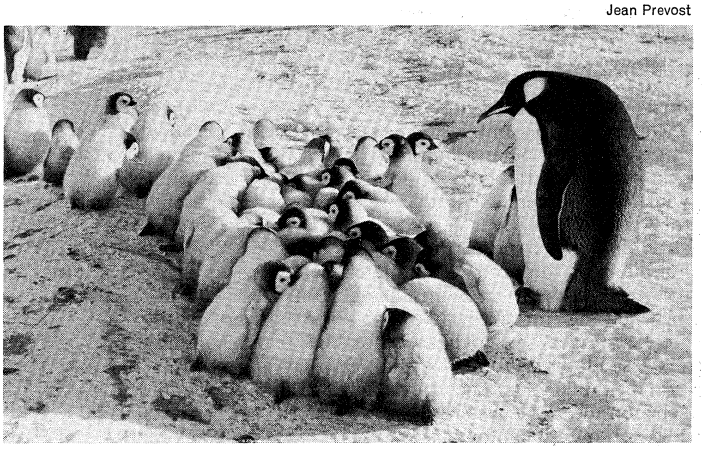


Figure 2: *Crèche of emperor penguins (Aptenodytes forsteri).* To guard against predators, the young are herded into groups attended by adults while the parents fish for food.

Crèche behaviour

group, or crèche, sometimes guarded by a few adults, while both its parents forage at sea. Upon returning with food, the parent calls its chick from the crèche and is able to distinguish it from other chicks (which frequently also respond) by voice and appearance. During the breeding season the number of "unemployed" adults in the colony increases with the addition of those who have lost eggs or chicks. In emperor penguin colonies, these unemployed birds often interfere with parents that have young and cause increased mortality. During the crèche stage the fuzzy down that has covered the chick since hatching is replaced by a coat of short stiff feathers, like those of the adult but usually somewhat different in colour. Once this molt is complete, the juvenile leaves the colony to seek its own food at sea.

The period of growth of the young bird from hatching to complete independence varies from two months, in the smallest species of the genus *Eudyptula*, to 5½ months in the emperor and 12 to 14 months in the king penguin.

Molt. Adult penguins molt all of their feathers once a year, following the breeding period. While in molt the bird is unable to enter the water and instead withdraws to a communal molting site usually situated in a sheltered area away from the colony. The duration of the molt varies from about two weeks in small species to more than a month in the larger ones.

Locomotion and orientation. Penguins are adapted for rapid locomotion in water, in which the wings, or flippers, are used for propulsion; the birds "fly" under water. When moving at high speed they frequently leave the water in leaps that may carry them a metre or more through the air; during this time they breathe.

On land, penguins are much more awkward, and their rocking walk is highly amusing to watch. Despite their short legs, however, penguins can run with surprising speed. Some species, such as the rockhopper (*Eudyptes chrysocome*) and Adélie penguins, move among rocks with agility, using the flippers for balance. When moving rapidly on snow or ice, many penguins "toboggan," sliding on the belly as they propel themselves by the feet and flippers. The flippers, along with the beak, are the prime weapons in defense and attack.

"Tobogganing"

Scientists have long wondered how penguins are able to find their way back to their colonies from far out at sea, where currents may have carried them great distances, and how they are able to direct themselves correctly on land, in the absence of clear-cut landmarks. Recent studies have shown that penguins transported into the interior of Antarctica are able to find their way back to the ocean by using the sun as a directional aid. It is probable that the same means of orientation is used at sea, and that upon approaching the coast they are able to recognize features of the shoreline and the ocean bottom.

Food habits. The type of food utilized varies with the species, the geographical region, and the time of year. Most of the smaller southern penguins feed primarily upon krill (crustaceans of the family Euphausidae), which attain high densities in the rich, well oxygenated Antarctic waters. Cephalopods (squid and cuttlefish) and small fishes may form substantial fractions of the food and, in a few, such as the Cape penguin, comprise the basic elements of the diet. The total weight (biomass) of food consumed by a large penguin colony is prodigious, often exceeding several tons per day.

Natural enemies. The principal enemies of penguins at sea are the leopard seal, or sea leopard (*Hydrurga leptonyx*), and the killer whale or grampus (*Orcinus orca*). The eared seals (*Otariidae*) of the genera *Arctocephalus* and *Neophoca* also take penguins in the Australian–New Zealand and subantarctic regions.

Form and function. The penguins are highly specialized for their flightless aquatic existence. The feet are located much farther back than those of other birds, with the result that the bird carries itself mostly upright; its walk can thus be described as plantigrade (*i.e.*, on the soles). It rests on the ground on a sole, comprising the whole foot, instead of one formed of the toes, as in other birds. The most notable character of the group is the transformation of the forelimb into a swimming member, accompanied by a body morphology particularly adapted

From R.W. Storer, "Diving Birds," International Ornithological Congress (1958)

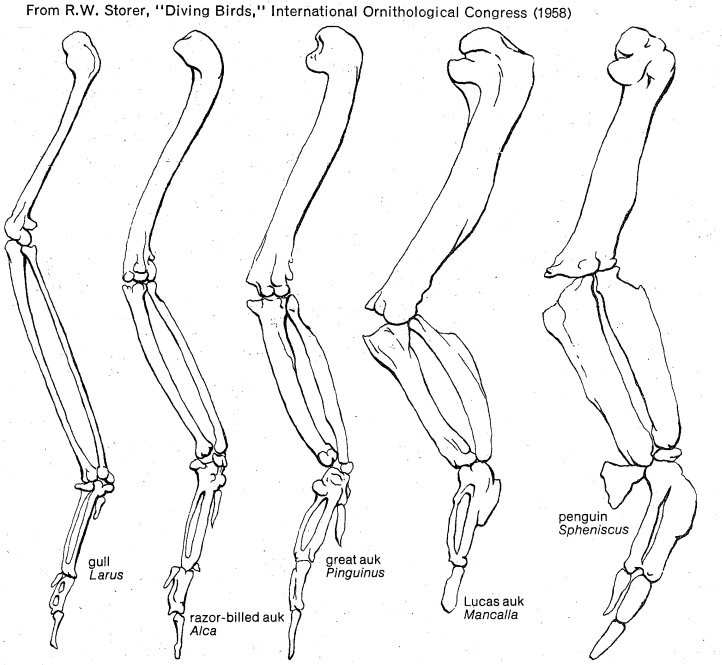


Figure 3: Evolution of the wing in diving birds.

to movement in a liquid medium. The thoracic (rib) cage is well developed, and the sternum bears a pronounced keel for the attachment of the pectoral muscles, which move the flippers. The flipper has the same skeletal base as the wing of flying birds but with its elements shortened and flattened, producing a relatively rigid paddle-like limb covered with very short feathers, which makes an ideal organ for rapid propulsion. The body plumage likewise consists of very short feathers, which minimize friction and turbulence; by their density and the layer of air which they retain, the feathers provide almost ideal insulation of the body.

Structure of the forelimb, or flipper

The insulation of the bird's body is particularly important for Antarctic species that move in water that is always below 0° C (32° F). The cooling power of seawater at −1.9° C (28.6° F) is equal to that of a temperature of −20° C (−4° F) with a wind of 110 kilometres (70 miles) per hour. The velocity of movement of the birds in the water serves to increase this heat loss. In fact, the skin is insulated by a layer of air trapped under the plumage, and the only bare skin in direct contact with the water is that of the feet. In the case of the emperor penguin on land in winter, the feet may also be in close and permanent contact with ice. The skin temperature is then in the neighbourhood of 0° C, and snow does not melt upon contact. This low temperature is made possible by remarkable anatomical arrangements in the lower limb, in which closely adjacent arteries and veins form a system of heat exchange between opposing flows; in the upper member a remarkable heat-exchange network permits the cooled returning blood to absorb heat from the outflowing blood, providing maximum economy of heat consistent with the functioning and survival of the foot.

Like the petrels, the penguins have supraorbital glands or salt glands, which permit the ingestion of hyperosmotic substances (*i.e.*, with more salt than the body tissues); excess chloride is excreted in the form of a solution the concentration of which is greater than that of seawater. These glands are functional in the young penguin chick, which consumes food of marine origin from its first day of life.

Recent research has shown that the species most isolated geographically, such as the emperor penguin, can be subject to viral diseases of the psittacosis–*Lymphogranuloma* type. Some, such as the Adélie penguin, carry in their bodies trace amounts of pesticides such as DDT, albeit in lower quantities than are found in many birds that live closer to man.

Evolution and classification. *Fossil record.* Evidence from paleontology indicates that the penguins and tubinares (Procellariiformes: albatrosses, shearwaters, and petrels) had a common origin. Both groups are represented by well-defined fossils from the Eocene (about 50,000,000 years ago). During the Tertiary Period (from about 25,000,000 to 65,000,000 years ago) the sphenisciform line produced a number of distinctive side branches, all recognizably penguins. From the taxonomic point of view the radiation (evolutionary differentiation) had reached the level of the subfamily in the Miocene (between 7,000,000 and 26,000,000 years ago), and the diverse subfamilies of that epoch have been replaced in the present-day group by one in which the radiation has reached the level of genera. All of the fossil remains of penguins have been collected within the zone of the present-day distribution of the Sphenisciformes. Some apparently lived in warmer regions than do most of today's penguins. Certain species, such as *Palaeospheniscus patagonicus*, had a height intermediate between the large and the medium-sized modern species; others, such as *Pachydypees*, *Anthropornis*, and *Palaeeudyptes*, reached heights of between 120 to 160 centimetres (about 50 to 60 inches).

Classification. One authority has divided the *Spheniscidae*, fossil and modern, into five subfamilies: *Palaeospheniscinae* and *Paraptenodytinae* of the Patagonian Miocene; *Anthropornithinae* of New Zealand, Patagonia, and Antarctica; *Palaeeudyptinae* of New Zealand and Australia; and finally *Spheniscinae*, which includes all the present-day genera.

BIBLIOGRAPHY. JOHN SPARKS and TONY SOPER, *Penguins* (1967); and BERNARD STONEHOUSE, *Penguins* (1968), are well-illustrated popular books. R.C. MURPHY, *Oceanic Birds of South America*, 2 vol. (1936), is a classic work with extensive information on ten species of penguins. J.T. EMLEN and R.L. PENNEY, "Distance Navigation in the Adélie Penguin," *Ibis*, 106:417–431 (1964), is a short technical paper concerned with long-range movements. R.L. PENNEY, "Territorial and Social Behaviour in the Adélie Penguin," the "Antarctic Research Series," vol. 12, *Antarctic Bird Studies*, pp. 83–131 (1968), reviews many studies of the behaviour of this species. JEAN PREVOST, *Écologie du Manchot empereur, Aptenodytes forsteri Gray* (1961), is a detailed study, in French, of the largest penguin species. Also of interest is L.E. RICHDALE, *Sexual Behavior in Penguins* (1951). G.G. SIMPSON, "Fossil Penguins," *Bull. Am. Mus. Nat. Hist.*, 87:1–99 (1946), is a basic work on the evolution of penguins.

(J.Pr.)

Sphenopsida

The Sphenopsida is generally considered to constitute a class of primitive, spore-bearing, vascular plants. The sole living genus, *Equisetum*, the horsetails or scouring rushes, comprises about 30 species of herbaceous plants and constitutes the order Equisetales. Extinct members, some of which appeared as early as the Devonian Period (beginning 395,000,000 years ago), include many herbaceous plants of the Equisetales, shrubby Hyeniales, vine-like Sphenophyllales, and trees of the order Calamitales.

General features. Sphenopsids, fossil and living, characteristically have whorled leaves and conspicuously jointed stems, which in many cases are also ribbed. When branches occur in *Equisetum*, they alternate with the leaves in position on the stem, an arrangement not known to occur in any other group of living plants. The buds in seed plants usually arise in the angle that the leaf makes with the main stem. Reproductive structures are present in the form of greatly compressed stems called cones, or strobili, which form at the ends of branches. The giant horsetails of the order Calamitales were trees up to one metre (three feet) in diameter and 30 metres (100 feet) in height. Their leaves—like those of extant horsetails—were arranged in spokelike whorls at regular intervals (nodes) along the jointed stems.

Size range

In the Sphenophyllales, an extinct order of climbing sphenopsids, the leaves were wedge shaped, with a repeatedly forking (dichotomous) venation system. The order Hyeniales included shrublike plants with inconspicuous leaves arranged in rather indistinct whorls.

Of the living species of *Equisetum*, most are less than one metre (three feet) tall; they are widely distributed in the Northern Hemisphere. There are reports of specimens of *E. giganteum*, from the American tropics, that attain a height of about ten metres (30 feet), with a stem diameter of only two centimetres (less than an inch); they are apparently supported by tall grasses in their environment. The majority of *Equisetum* species are found in wet or damp habitats, often in shaded locations along streams, ditches, and canals; some species, however, have become adapted to drier and sunnier conditions.

The extant sphenopsids have had little economic importance. The extinct giant forms contributed to the coal beds of the Carboniferous Period (345,000,000–280,000,000 years ago). Living horsetails have been used as scouring agents, their cleaning value being attributed to the abrasive action of the silica-laden walls of certain of their cells. Silica is only one of several minerals that horsetails selectively accumulate in their bodies. Gold is another—up to 4½ ounces per ton of plants—not economically feasible to mine but a certain indication of the availability of such ore deposits in the soil. Horses foraging on stands of *Equisetum* have been known to die from severe intestinal inflammation.

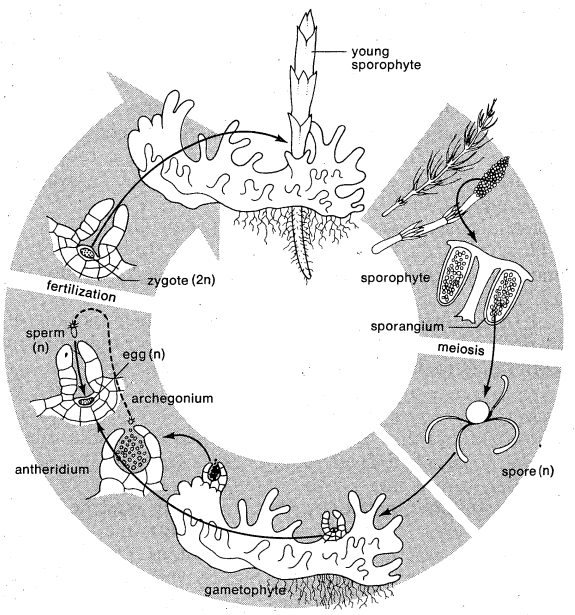
Life cycle. Horsetails, like other vascular plants, display an alternation of generations: an asexual phase, represented by a sporophyte (the horsetail plant); and a sexual phase, the gametophyte, an inconspicuous, delicate, green plant. Each year thousands of gametophytes are probably initiated from spores, but very few survive in nature. Horsetails apparently survived as a small group of very ancient plants mainly by vegetative reproduction

Fragility of gametophytes

rather than by a regular dependence on an alternation of sexual and asexual generations. It is assumed that fossil members had an alternation of generations, although only sporophytes have been described for them.

Some horsetails have terminal cones (strobili) on green aerial branches. Other species, however, have separate, upright, aerial branches for vegetative and for reproductive shoots. In the latter, the strobilate branches appear first, and, after the spores are shed, the green, vegetative shoots develop. The component fertile appendages of the strobilus are called sporangiophores; each consists of a stalk bearing a flattened disk at its apex, on the lower edge of which is found a circle of sporangia, each opening and shedding spores by a longitudinal slit on its inner side. (The Carboniferous, treelike horsetails and their smaller allies are believed to have possessed the most elaborate reproductive strobili known among the vascular plants.)

The spores have four bands (elaters); these elaters, which coil and uncoil in response to changes in moisture content of the air, assist in the dispersal of the spores. Under low light intensity and high humidity, the spores grow into small, flattened, green prothallia (early gametophytes). After a period of development these gametophytes come to resemble miniature green pincushions, being up to three centimetres (slightly over an inch) in diameter. Eggs are produced in structures called archegonia, at the bases of the upright lobes on the small gametophytes, and sperm are produced in structures called antheridia, present on the lobes. The egg is fertilized in the archegonium by a sperm; then the zygote, by continued divisions, forms an embryo within the archegonium. The embryo (young sporophyte) is nourished by the gametophyte until it develops its own shoot and roots. One gametophyte may support two or more young sporopyhtes before it ultimately dies and decays.



Life cycle of *Equisetum*.

Form and function. The shoot system of the sporophyte consists of an underground stem (rhizome) and an aerial stem, both of which show the same basic organization.

The slender, herbaceous stems of *Equisetum* have hollow internodes. The whorled leaves are greatly reduced and are united laterally at each node to form a toothed sheath around the stem. Each leaf has a single, unbranched midvein. Secondary growth, during which girth increases annually, was characteristic of the extinct Calamitales and Sphenophyllales. In *Equisetum* the vascular strands are organized into a cylinder surrounding a large pith cavity. In the majority of species, the cell walls of the outer cell layer (epidermis) are thick and contain silica deposits. Roots arise at nodes of the rhizome. Branch buds also are initiated at nodes, but their subsequent growth is dependent on the growth characteristics of the particular species. In relatively unbranched species, the potential buds remain inhibited.

The sporophyte consists of stem, leaves, and roots. Spores are produced in spore cases (sporangia), which are borne on sporangiophores. The latter are organized into strobili associated with sterile bracts (much reduced leaves)—as in Sphenophyllales and Calamitales—or without them (Hyeniales, Equisetales). Each sporangiophore in *Equisetum* has five to ten sporangia. The entire sporangiophore appears to be a condensed, dichotomous branch system (*i.e.*, dividing into two parts), each sporangium occupying the end of a branch but lying parallel to the stalk of the sporangiophore.

Evolution and classification. Certain Sphenopsida flourished as trees (for example, *Calamites* species) during the coal-forming Carboniferous Period, but the earliest sphenopsids appeared as early as the Devonian. In its fossil history the class is believed to have constituted a larger portion of the flora of the Earth than at the present time.

Equisetum, which may have been present also during the Carboniferous, may be one of the oldest living genera of vascular plants. The evolutionary tendency in this genus has not been noticeably radical. The more primitive species, which now range in South America, Cuba, and Mexico, have perennial, green shoots with green, pointed cones. The advanced species, northward to the Arctic, have annual, green, branched, vegetative shoots and nongreen, unbranched, fertile shoots with brownish cones. Intermediate combinations of these features occur in many species.

The evolutionary trend

Annotated classification. Some botanists recognize five or six different orders in the class; others restrict the class to four orders. There is only one order, Equisetales, that has both living and extinct species; all others comprise extinct sphenopsids. The latter are indicated by a dagger (†) in the listing below.

CLASS SPHENOPSIDA (OR ARTICULATAE)
Extinct and living primitive, seedless, vascular plants with jointed, ribbed stems and whorls of leaves at regular intervals along the stem.

† Order Hyeniales (Protoarticulatae)
Extinct, shrublike plants, with short, forked leaves in whorls. One family: Hyeniaceae.

† Order Pseudoborniales
One family, Pseudoborniaceae, with a single extinct species, *Pseudobornia ursina.*

† Order Sphenophyllales
Extinct, vinelike sphenopsids, with small, wedge-shaped leaves. Two families: Sphenophyllaceae and Cheirostrobaceae.

† Order Calamitales
Extinct tree horsetails in two families: Asterocalamitaceae and Calamitaceae.

Order Equisetales
Herbaceous horsetails and fossil allies with needlelike leaves in whorls along the stem. About 30 extant species in the genus *Equisetum* and several extinct species in the genus *Equisetites*. *Equisetum* is distributed predominantly in the Northern Hemisphere.

Critical appraisal. Whereas the dominant opinion is that the Sphenopsida is a class, also called Articulatae, in reference to the jointed, or articulated, stems, some authorities raise the group to the level of a division, Arthrophyta or Sphenophyta, or a subdivision, retaining the term Sphenopsida. These differences in viewpoint reflect a lack of information basic to the erection of a stable classification. Additional paleobotanical evidence is needed to establish a classification acceptable to most botanists.

BIBLIOGRAPHY. THOMAS E. WEIER, C. RALPH STOCKING, and MICHAEL G. BARBOUR, *Botany: An Introduction to Plant Biology*, 4th ed. (1970); and ROBERT F. SCAGEL *et al.*, *An Evolutionary Survey of the Plant Kingdom* (1965), general textbooks of college-level difficulty; ADRIANCE S. FOSTER and ERNEST M. GIFFORD, JR., *Comparative Morphology of Vascular Plants* (1959); and DAVID W. BIERHORST, *Morphology of Vascular Plants* (1971), textbooks with detailed treatment of

vascular plants together with theory and interpretation; HENRY N. ANDREWS, JR., *Studies in Paleobotany* (1961); and THEODORE DELEVORYAS, *Morphology and Evolution of Fossil Plants* (1962), textbooks with detailed accounts of fossil forms and their relationships with living plants. Journals that publish articles on vascular plant morphology include: *American Journal of Botany* (10/year), *The Botanical Gazette* (quarterly), *Phytomorphology* (quarterly), *Annals of Botany* (5/year), and *Palaeontographica* (1–3/year).

(E.M.G.)

Spices, Herbs, and Flavourings

Spices and herbs are dried parts of various plants cultivated for their aromatic, pungent, or otherwise desirable substances. They consist of rhizomes, bulbs, barks, flower buds, stigmas, fruits, seeds, and leaves. They are commonly spoken of loosely as spices, spice seeds, and herbs. Spices are the highly esteemed, fragrant or pungent plant products of tropical and subtropical regions, the dominant species of the trade including cardamom, cinnamon, cloves, ginger, and pepper. Spice seeds are the tiny aromatic fruits and oily seeds of herbaceous plants including anise, caraway, cumin, fennel, poppy, and sesame. Herbs are the fragrant leaves of such plants as marjoram, mint, rosemary, and thyme.

Spices, spice seeds, and herbs are employed as adjuncts to impart flavour and aroma or piquancy to foods. In the small quantities used to prepare culinary dishes they have little or no nutritive value, but they stimulate the appetite, add zest to food, enhance the taste, and delight the gourmet. For individual spices and herbs see RELATED ENTRIES under SPICES, HERBS, AND FLAVOURINGS in the *Ready Reference and Index*.

HISTORY OF THE SPICE TRADE

The role of spices and herbs in shaping the course of history is a long and romantic story.

The spice trade is one of the oldest known to man. Overland trade routes across Asia antecede recorded history, and merchandise moved over coastal sea routes in both directions between Arabia and southern Cathay (China) in very early times. Cinnamon, cassia, cardamom, ginger, and turmeric were known and employed by Eastern peoples thousands of years ago, and they became important items of commerce early in the evolutionary course of trade.

Ancient spice trade routes

Cinnamon and cassia found their way to the Middle East at least 2,000 years before the Christian Era. There are many references to spices in the Bible, including four to cinnamon and three to cassia. The spice trade in those days was extensive and profitable. Part of the immense yearly income of King Solomon came from "the traffic of the spice merchants" (1 Kings 10:15).

From time immemorial, southern Arabia (Arabia Felix of antiquity) had been a trading centre for frankincense, myrrh, and other fragrant resins and gums. It is logical to conclude that her merchants would seek to add to their store of aromatics and doubtless it was the Arabs who first brought the highly esteemed cinnamon and cassia to the Middle East. Shrewd traders as they were, they artfully withheld the true source of these spices. To satisfy the curious, to protect their market, and to discourage likely competitors, they spread abroad fantastic tales to the effect that cassia grew in shallow lakes guarded by winged animals, that cinnamon grew in deep glens infested with poisonous snakes, and that, in consequence, they obtained these spices only at great risk. Pliny the Elder (AD 23–79) ridicules these stories of "fabulous antiquity" and boldly declares that "all these tales, however, have been evidently invented for the purpose of enhancing the price of these commodities." Strabo (*c.* 64 BC–AD 20) in his *Geography* speaks of the extensive trade in spices carried on by the Arabians, of the wealth of the Arabian spice merchants, and observes that even a camel driver "might attain to some sort of wealth" by trading in spices.

Whatever part the overland trade routes across Asia played, it was mainly by sea that the spice trade grew. By rudimentary navigational skill, and aided by the winds called monsoons, Arabians were making direct sailings before the Christian Era. In the Far East, parallel in time and not less skilled, the Chinese were moving through the waters of the Malay Archipelago and trading in the Spice Islands. Ceylon was a central trading point.

With the advent of the Ptolemies, Egyptian trade with India began and was developed by the Romans who succeeded them. From its early, hesitant start this trade was well developed by the time of Strabo, who says that on a visit to Syene he found that about 120 ships sailed regularly from the Red Sea port of Myos-Hormos to India.

Under the Ptolemies, Alexandria, founded in 332 BC, had become a great seaport and centre of trade. Its revenues from port dues were already enormous when Ptolemy Alexander bequeathed this ancient city to the Romans in 80 BC. Under the Romans Alexandria became the greatest commercial centre of the world and the emporium for the aromatic and pungent spices of India that found their way to the markets of Greece and the Roman Empire.

Roman trade with India was extensive for over three centuries and then began to decline. It had, no doubt, weakened but not broken the Arabian hold on the spice trade. The Roman trade revived in the 5th century but dwindled in the 6th, whereas the Arabian trade endured into and through the Middle Ages.

By the 6th century Ceylon had become a notable trading centre to which came, according to Cosmas Indicopleustes of Alexandria, the "richest merchandise of the Indies, and all sorts of perfumes and spices."

At about the same time, Venice came into existence. By the 10th century it was beginning to prosper in the trade of the Levant; by the early part of the 13th century it enjoyed a monopoly of the trade of the Near East; and by the 15th century it was a formidable power in Europe. Part of Venice's great wealth came from trading in the spices of India, Ceylon, the East Indies, and Cathay, which it obtained in Alexandria and sold to northern and western European buyer-distributors at exorbitant prices.

Role of spices in the voyages of discovery

The Europeans knew full well the origin of the spices reaching Alexandria and, unable to break the hold of Venice, determined in the last third of the 15th century to build ships and venture abroad in search of a route to the spice-producing countries. So began the famed voyages of discovery. The Portuguese were first in the race, and the first to bring spices from India to Europe by way of the Cape of Good Hope.

Bartholomeu Dias doubled the Cape of Good Hope in 1488; Vasco da Gama reached Calicut, India, in 1498; and Pedro Álvares Cabral brought spices to Lisbon in 1501.

In 1492 Christopher Columbus sailed under the flag of Spain, and in 1497 John Cabot sailed for England; but both failed to find the fabulous spice lands.

Columbus had read, probably, that Eratosthenes of Cyrene (about 250 BC) had calculated the circumference of the world and that Strabo had observed that India might easily be reached from Spain, "still keeping the same parallel were it not for the extent of the Atlantic." Columbus thought, no doubt, that by sailing west he might reach the East before the Portuguese. But the first land he reached was not that of the fabulous Orient but a small island, which he named San Salvador. He thought he had arrived in the Indies "where the spices grow" (hence the name West Indies for the Caribbean islands he discovered). His four voyages resulted in no significant contribution to spices beyond his finding that the natives used capsicums, which they called *aji*, to season their food.

John Cabot fetched up on the barren coast of Cape Breton Island and took possession of the land in the name of Henry VII. He, too, thought he had arrived in the land where the spices grow, but he found none at all.

Ferdinand Magellan took up the quest for Spain in 1519. He discovered the straits that bear his name, and although he was killed in the Philippines en route, he is credited with the first circumnavigation of the world. Of the five vessels that set out from San Lucar under his command only one, "Victoria," returned to Spain, but triumphantly, laden as she was with the cloves of Tidore.

In 1577 Sir Francis Drake began his adventurous voy-

age around the world by way of Magellan's straits and the Spice Islands and brought the "Golden Hind" heavily laden with the cloves of Ternate and other treasures into Plymouth harbour in 1580.

For Holland, a fleet under the command of Cornelis de Houtman sailed for the Spice Islands in 1595; another, commanded by Jacob van Neck, put to sea in 1598; and their ships returned home with rich cargoes of cloves, mace, nutmegs, and pepper.

The market in Europe for spices was considerable; the prices were high and the profits great. The early success of the Dutch gave them a temporary monopoly on pepper, forcing the price of that spice in England to rise from two shillings and eightpence to eight shillings per pound in 1599. The London merchants acted quickly, and on the formation of the East India Company under charter of Queen Elizabeth, December 31, 1600, Capt. James Lancaster, later Sir James, sailed for the East Indies. His ships returned to England in September 1603 with 1,000,000 pounds of pepper, which had cost him less than sixpence a pound in Sumatra.

East India Company

In 1608 a cargo of cloves bought in the Spice Islands for £2,948 realized £36,287 in England.

In 1602 the Dutch East India Company came into existence by authority of the Estates-General of the Netherlands. In 1664 the French East India Company was organized by state authorization under Louis XIV. Other European nations granted charters to East India companies with varying success.

From the time in the latter part of the 15th century when the Portuguese put to sea in search of a southern route to India, European powers vied with one another in the quest for Eastern spices. Numerous vessels set out in search of spice cargoes. There followed struggles and conquests to gain advantage and monopolistic control of the trade. For over 100 years Portugal was the dominant power, then yielding to English and Dutch enterprise and conquest, and this to end in the 19th century with British interests firmly rooted in India and Ceylon and the Dutch in control over the greater part of the East Indies.

For mariners it was an age of adventure, risk, hardship, disease, and death; for nations an age of struggle, defeat, or conquest; an age of acquisition of new, near-primitive lands and of colonizing and gaining dominion over civilized foreign territories. For European commercial interests it was an age of rewarding success, which broke the monopoly of Venice, overcame the Muslim domination of the spice trade that extended from the Red Sea and the Persian Gulf to the China Sea, created a voluminous trade in a great variety of merchandise between Europe and the Far East, and opened up a New World.

The United States entered the spice trade at the close of the 18th century. Capt. Jonathan Carnes is credited with founding the pepper trade with the East Indies. Financed by Salem interests, he put to sea in the brigantine "Rajah," July 20, 1795, and returned in the following year with a cargo of 158,544 pounds of pepper, which sold at 37 cents per pound and brought a handsome profit of 700 percent.

This historic voyage began a lucrative pepper trade with Sumatra that flourished for 69 years, bringing pepper to Salem, Boston, and New York. It is recorded that between 1795 and 1845 some 200 merchantmen laden with the pepper of the East Indies entered Salem harbour.

By the 20th century the world had come to know an orderly and stable pattern of trade in spices, and London had become its centre. Today, New York has taken the place of London.

USES, ANCIENT AND MODERN

The most notable uses of spices and herbs in very early times were in medicine, in the making of holy oils and unguents, and as aphrodisiacs. Priests employed them in worship, incantations, magical rites, and rituals.

Beliefs in the healing powers of spices

Ancient herbals, including those of Cathay, Sumer, Assyria, Egypt, Greece, and Rome, testify to the use of spices and herbs in the treatment of disease. Hippocrates, Galen, and Pedanius Dioscorides, among others, employed them. In the first century of the Christian Era Pliny the Elder, in his *Natural History*, extols at length the efficacy and healing powers of spices and herbs in the treatment of just about every ailment known in his day. Such virtues, tempered and moderated, filtered down into the Middle Ages and early modern times.

It is not known when man first used spices and herbs in food. Sesame seems to have been known and employed as food, for making wine, and for its oil from time immemorial. Garlic and onions were employed as part of the diet in very early times. Certainly by the time of the ancient Greeks and Romans many spices and herbs had come into use to flavour food and beverages.

Herodotus (*c.* 484–*c.* 425 BC) describes the food of peoples he had met or heard about and relates that some ate fish and meats raw, sun-dried, salted, or pickled. Dried, salted, and pickled meats were staple items of food in many lands centuries after his time, and it requires no effort to imagine the welcome change the gratifying flavour of spices and herbs brought to foods, not only to opulent Greeks and Romans but to the affluent of society everywhere long after them.

Only the wealthy could indulge the use of imported aromatic spices and this was true, apparently, all over the world. Marco Polo observed, in 13th-century Cathay, that the higher class people ate meat that had been preserved in several of their spices, but the poor had to be content with meat steeped in garlic juice.

In Europe, in the course of time, knowledge slowly spread of the use of spices and herbs to aid in the preservation of food by retarding or preventing rancidity or other deterioration caused by oxidation and to flavour dishes. By medieval times large quantities of culinary herbs were in use. Eastern spices were beyond the purse of the greater number of people, but with the ascendency of the western European nations in the Oriental spice trade these conditions gradually changed, and the aromatic and pungent spices finally came into general use by rich and poor alike.

Modern uses of spices, spice seeds, and herbs are legion. There are few culinary recipes that do not include them, and their judicious use brings a delectable, distinctive aroma and taste to a host of dishes (see Table 1).

In the food industry they are employed in the preparation of numerous products including processed meats, fancy sausages, sauces and vinegars, prepared mustard, pickles, chutneys, preserves, mayonnaise, salad dressings, biscuits, cookies, cakes, confections, and beverages.

Their essential oils and oleoresins are the basis of a number of spice flavourings and seasonings employed in food manufacturing, where oil spices are preferred to the whole or ground spices for the preparation of certain products.

Spice and herb liqueurs

Spices and herbs, or their oils where processing temperature permits, go into the preparation of a number of liqueurs including absinthe, anisette, benedictine, crème de menthe, curaçao, and kümmel.

Their essential oils are the flavouring components of a number of extracts; and they are employed in perfumery, cosmetics, toilet preparations, hair oils, toothpastes, toilet soaps, and tobacco.

In medicine the spices and herbs have not entirely lost their reputation. In India and other Asiatic countries their curative virtues enjoy respect. They still have a place, though limited, in Western medicine. Present-day herbalists extol the efficacies of some of the spices, spice seeds, and herbs in the treatment of certain ailments.

MODERN TRADE AND PRODUCTION

In the passage of time, by intentional transplanting and seeding, and by natural dissemination, many of the plants from which we obtain spices, spice seeds, or herbs have become established in lands near to, or far from, their natural habitat. Ginger, native to southeastern Asia, is now cultivated in many countries across the world from East to West; capsicum spices, native to Central America, enjoy almost universal production; pepper, native to and an export spice of southern India, is also cultivated for export in Indonesia, Cambodia, Sri Lanka (Ceylon), Malagasy Republic, Malaysia, Thailand, and Brazil; carda-

Table 1: Uses of Spices and Herbs With Foods

	beverages	breads, rolls	cakes, cookies, biscuits	cheese dishes	chutneys	desserts	dressings	egg dishes	fish and seafood	fruit dishes	gravies	meat dishes	pickles	poultry	salads	sauces and relishes	soups	vegetable dishes
Allspice			+		+	+	+						+			+		
Aniseed	+	+	+										+					
Basil, sweet								+	+			+			+		+	+
Caraway seed		+	+	+								+		+	+			+
Cardamom		+	+			+				+			+					
Cayenne pepper					+			+	+			+				+		
Celery salt								+	+			+					+	
Celery seed							+		+				+		+			+
Chervil							+	+	+			+		+	+	+	+	
Chili peppers					+								+			+		
Chili powder								+	+		+	+				+	+	
Cinnamon	+	+	+			+				+			+					
Cloves			+		+	+	+			+		+	+					
Coriander seed		+	+				+					+	+	+	+			
Cuminseed				+			+	+				+		+			+	
Curry powder								+	+			+		+		+		+
Dillseed							+				+	+	+		+	+	+	+
Fennel seed	+		+			+			+				+					
Fenugreek seed					+													
Garlic powder							+					+						
Garlic salt								+				+						
Ginger		+	+		+	+				+		+	+	+				
Laurel (bay) leaves									+			+	+			+	+	
Mace			+			+				+						+		
Marjoram, sweet								+	+			+		+	+	+	+	
Mint	+								+			+				+	+	+
Mixed pickling spice													+					
Mustard					+		+		+		+	+	+			+		
Nutmeg			+			+				+						+		+
Onion powder							+					+			+	+	+	
Onion salt												+			+	+	+	
Origanum (oregano)				+			+	+	+		+	+			+	+	+	+
Paprika							+	+	+		+	+						+
Parsley									+			+		+	+	+	+	+
Pepper							+	+	+		+	+	+	+	+	+	+	+
Poppy seed		+	+				+								+			+
Red pepper							+	+	+			+				+		
Rosemary									+			+		+			+	+
Saffron			+															
Sage									+			+		+				
Savory							+	+	+			+		+	+	+		
Sesame seed		+	+															
Tarragon								+				+		+	+	+		+
Thyme							+					+		+		+	+	+
Turmeric				+			+		+				+			+	+	

mom and sesame, products of the Old World, now flourish in the New World. Today spices, spice seeds, and herbs come from many parts of the world (see Table 2). This has led to increased production, but not overproduction; the consumption of spices and herbs throughout the world is considerable.

Variations of quality have resulted because of differences in altitude, temperature, soil composition, precipitation, and other environmental factors and because of differences in methods of harvesting, drying, processing, and storing in the various regions.

Spices, spice seeds, and herbs are purchased directly from producers or, more commonly, through brokers specializing in the trade and operating, usually, in major seaports such as New York, London, Hamburg, and Amsterdam.

Organizations or associations exist to promote, facilitate, and regulate trading: the Spices Export Promotion Council, India; Fachverband der Gewürzindustrie c.V., West Germany; and the American Spice Trade Association, United States.

Importation and sale. Spices and herbs destined for ports overseas are transported in general cargo vessels. On arrival at their destination, these products are subject to examination for adulteration and contamination by port health authorities acting under the imported food laws of their respective countries. For example, spices, spice seeds, and herbs imported into the United States are subject to sampling and examination by the Food and Drug Administration; spice lots found to be contaminated are reconditioned under their supervision.

Inspection at ports of entry

Following inspection, spices and herbs imported for or by merchants and manufacturers located in the port of entry are cleared through customs on payment of duty, if any (a procedure usually carried out by customs brokers), and picked up at the pier. Those en route to other areas are shipped in bond to their destination, where they are similarly cleared and then picked up at the bonded warehouse by the merchants concerned.

Spices, spice seeds, and herbs are usually imported whole, some to be sold whole, most to be further processed by grinding. Spice grinding is a specialized business requiring ability to identify the various spices, spice seeds, and herbs; to assess their quality; and to know their structure and principal chemical components.

Bark and root (rhizome) spices must be reduced to a size suitable for introduction into the grinding machine. This is usually accomplished with the aid of a knife-cutter comminuting machine.

Grinding machines commonly in use in the United States include pulverizers, roller mills, hammer mills, and comminuters.

The ground and sifted spices are mixed to uniformity usually in a ribbon blender or mixer. This machine also prepares spice and herb mixtures (see below *Blends*).

Excessive heat must be avoided in the grinding process, as it will cause a loss of valuable volatile oil.

Spices, spice seeds, and herbs are sold whole or ground, in small containers for the retail trade, or in bulk quantities. Bulk buyers include meat packers, fancy sausage specialists, pickle manufacturers, producers of condiments, commercial bakers, confectioners, and the catering industry.

The United Nations and the Organization for Economic Co-operation and Development provide statistics of imports and exports reported by several but not all countries. Many countries provide only a partial report; world totals would be greater if all information were included. The international category 075.0 covers a limited number of spice items (see Table 3).

Production. Spices in general, after harvesting and collection, are cleansed of extraneous matter, cured if necessary, dried in the sun or by artificial heat or both, sorted, and graded. Spice seeds are obtained by threshing

Table 2: Spices and their Regions of Origin

Allspice	Guatemala†
	Honduras, British†
	Honduras, Republic†
	Jamaica*
	Mexico*
Aniseed	Bulgaria†
	Lebanon†
	Mexico*
	The Netherlands*
	Spain*
	Syria†
	Tanzania†
	Turkey†
	U.S.S.R.†
	Yugoslavia†
Basil, sweet	Belgium†
	France*
	Hungary*
	Italy†
	U.S.*
	Yugoslavia*
Caraway seed	Denmark*
	Ethiopia†
	Lebanon*
	The Netherlands*
	Poland*
	U.S.S.R.†
Cardamom	Cambodia†
	Costa Rica†
	El Salvador†
	Guatemala*
	India*
	Malaysia†
	Sri Lanka (Ceylon)†
	Tanzania†
	Thailand†
Celery seed	France*
	India*
	Indonesia†
Chervil	France†
	U.S.*
Chili peppers	Chile†
	China†
	Ethiopia*
	Fiji†
	Greece†
	Haiti†
	India*
	Indonesia†
	Israel†
	Japan*
	Kenya*
	Malaysia†
	Mexico*
	Nigeria*
	Pakistan*
	Sierra Leone†
	Sudan†
	Tanzania*
	Taiwan†
	Thailand†
	Turkey†
	Uganda†
	U.S.*
Cinnamon, Ceylon	Kenya†
	Malagasy Republic†
	Seychelles†
	Sri Lanka (Ceylon)*
Cinnamon, cassia	China*
	India†
	Indonesia*
	Taiwan†
	South Vietnam*
Cassia buds	China*
Cloves	Indonesia*
	Malagasy Republic
	Malaysia*
	Mozambique†
	Sri Lanka (Ceylon)†
	Tanzania*
Coriander seed	Algeria†
	Argentina*
	Bulgaria†
	Cyprus†
	Czechoslovakia†
	Ethiopia†
	India*
	Iran†
	Mexico†
	Morocco*
	Nigeria†
	Pakistan†
	Poland†
	Romania*
	Spain*
	U.S.†
	U.S.S.R.†
	Yugoslavia*
Cuminseed	Argentina†
	China†
	Cyprus†
	India*
	Indonesia†
	Iran*
	Japan†
	Lebanon*
	Mexico†
	Morocco†
	Syria†
	Turkey†
Dillseed	China†
	India*
	Mexico†
	U.S.†
Fennel seed	Argentina*
	Brazil†
	Bulgaria*
	China†
	Denmark†
	France†
	Germany, West*
	Greece*
	India*
	Italy†
	Korea, South†
	Lebanon*
	Romania†
	Taiwan†
	Yugoslavia†
Fenugreek seed	Egypt†
	Ethiopia†
	France†
	India*
	Iran†
	Lebanon†
	Morocco†
Garlic	Argentina*
	Bulgaria‡
	China†
	Egypt‡
	Ethiopia†
	Hungary‡
	India†
	Iran†
	Italy†
	Mexico‡
	Peru†
	Portugal‡
	Spain†
	Taiwan‡
	U.S.‡
	U.S.S.R.†
Ginger	China†
	Fiji†
	Guatemala†
	Haiti†
	Hong Kong†
	India*
	Jamaica*
	Japan†
	Kenya†
	Nigeria*
	Puerto Rico†
	Sierra Leone*
	Taiwan†
	Thailand†
	Uganda†
Laurel (bay) leaves	Cyprus†
	Greece†
	Mexico†
	Portugal*
	Turkey*
	Yugoslavia†
Marjoram, sweet	Chile†
	France†
	Lebanon†
	Mexico†
	Peru†
	Romania†
	U.S.†
Mint	Bulgaria*
	Egypt*
	France*
	Germany, West†
	Greece*
	Morocco*
	Romania*
	South Africa†
	U.K.*
	U.S.†
	U.S.S.R.*
Mustard seed	Bulgaria†
	Canada*
	Chile†
	Denmark*
	Ethiopia*
	Germany, West†
	Italy†
	The Netherlands†
	U.K.*
	U.S.S.R.†
	U.S.†
Mace	Grenada*
	India†
	Indonesia*
	Jamaica†
	Leeward Islands†
	Malaysia†
	Sri Lanka (Ceylon)†
	Trinidad and Tobago†
Nutmeg	Grenada*
	India†
	Indonesia*
	Jamaica†
	Malaysia†
	Sri Lanka (Ceylon)†
	Trinidad and Tobago†
Onion	Argentina*
	Bulgaria‡
	Egypt‡
	Ethiopia†
	France†
	Germany, West†
	Guatemala†
	Hungary‡
	Iran†
	Israel‡
	Lebanon†
	Peru†
	Portugal†
	Romania*
	Spain†
	Taiwan†
	U.S.‡
	U.S.S.R.†
Origanum (Oregano)	Cyprus†
	Dominican Republic†
	France†
	Germany, West†
	Greece*
	Italy†
	Japan†
	Lebanon†
	Mexico*
	Portugal†
	Spain†
	Turkey†
	U.S.†
Paprika	Bulgaria*
	Chile†
	Czechoslovakia†
	Egypt†
	Hungary*
	Mexico†
	Morocco*
	Portugal*
	Romania†
	Spain*
	Turkey†
	U.S.†
	U.S.S.R.†
	Yugoslavia*
Parsley	Belgium†
	Canada†
	France†
	Germany, West†
	Hungary†
	Spain†
Parsley (cont)'d	U.K.†
	U.S.†
Pepper	Brazil*
	Cambodia†
	Ivory Coast†
	India*
	Indonesia*
	Jamaica†
	Malagasy Republic†
	Malaysia*
	Sarawak*
	Sri Lanka (Ceylon)*
	Thailand†
Poppy seed	Bulgaria†
	Denmark†
	Iran†
	The Netherlands*
	Poland*
	Romania*
	Turkey*
	U.S.S.R.*
Rosemary	France†
	Spain†
	U.S.†
	Yugoslavia†
Saffron	Egypt†
	Greece†
	Spain*
	Turkey†
Sage	Albania*
	Cyprus†
	Czechoslovakia†
	Greece†
	Italy†
	Lebanon†
	Turkey†
	U.S.†
	U.S.S.R.†
	Yugoslavia*
Savory	France*
	Spain*
	U.S.*
	Yugoslavia*
Sesame seed	Brazil†
	China*
	Colombia†
	Egypt†
	El Salvador*
	Ethiopia*
	Guatemala*
	India*
	Iraq†
	Japan†
	Lebanon†
	Mexico*
	Nicaragua*
	Sudan†
	Syria†
	Thailand†
	U.S.†
	U.S.S.R.†
Star anise	China*
	North Vietnam*
Tarragon	France*
	The Netherlands†
	U.S.*
Thyme	Cyprus†
	France*
	Greece†
	The Netherlands†
	Spain*
	Turkey†
	U.S.†
Turmeric	China†
	Haiti†
	Honduras, Republic*
	India*
	Indonesia†
	Jamaica*
	Pakistan†
	Peru†
	Sri Lanka (Ceylon)†
	Taiwan†
Vanilla	Indonesia*
	Malagasy Republic*
	Mexico*
	Mozambique†
	Réunion†
	Seychelles†
	Tahiti†
	Uganda†

*Major source. †Other sources. ‡Includes dehydrated garlic, onion.

Table 3: Spice and Herb Production

	imports (metric tons)	value (U.S. dollars)	exports (metric tons)	value (U.S. dollars)
Barbados	57	50,338	7	8,099
Belgium	1,788	1,502,503	83	135,061
Canada	5,996	4,863,889	60,292*	8,896,296
Costa Rica	94	117,497	2	1,334†
Cyprus	904	42,360	179	11,844
Ethiopia	797	238,953†	30,861	7,727,351†
Fiji			522	154,840†
France	9,899	7,935,088	713	1,615,038
Guatemala	111	80,567	3,990	2,449,257
Honduras, British	29	22,879	5	3,682
Honduras, Republic of	115	127,055	805	232,306
India			60,120	37,572,485
Iran	5,031	2,078,228	371	261,030
Jamaica	489	63,552	3,003	690,655
Mauritius	822	307,233	901	100,710
Mexico	2,462	3,953,901	4,730	3,079,419
Netherlands, The	11,439	3,963,900	12,517	3,608,800
Peru	1,214	977,044	542	115,860
Portugal	424	410,228	229	144,000
Seychelles			3,107	1,347,432
Singapore	42,049	23,288,880	51,449	29,218,469
Spain	1,648	1,337,500	10,000	10,359,100
Sri Lanka (Ceylon)			5,039	7,118,758
Thailand	337†	296,390	5,690	1,569,866
Trinidad and Tobago	437	236,031	59	49,392
Turkey	415	238,719	1,847	762,862
United Kingdom	11,402	6,824,054	1,120	972,969
United States	113,559	62,548,321	3,080	4,547,065
Vietnam, South	116	32,367	22	56,225
From OECD 075.0‡				
Austria	2,882	1,903,000	49	43,000
Germany, West	23,295	17,054,000	1,112	1,923,000
Ireland	292	235,000		
Norway	680	676,000	3	5,000
Sweden	1,983	3,121,000	117	247,000
From United Nations 075.0‡				
Australia	1,962	1,466,000		
Denmark	1,169	1,294,000	268	156,000
Finland	742	1,463,000		
Greece	807	659,000	158	127,000
Iceland	127	143,000		
Israel	487	387,000		
Italy	3,954	3,391,000	431	359,000
Japan	8,567	5,151,000	2,517	1,787,000
Libya	1,803	816,000		
Switzerland	2,063	2,445,000	21	114,000
Taiwan	955	484,000		
Tunisia	360	287,000	2,099	729,000
Yugoslavia	1,021	809,000	2,138	952,000
Calculated from exports to OECD countries 075.0‡				
Argentine	§	§	143	42,000
Bolivia	§	§	100	50,000
Brazil	§	§	6,533	3,917,000
Egypt	§	§	202	80,000
Hong Kong	§	§	274	213,000
Indonesia	§	§	31,467	24,527,000
Iraq	§	§	56	30,000
Israel	§	§	90	156,000
Kenya	§	§	317	158,000
Malagasy Republic	§	§	3,568	15,358,000
Malaysia	§	§	10,149	7,133,000
Morocco	§	§	6,327	1,656,000
Mozambique	§	§	74	26,000
Nigeria	§	§	2,149	790,000
Oceania, British	§	§	492	134,000
Oceania, French	§	§	61	752,000
Pakistan	§	§	315	135,000
Philippines	§	§	58	45,000
Port. Asia	§	§	245	263,000
Réunion	§	§	147	814,000
Senegal	§	§	57	54,000
Sierra Leone	§	§	929	341,000
Sino-Soviet Bloc	§	§	18,034	9,761,000
South Korea	§	§	152	54,000
Syria	§	§	148	66,000
Tanzania	§	§	1,233	1,345,000
Uganda	§	§	406	252,000
Vietnam, North	§	§	113	93,000
Total	264,783	163,321,477	354,037	197,497,205

*Mostly mustard seed. †Estimated. ‡075.0 includes aniseed, allspice, capsicum spices, caraway, cardamom, cassia, cassia buds, cinnamon, cloves, clove stems, coriander, cumin, curry powder, fennel, ginger, laurel (bay) leaves, mace, nutmeg, pepper, saffron, star anise, thyme, vanilla beans. §No record. All data 1968.

the harvested plants after they have been stacked in the field to mature and dry. Herbs are harvested, suspended, or stacked in the open to dry, and the dried leaves rubbed from the stalks; or following harvesting they are dried by exposure to controlled artificial heat and then processed mechanically to produce herbs of good colour, full flavour, and desired fragment size. Consumer demand at retail differs the world over because cuisine differs.

Blends. Ground mixtures of spices, some long established and some newly introduced, are available.

Meat tenderizers. Meat tenderizers may be seasoned or unseasoned. The seasoned tenderizers contain appropriate ground spices. Both types usually contain salt and dextrose, some contain monosodium glutamate, but the principal tenderizing substance in all of them is papain. Papain is derived from the latex or the unripe fruit of the papaya, or papaw, tree, *Carica papaya*, native to tropical America. Papain is a proteolytic enzyme that acts as a protein digestive.

Peels. Peels such as lemon and orange produced by spice manufacturers consist of the dried and finely comminuted peels blended, respectively, with lemon oil and orange oil.

Condiments. Condiments (Latin *condire*, "preserve") are, loosely, any kind of relish or seasoning added to prepared food dishes. In this sense pepper could be classified as a condiment, but the term usually includes such preparations as chutneys, catsup, tabasco sauce, Worcestershire sauce, barbecue sauce, soya sauce, mayonnaise, salad dressing, prepared mustard, etc. The last mentioned item is more closely allied to spices; the others contain spices among a number of ingredients. The major ingredient of prepared mustard is mustard flour (ground mustard seed), although other spices are present as well as vinegar, salt, and sugar or dextrose. Prepared mustard with horseradish added is also available.

Essential oils. Essential oils are so called because they contain the aroma and flavour of the plants. They are also called volatile oils because they evaporate when exposed to air.

In addition to the principal flavour component or components of the oils, lesser amounts of other natural organic chemical constituents present in the essential oil contribute to the full flavour of the spice, seed, or herb.

Oleoresins. Oleoresins, as their name implies, are natural combinations of essential oils and resins, more or less liquid in consistency. Resins are complex natural oxidation products of essential oils. Oleoresins occur naturally in capsicums, ginger, laurel (bay) leaves, pepper, star anise, and turmeric.

Commercially, an oleoresin is prepared from a spice or herb by extraction with a selected volatile, organic solvent. This process yields a mixture of the essential oil, soluble resins, and related substances present in the spice or herb depending on the kind of solvent employed. These mixtures contain the flavouring principles present in the original spice or herb.

By combining essential oils and oleoresins in various processes, products are obtained that replace the natural spice or herb and can be used in the manufacture of soluble spice flavourings, stable flavourings for use by commercial bakeries, and complete flavourings for all types of processed meats and fancy sausages.

Soluble spice flavourings are made by distributing the oleoresin product on a dry, soluble, edible substance such as salt or dextrose. When other substances such as ground spices are added, the complete mixture is known commercially as a dry soluble seasoning.

Maintenance of quality. *Quality control.* All bulk spices, whole and ground, are stored in dry, clean, sanitary, rodent-proof, cool premises. Temperature and humidity are controlled to inhibit mold growth and discourage insect infestation in the interest of consumer protection, and to avoid loss to the manufacturer.

Whole spices and herbs intended for grinding should not be stored too long. A slow deterioration of essential oil

Table 4: Mixed Seasonings and Flavour Enhancers

blend	typical ingredients
Apple pie spice	cinnamon, allspice or cloves, ginger, nutmeg or mace
Barbecue spice	cayenne or red pepper, cumin, cloves or allspice, garlic
Celery pepper	ground celery seed, black pepper
Celery salt	ground celery seed, free-running salt
Chili powder	capsicums
Cinnamon sugar	cinnamon, sugar
Curry powder	turmeric, cumin, fennel, coriander, fenugreek, cayenne or red pepper, black or white pepper, with or without ginger, allspice, cloves, cardamom (depending on quality); curry powder is "Hot" or "Mild" and formulae, accordingly, include or omit the more pungent spices
Garlic pepper	garlic powder, black or white pepper
Garlic salt	garlic powder, free-running salt, minced garlic
Herbs, fine	sweet marjoram, sweet basil, rosemary, thyme
Hamburger seasoning	allspice or cloves, celery, onion powder, white pepper, sage
Italian seasoning	sweet basil, sweet marjoram, origanum (oregano), rosemary, savory
Mincemeat spice	cinnamon, allspice or cloves, nutmeg, mace
Onion pepper	onion powder, black or white pepper
Pastry spice	cinnamon, allspice or cloves, nutmeg or mace, cardamom, anise, sugar
Poultry seasoning	white pepper, sage, origanum (oregano), savory, thyme, allspice or cloves, onion powder
Pumpkin pie spice	cinnamon, allspice or cloves, nutmeg, mace, ginger
Pickling spice	allspice, cracked bay leaves, cardamom seed, cracked cassia, whole small chilies, cloves, cracked ginger, broken mace, mustard seed, black pepper
Flavour enhancers	
Salad seasoning	a mixture of spices including onion and garlic powder, sesame and celery, salt, and monosodium glutamate
Seasoned pepper	a mixture of spices including garlic and onion powder, black pepper, and monosodium glutamate
Seasoned salt	a mixture of salt, spices, spice extractives, and monosodium glutamate
Monosodium glutamate	$HOOCCH(NH_2)CH_2CH_2COONa$, white crystalline powder, very soluble in water, employed to impart a meat flavour to foods; very little is required to give the desired effect

may occur with the passage of time although the cellular structure is intact. Spices and herbs with oil glands on or near the surface are more vulnerable than those with oil secretion cells deeply seated in the tissues.

Ground spices should not be left exposed to the air but kept in tightly closed containers while awaiting packaging. Grinding fractures and tears apart the tissues and cell walls, and loss of volatile constituents will occur in direct ratio to surrounding air temperature unless the products are properly contained.

All retail packages and containers should afford full protection against this kind of loss.

Laboratory tests

Laboratory quality-control checks can be made by sampling and routine analysis. For this purpose the spice is ground fine enough to pass through a sieve with circular openings one millimetre in diameter. The sample should be thoroughly representative of the lot.

The procedure usually involves the following determinations: (1) moisture, (2) total ash, (3) ash insoluble in hydrochloric acid, (4) volatile oil, (5) nonvolatile ether extract, (6) crude fibre, (7) starch.

In North America, spices and herbs offered for sale to the public are subject to spot sampling and analysis by government agencies responsible under the food laws.

Shelf life. The shelf life of spices depends largely upon the form of the spice and the material used for packaging. The shelf life from the date of manufacture varies as follows: flaked items in tins, 12 months; whole spices in cartons, 18 months; ground spices in tins, 24 months; and spices packed in glass containers, 36 months.

Standards. Standards of identity are promulgated by governments to regulate the industry and protect the public. Minimum and maximum limits for physical and chemical properties insure against inferiority or adulteration. These requirements must be met by foreign exporters and by importers at home.

Not all countries have standards relative to spices, spice seeds, and herbs, but there is universal recognition of the value of such standards to exporters, importers, and manufacturers, and of their necessity in the interests of consumer protection. A trend to the promulgation of standards by more nations is now evident.

FLAVOURINGS

Flavouring extracts, essences, and flavours are liquids added to foods to enhance taste and aroma. They are prepared from essential oils, such as almond and lemon; from vanilla; from fresh fruits by expression; from ginger by extraction; from mixtures of essential oils and synthetic organic chemicals; or entirely from synthetic chemicals, with alcohol, glycerol, propylene glycol, alone or in combination, as solvents. Water is added and sometimes certified food colour.

Extracts, essences, and flavours employing only natural flavouring agents are called pure; those employing synthetics (in part or entirely) are called imitation or artificial flavourings.

Essential oils, complex substances derived from plants, consist of a number of organic chemical components including alcohols, aldehydes, ethers, esters, hydrocarbons (terpenes, sesquiterpenes, etc.), ketones, lactones, phenols, and phenol ethers. Nearly all of these organic chemicals have been synthesized, and it is these synthetics that are employed in the manufacture of imitation flavourings.

Essential oil extracts. These extracts are prepared by dissolving an essential oil in alcohol of the proper strength, adding water and, where desirable and permitted by law, a small amount of certified food colour. They include almond, anise, celery, cassia or cinnamon, clove, lemon, nutmeg, orange, rosemary, savory, basil, sweet marjoram, thyme, and wintergreen.

Nonalcoholic flavouring extracts. Nonalcoholic flavouring extracts are prepared by employing glycerol or propylene glycol or a combination of these to bring the essential oil into solution, with added water, and sometimes food colour.

True fruit flavours. True fruit flavours are obtained by expression, concentration, or distillation of fresh fruits. The alcoholic content of the finished product is usually between 18 and 22 percent to prevent fermentation.

Imitation, artificial extracts, essences, flavours. Imitation, artificial extracts, essences, and flavours are prepared by bringing into solution with alcohol, glycerol, or propylene glycol various synthetic flavouring agents to formulate an extract, essence, or flavour with the likeness of the flavour of the fruit, spirit, or liqueur for which it is named. These preparations cover a wide range of flavourings including vanilla, lemon, lime, banana, cherry, butterscotch, brandy, and rum.

Some contain few ingredients, others many. Imitation vanilla extract may be prepared with synthetic vanillin, coumarin (this ingredient is now prohibited in the United States and Canada), heliotropin, glycerol, water, and caramel colouring.

Imitation strawberry flavour may contain 10, 12, or more synthetic organic chemicals including aldehyde C16, aldehyde C14, benzyl acetate, methyl anthranilate, amyl acetate, amyl butyrate, and ethyl propionate, with solvents and added colour, according to the particular manufacturer's art, skill, and experience in bringing out the fullest and most realistic flavour.

Imitation brandy flavour may contain, with other organic chemicals, ethyl oenanthate, methyl succinate, ethyl acetate, and ethyl propionate.

Certain synthetic organic chemicals lend a distinctive note to imitation flavourings—for example, allyl caproate and ethyl butyrate to pineapple, benzaldehyde to almond, benzyl butyrate to raspberry, and citral to lemon.

Imitation extracts, essences, and flavours must be labelled accordingly.

Quality control and analysis. In the United States the industry is presided over by the Flavor and Extract Manufacturers' Association, FEMA, incorporated in 1909. The association assisted in the framing of federal laws to ensure quality of product and continues in this work in the interests of the industry and for the protection of the consumer. FEMA played an important part in arriving at an efficient method of analysis for vanilla, the most widely employed flavouring product, including such modern techniques as paper and column chromatography.

The International Organization of Flavour Industries,

International Organization of Flavour Industries

IOFI, with headquarters in Brussels, was recently incorporated under the laws of Belgium. Its main purpose is the continuing evaluation of flavour ingredients for safety, and its aim is directed toward the establishment of a universal code of flavour laws. The Flavor and Extract Manufacturers' Association is a member of IOFI.

For analytical procedures relevant to flavouring extracts consult "Official Methods of Analysis" of the Association of Official Analytical Chemists, 10th ed. (1965), pages 286-305, or other standard analytical texts.

BIBLIOGRAPHY. General works on spices and herbs include: J.W. PARRY, *Spices*, vol. 1, *The Story of Spices and the Spices Described* (1969), a history of spices, with descriptions of individual spices, spice seeds, and herbs (illustrated); F. ROSENGARTEN, JR., *The Book of Spices* (1969), a comprehensive and illustrated work including a history of spices, a description of individual spices and herbs, and a number of culinary recipes using them; and AMERICAN SPICE TRADE ASSOCIATION, *A Glossary of Spices* (1966), an illustrated booklet describing and suggesting uses for various spices, spice seeds, herbs, and blends.

Food flavourings are treated in J. MERORY, *Food Flavourings: Composition, Manufacture, and Use* (1960), a guide to the manufacture of such flavouring extracts as vanilla, fruit flavours, and others; and the *Organic Chemicals Handbook* (1963), a booklet issued by Fritzsche Brothers, Inc. (New York) listing various aromatic organic chemicals and describing their colour, appearance, volatility, and application. More technical treatment is offered in J.W. PARRY, *Spices*, vol. 2, *Morphology, Histology, and Chemistry* (1969). *The Oleoresin Handbook* (n.d.), an illustrated booklet issued by Fritzsche Brothers, Inc. (New York) treating the commercial production of oleoresins, also covers various spices, spice seeds, herbs, and vanilla extract, and includes an appendix on U.S. Federal specifications. The *Official Analytical Methods of the American Spice Trade Association* (1965) treats accepted analytical methods of the U.S. industry; the *Official Methods of Analysis of the Association of Official Agricultural Chemists* (1965) treats accepted analytical methods for this organization.

Legislation, regulation, and standards of identity issued by various governments are treated in a number of official publications, such as the UNITED STATES DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE, *The Federal Food, Drug, and Cosmetic Act*, as amended (1968); and the CANADIAN DEPARTMENT OF NATIONAL HEALTH AND WELFARE, *The Food and Drugs Act and Regulations* (1969).

(J.W.P.)

Spinoza, Benedict de

Benedict de Spinoza (Despinoza) was a 17th-century Rationalist, author of one of the most consummate systems of metaphysics in the history of philosophy. Convinced that every feature of reality is necessary and that every logically coherent possibility must exist, Spinoza set out to deduce the nature of reality, taking Euclid's geometry as a model, from a set of primitive definitions and axioms. Spinoza's significance thus lies not as much in the particular merits of the system itself (which was pantheistic) as in the fact that it comprises a specific sample of a thoroughgoing metaphysical system, which scholars can examine in order to evaluate the validity and promise (or their opposites) of the metaphysical enterprise in general.

Early life and career. Spinoza was born of Jewish parents in Amsterdam on November 24, 1632. His Hebrew name was Baruch (meaning blessed), which by custom was rendered into its Latin equivalent, Benedictus.

Spinoza's grandfather and father were Portuguese and had been crypto-Jews after the Spanish Inquisition had compelled them to embrace Christianity. Later, after Holland's successful revolt against Spain and the granting of religious freedom, they found refuge in Amsterdam. His mother, who also came from Portugal, died when Benedict was barely six years old.

The Spinozas were prosperous merchants and respected members of the Jewish community, and it may be assumed that Spinoza attended the school for Jewish boys founded in Amsterdam about 1638. Its curriculum included the traditional Hebrew subjects: the Old Testament, the Talmud, the Hebrew codes, and the works of Ibn Ezra, Maimonides, Crescas, and others. Outside school hours the boys had private lessons in secular subjects. Spinoza was taught Latin by a German scholar, who may also have taught him German; and he knew to some extent all of the other significant continental languages. When he learned mathematics and physics is not known.



Spinoza, oil painting by an unknown artist, c. 1665. In the Herzog-August Bibliothek, Wolfenbüttel.
By courtesy of the Herzog-August Bibliothek, Wolfenbuttel

In March 1654 Benedict's father died. There was some litigation over the estate, with Benedict's only surviving stepsister claiming it all. Benedict won the lawsuit but allowed her to retain nearly everything.

His studies so far had been mainly Jewish, but he was an independent thinker and had found more than enough in his Jewish studies to wean him from orthodox doctrines and interpretations of Scripture; moreover, the tendency to revolt against mere tradition and authority was very much in the air in the 17th century. But the Jewish religious leaders in Amsterdam were fearful that heresies (which were no less anti-Christian than anti-Jewish) might give offense in a country that did not yet regard the Jews as citizens.

Difficulties with Jewish authorities

Spinoza soon incurred the disapproval of the synagogue authorities. In conversations with other students, he had held that there is nothing in the Bible to support the views that God had no body, that angels really exist, or that the soul is immortal; and he had also expressed his belief that the author of the Pentateuch (the opening books of the Bible) was no wiser in physics or even in theology than were they, the students. The Jewish authorities, after trying vainly to silence Spinoza with bribes and threats, excommunicated him in July 1656, and he was banished from Amsterdam for a short period by the civil authorities. There is no evidence that he had really wanted to break away from the Jewish community, and indeed the scanty knowledge available would suggest the opposite. On December 5, 1655, for example, he had attended the synagogue and made an offering that, in view of his poverty, must have been a rare event for him; and, about the time of his excommunication, he had addressed a defense of his views to the synagogue.

Among Spinoza's Christian acquaintances was Francis van den Enden, who was a former Jesuit, an ardent classical scholar, and something of a poet and dramatist and who had opened a school in Amsterdam. For a time, Spinoza stayed with him, helping with the teaching of the schoolchildren and receiving aid in his own further education. In this way he improved his knowledge of Latin, learned some Greek, and was introduced to the Neoscholastic philosophy of such writers as Franco Burgersdijck and Adrianus Heereboord. It may have also been through van den Enden's school that Spinoza became acquainted with the "new philosophy" of René Descartes, later acknowledged to be the father of modern philosophy. Spinoza's other Christian acquaintances were mostly of the Collegiants, a brotherhood that later merged with the

Mennonites; they were especially interested in Cartesianism, the dualistic philosophy of Descartes and his followers (see below).

At the same time, he was becoming expert at making lenses, supporting himself partly by grinding and polishing lenses for spectacles, telescopes, and microscopes; he also did tutoring. One of Spinoza's friends at this period was Simon Joosten de Vries, whose admiration for him was so great that he pressed him to accept a gift of 2,000 florins and wanted to make him his sole heir. This Spinoza refused, and, when he was bequeathed an annuity of 500 florins, he would accept only 300. A kind of reading and discussion circle for the study of religious and philosophical problems came into being under the guidance of Spinoza. In order to collect his thoughts, however, and reduce them to a system, he withdrew in 1660 to Rijnsburg, a quiet village on the Rhine, near Leiden.

It was in 1661, during his stay at Rijnsburg, that Spinoza met Heinrich Oldenburg, soon to become one of the two first secretaries of the Royal Society in London, with whom Spinoza was later to conduct the most important part of his correspondence. Rijnsburg was the headquarters of the Collegiants, and Spinoza's lodgings there were with a surgeon named Hermann Homan. In Homan's cottage Spinoza wrote *Korte Verhandeling van God, de Mensch en deszelfs Welstand* (written *c.* 1662; *Spinoza's Short Treatise on God, Man, and His Well-Being*, 1910) and *Tractatus de Intellectus Emendatione* ("Treatise on the Correction of the Understanding"), both of which were ready by April 1662. He also completed the greater part of his geometrical version of Descartes's *Principia Philosophiae* (with the appendix of "Cogitata Metaphysica" ["Metaphysical Thoughts"]) and the first book of his *Ethica*. Spinoza's attitude in these works already showed a departure from Cartesianism. Part of the *Korte Verhandeling*, it is conjectured, was based on notes that he had prepared in connection with the Amsterdam study circle, to which he seems also to have sent periodically the various parts of the *Intellectus Emendatione* and other works for study and discussion.

Early works in Rijnsburg

Influence of Descartes and the geometrical method. His version of Descartes's *Principia* (which covered at first only its second part) was prepared while Spinoza was giving instruction in the philosophy of Descartes to a private pupil. When he showed this work to his Cartesian friends in 1663, they persuaded him to do the same with the first part, which he then did in a fortnight. These two parts, together with its appendix, "Cogitata Metaphysica," were accordingly published by his friends under the title *Renati des Cartes Principiorum Philosophiae Pars I et II, More Geometrico Demonstratae, per Benedictum de Spinoza* (1663), with an introduction explaining that Spinoza did not share the views expressed in the book. This was the only book published in Spinoza's lifetime with his name on the title page.

The philosophy of Spinoza may thus be regarded as a development from and a reaction to that of his contemporary Descartes (1596-1650). Though it has been argued that Spinoza was also much influenced by medieval philosophy (especially Jewish), he seems to have been much more conscious of the Cartesian influence, and his most striking doctrines are most easily understood as solutions of Cartesian difficulties. Clearly, he had studied Descartes in detail. He accepted Descartes's physics in general, though he did express some dissatisfaction with it toward the end of his life. As for the Cartesian metaphysics, he found three very unsatisfactory features, viz., the transcendence of God, the substantial dualism of mind and body, and the ascription of free will both to God and to human beings. In Spinoza's eyes, those doctrines made the world unintelligible. It was impossible to explain the relation between God and the world or between mind and body or to account for events occasioned by free will.

The publication of Spinoza's version of Descartes's *Principia* had been intended to prepare the way for that of his own philosophy, for he had both to secure the patronage of influential men (as a shield against the hostility of the Calvinists) and to show the more philosophically minded that his rejection of Cartesianism was not out of ignorance.

Spinoza became dissatisfied with the informal method of exposition that he had adopted in the *Korte Verhandeling* and the *De Intellectus Emendatione* and turned instead to the geometrical method in the manner of Euclid's *Elements*. He assumed without question that it is possible to construct a system of metaphysics (*i.e.*, an account of the world as a whole) that will render it completely intelligible. It is therefore possible, in his view, to present metaphysics deductively—that is, as a series of theorems derived by necessary steps from self-evident premises expressed in terms that are either self-explanatory or defined with unquestionable correctness. His masterpiece, the *Ethica*, was set out in this manner—*Ordine Geometrico Demonstrata*, according to the reading of its subtitle. Its first part, "De Deo" ("Concerning God"), was finished and in the hands of his friends early in 1663. Initially the work was intended to have three parts only, but it eventually appeared (in 1677) in five parts. Spinoza's desire for an impersonal presentation was probably his chief motive for adopting the geometrical method, appreciating that the method guarantees true conclusions only if the axioms are true and the definitions correct. Spinoza, like his contemporaries, held that definitions are not arbitrary but that there is a sense in which they may be correct or incorrect.

Adoption of the geometrical method

The question was discussed at length in his unfinished *Tractatus de Intellectus Emendatione*. A sound definition, he held, should make clear the possibility or the necessity, as the case may be, of the existence of the object defined. Because the *Ethica* begins with the definition of "substance," the necessary existent, the entire system is vulnerable to anyone disputing that definition, however cogent the subsequent reasoning may be. In fact, as Gottfried Wilhelm Leibniz, a Rationalist philosopher and mathematician, pointed out, though the system is closely knit, its demonstrations do not proceed with mathematical rigour.

Period of the "Ethica." In June 1663 Spinoza moved to Voorburg, near The Hague, and it appears that by June 1665 he was nearing the completion of the *Ethica*. During the next few years, however, he was at work on his *Tractatus Theologico-Politicus*, which was published anonymously at Amsterdam in 1670. This work aroused great interest and was to go through five editions in as many years. It was intended "to show that not only is liberty to philosophize compatible with devout piety and with the peace of the state, but that to take away such liberty is to destroy the public peace and even piety itself." As this work shows, Spinoza was far ahead of his time in advocating the application of the historical method to the interpretation of the biblical sources. He argued that the inspiration of the prophets of the Old Testament extended only to their moral and practical doctrines and that their factual beliefs were merely those appropriate to their time and are not philosophically significant. Complete freedom of scientific and metaphysical speculation is therefore consistent with all that is important in the Bible. Miracles are explained as natural events misinterpreted and stressed for their moral effect.

Critico-historical treatment of the Bible

Spinoza also told his friend Oldenburg that he wanted to clear himself of the charge of atheism. Apparently, with the completion of the *Ethica* in sight, Spinoza was considering the question of its publication but found it impracticable in view of the contemporary state of public opinion; he therefore set himself to show that the Bible, properly understood, gives no support to ecclesiastical intolerance or to clerical interference in civil and political affairs.

In May 1670 Spinoza moved to The Hague, where he remained until his death. He began to compose a Hebrew grammar, *Compendium Grammatices Linguae Hebraeae*, but did not finish it; instead, he returned to the *Ethica*, although the prospect of its publication became increasingly remote. There were many denunciations of his *Tractatus Theologico-Politicus* as an instrument "forged in hell by a renegade Jew and the devil, and issued with the knowledge of Mr. Jan de Witt" (a promi-

nent but controversial Dutch statesman, who had shown sympathy for Spinoza). De Witt's murder by a Dutch mob in 1672 during a war with France provoked the philosopher to plan to post placards decrying "the very lowest of barbarians" responsible for the murder—an act that might have cost him his life had he not been forcibly prevented from carrying it out by his landlord. Further, Spinoza's political situation in The Hague was compromised in May 1673, when he went to Utrecht (then under French occupation) with a view to possible negotiations for peace. On his return several weeks later, he was badly received by the suspicious populace of The Hague. When the *Ethica* was finally completed in 1675, Spinoza had to abandon the idea of publishing it, though manuscript copies were circulated among his intimate friends.

Substance, attributes, and modes of the *Ethica*

The Ethica begins by distinguishing substance, attributes, and modes. Substance is that which can be conceived as existing in complete independence. Because it cannot be explained by anything else, it must be its own cause, or necessarily existent. There can be only one such substance, and this is quickly identified with God and with the whole of Nature. Spinoza thus arrives at a pantheism. The most shocking feature of his philosophy was his denial not only of the transcendence of God but also of God's personality, providence, free will, and purposes. The good man, although he loves God, will not expect God to love him in return.

Substance, in Spinoza's view, has an infinity of attributes. By an "attribute" is understood "that which the intellect perceives of substance, as constituting its essence." Of these attributes, only thought and extension are known to man. Thus, the universe is a spatial whole that is conscious throughout, and "All things . . . ," as he says, "are alive"—a position known as panpsychism. The theory of infinite attributes, however, poses difficulties. Since all attributes characterize substance throughout its extent, it seems to follow that, underlying every attribute (not merely mind and matter), there should be some entity related to the human mind in the same way that the body is related to it. But, if this were so, one would expect that all of the attributes would be known to man—as is the extension of body—though, in fact, they are not. This difficulty has led some commentators to view Spinoza as saying that the attributes are simply the infinitely many possible ways in which substance could be conceived. It seems clear, however, from Spinoza's 66th letter that he thought of the attributes as objective.

Substance is also diversified into an infinity of finite modes, best understood as "parts"—which cannot, however, be separated from the whole. Human beings and other commonsense objects, which are typically modes, are parts of nature; and Spinoza once referred to the human mind as part of the infinite intellect of God. There is a one-to-one correspondence between modes under the several attributes. There are also infinite modes, both immediate and mediate. The immediate infinite mode under the attribute extension is motion and rest, and the mediate infinite mode is the "aspect of the whole universe"; this last may be a reference to the conservation of motion in the total universe. The immediate infinite mode under thought is said to be the "absolutely infinite intellect"; by this Spinoza may refer to the sum total of consciousness of knowledge.

The whole scheme is completely deterministic. Since the nature of God cannot be other than it is, the whole system of modes, which follows from or expresses the nature of God, must necessarily be as it is. A mode under one attribute cannot be caused by a mode under another attribute, for they have no property in common. Any modal happening under one attribute has its explanation entirely in terms of the general laws that govern modal happenings under that same attribute.

Spinoza's views of body, knowledge, will, emotion

In applying his metaphysical scheme to human beings, Spinoza argued that man's body is a mode under extension, a complex mode owing its unity to the maintenance of a constant pattern of relations between changing parts. The human mind is likewise a system maintaining the same pattern of relation through change of parts. Spinoza sometimes spoke of the mind and body as one and the same thing, conceived now under thought and now under extension. Although there is no causal connection between mind and body, there is a mental correlate for every physical event and vice versa. Thus, for Spinoza, every change in the human body is associated with some change of awareness, however dim and vague it may be, in the human mind; and a human being can know external physical objects only insofar as they affect his own body.

Spinoza held that all sense experience and all unscientific generalization from it is inadequate. No object can be isolated from the rest of nature; hence, no one can state the whole truth about it, since this would involve the whole of nature. Knowledge of this kind—the first of three different grades—was called by Spinoza opinion or imagination. If, however, the consciousness is directed solely toward those properties that all objects have in common, there is no distortion such as occurs in sense experience. This second kind of knowledge is called reason. In this way Spinoza accounted for the possibility of a priori knowledge in geometry, general physics, and general psychology. The third kind of knowledge, intuition, seems to be the adequate knowledge of individual objects. It is possible through intuition, he held, to state of any individual thing that it depends on God as its complete, immanent cause. Perhaps by the term intuition Spinoza referred to a kind of mystical experience, the insight—accompanied by a strong emotion that he called "the intellectual love of God"—into the dependence of all things, including the human being himself, on the whole of nature.

Though the distinction between truth and falsehood about particular things occasioned Spinoza great difficulty, it is at least clear that a man's understanding of an individual thing or event will be adequate in proportion as he can explain its features in terms of general laws, which are the concern of reason, or it will be adequate insofar as he can make correct causal statements about it.

Spinoza maintained that, though deliberately chosen action indeed occurs, it does not involve free will in the ordinary sense. Willing is primarily *thinking* of the course of action to be followed, and, when there are no hindering factors, the action in question inevitably follows. The illusion of undetermined choice arises from one's ignorance of the preceding causes of thought and action. Spinoza's alternative account of freedom is to be found in his distinction between the active and the passive emotions. In human beings, conatus (appetite and desire) is one of the primary emotions. The only other primary emotions are joy, the psychic accompaniment of enhanced power, and grief, that of enfeebled power. Insofar as men are able to understand themselves and their environment, they can control events rather than be controlled by them; they then experience desire and joy in their active forms. A man is free insofar as his emotions are active rather than passive and in proportion as he is able to think clearly and to preserve himself for the maximum time in an active state. Spinoza was the first to emphasize that the complete understanding of emotions and of their causation deprives them of their overmastering quality.

Moral and political philosophy

Spinoza's moral philosophy, as he presented it in the *Ethica,* defines "the good" in largely subjective terms: the good for different species (*e.g.*, for man and for horse) is different. Thus, for God the distinction between good and evil would be meaningless, as it is essentially relative to the purposes of finite creatures. Human beings, however, have enough in common to suggest what is really good for them, however different their views may be. Thus, "good" may be said to have an objective significance for human beings, especially as reflecting the "knowledge of God." Spinoza's ethics are essentially egoistic, but the conduct that he enjoined is that which would be commended by an altruistic Utilitarian: he tried to prove, for example, that hatred may be overcome by love. His ideal man is in many respects comparable to the Stoic sage, for his aim is the maximum self-sufficiency; but Spinoza had a stronger realization of the necessity for cooperation and friendship between rational men in a civil society.

Last years and posthumous influence. After Spinoza had completed the *Ethica*, he concentrated his attention on political problems and began his *Tractatus Politicus*, which he did not live to finish. During the post-*Ethica* period, he was visited by several important people, among them Ehrenfried Walter von Tschirnhaus (in 1675), a scientist and philosopher, and Gottfried Wilhelm Leibniz (in 1676), like Spinoza, one of the foremost Rationalists of the time. Leibniz, having heard of Spinoza as an authority on optics, had sent him an optical tract and had then received from Spinoza a copy of the *Tractatus Theologico-Politicus*, which deeply interested him. According to Leibniz' own account, he "conversed with him often and at great length." Spinoza, however, was now in an advanced stage of consumption, aggravated by the inhaling of glass dust from the polishing of lenses in his shop. On February 21, 1677, he died and was buried four days later in the New Church on the Spui. He was unmarried and left no heir, and his few possessions were sold by auction. These included about 160 books, the catalog of which has been preserved.

In accordance with Spinoza's previous instructions, several of his friends prepared his manuscripts secretly for the press, and they were sent to a publisher in Amsterdam. The *Opera Posthuma* (Dutch version: *Nagelate Schriften*), which was published before the end of 1677, comprised the *Ethica, Tractatus Politicus,* and *Tractatus de Intellectus Emendatione*, as well as letters and the Hebrew grammar. His *Stelkonstige reeckening van den regenboog* ("On the Rainbow") and his *Reeckening van kanssen* ("On the Calculation of Chances") were printed together in 1687. The *Korte Verhandeling* was lost to the world until E. Boehmer's publication of it in 1852.

Spinoza's place in Western intellectual history

Spinoza has an assured place in the intellectual history of the Western world, though his direct influence on technical philosophy has not been great. Throughout the 18th century he was almost universally decried as an atheist—or sometimes used as a cover for the detailing of atheist ideas. The tone had been set by Pierre Bayle, a Skeptical philosopher and encyclopaedist, in whose *Dictionnaire historique et critique* Spinozism was described as "the most monstrous hypothesis imaginable, the most absurd"; and even David Hume, a Scottish Skeptic and historian, felt obliged to speak of the "hideous hypothesis" of Spinoza.

Spinoza was rendered intellectually respectable by the efforts of literary critics, especially of the Germans Lessing and Goethe and the English poet Coleridge, who admired the man and found austere excitement in his works, which they saw as the expression of an intensely religious attitude entirely divorced from dogma. Spinoza has also been much studied by professional philosophers since the beginning of the 19th century. Both absolute Idealists and Marxists have read their own doctrines into his work, and Empiricist philosophers, while rejecting his metaphysical approach, have developed certain detailed suggestions from his theory of knowledge and psychology.

MAJOR WORKS

Korte Verhandeling van God, de Mensch, en deszelf's Welstand (written *c.* 1662, first pub. as *Tractatus de Deo et Homine Ejusque Felicitate*, 1852; *Short Treatise on God, Man, and His Well-Being*, trans. by A. Wolf, 1910); *Tractatus de Intellectus Emendatione* (written 1662, first pub. 1677; trans. by W.H. White, 1899); *Renati des Cartes Principiorum Philosophiae Pars I et II, More Geometrico Demonstratae, per Benedictum de Spinoza* (1663; *Principles of Cartesian Philosophy*, trans. by H.E. Wedeck, 1961); *Ethica in Ordine Geometrico Demonstrata* (written 1662–75, pub. 1677; *Ethics*, trans. by W.H. WHITE and A. STIRLING, 1894); *Tractatus Theologico-Politicus* (written 1665–70, pub. 1670; trans. by R.H.M. ELWES, 1889); *Tractatus Politicus* (unfinished, written from 1665 onward, first pub. 1677; trans. by A.G. Wernham in *The Political Works*, 1958).

BIBLIOGRAPHY. The best critical edition of Spinoza's work, known as the "Heidelberg edition," is by CARL GEBHARDT, 4 vol. (1924); there is also a good edition of the text by J. VAN VLOTEN and J.P.N. LAND, 3rd ed., 4 vol. (1914). For English versions, see *Spinoza's Chief Works*, ed. by R.H.M. ELWES, 2 vol. (1883–84); *Ethics*, trans. by W. HALE WHITE and A.H. STIRLING, 4th ed. rev. (1910); *Spinoza's Short Treatise on God, Man and His Well-Being*, trans. by A. WOLF (1910, reprinted 1963); and *The Correspondence of Spinoza*, trans. by A. WOLF (1928, reprinted 1966). For general works, see the short biography by FREDERICK POLLOCK, *Spinoza: His Life and Philosophy*, 3rd ed. (1912); and JACOB FREUDENTHAL, *Spinoza, Leben und Lehre*, ed. by CARL GEBHARDT, 2 pt. (1927). For biography, see STANISLAUS DUNIN-BORKOWSKI, *Der Junge de Spinoza* (1910); LEON BRUNSCHVICG, *Spinoza et ses contemporains*, 3rd ed. rev. (1923); and the detailed biography by A. WOLF in the introduction to his *Spinoza's Short Treatise . . .* (above). The original materials for a biography are contained in CARL GEBHARDT, *Spinoza: Lebensbeschreibung und Gespräche* (1914). The Jewish background of his thought is explored in LEON ROTH, *Spinoza, Descartes, and Maimonides* (1924, reprinted 1963) and *Spinoza* (1929), a more introductory work. See also RICHARD MCKEON, *The Philosophy of Spinoza* (1928); HARRY A. WOLFSON, *The Philosophy of Spinoza*, 2 vol. (1934, reprinted 1969), who argues that Spinoza was the last of the medievals and places him in the context of medieval Jewish and Scholastic thought. A Hegelian approach is presented in HAROLD H. JOACHIM, *Spinoza's Tractatus de Intellectus Emendatione: A Commentary* (1940) and *A Study of the Ethics of Spinoza* (1901, reprinted 1964). See also GABRIEL HUAN, *Le Dieu de Spinoza* (1914); HAROLD F. HALLETT, *Benedict de Spinoza: The Elements of His Philosophy* (1957); PAUL SIWEK, *Spinoza et le panthéisme religieux*, new ed. (1950) and *Au Coeur du spinozisme* (1952); and AUGUSTO GUZZO, *Il pensiero di Spinoza*, 2nd ed. (1964). His views on method, truth, and knowledge are given in G.H.R. PARKINSON, *Spinoza's Theory of Knowledge* (1954). STUART HAMPSHIRE, *Spinoza* (1956), discusses Spinoza from the standpoint of Analytic philosophy. The following collections of articles on Spinoza contain useful material: *Spinoza*, ed. by MARJORIE GRENE (1972); *Studies in Spinoza*, ed. by P. KASHAP (1972); and *Spinoza: Essays in Interpretation*, ed. by E. FREEMAN and M. MANDELBAUM (forthcoming). Of various bibliographies of Spinoza, mention may be made of ADOLPH S. OKO, *The Spinoza Bibliography* (1964); and JON WETLESEN, *A Spinoza Bibliography* (1968).

Spiritualist Groups

A Spiritualist is a person who believes that communication with those who have "passed into the higher life" (have died) is possible and makes such communication the centre of his religion. This article deals with the spiritualist religion, not with psychic research, which is a study on what are believed to be scientific lines of communication with the dead, telepathy, precognition, and kindred subjects.

The important terms in spiritualism are defined in the following ways. *Spirit* is the essential part of man. After the death of the body the spirit lives on. The *spirit world* is the world of disembodied spirits. A *medium* is a person on earth who is sensitive to vibrations from the spirit world and is able to convey messages between that world and this one and to produce other Spiritualist phenomena. A *control* is a disembodied spirit who gives messages to a medium who in turn gives them to men and women on earth. A *séance* is a meeting (usually small) around a medium to seek messages from the spirits. *Telepathy* is the communication of ideas through other than physical means. *Clairvoyance* is the power of seeing through means other than the physical eye. *Clairaudience* is the power of hearing through means other than the physical ear. *Levitation* is the lifting of an object by other than physical means. *Materialization* is the appearance of a spirit in matter. *Apport* is the production of objects without physical means or the passing of objects through objects (*e.g.*, walls) which have no openings.

HISTORY

Spiritualists believe that throughout human existence spirits have communicated with mortals. The Bible, they claim, is an outstanding record of Spiritualist phenomena. Well-known examples from the Bible are King Saul's clandestine visit to the woman at Endor, through whose mediumship Samuel, in the spirit, spoke to Saul, and Paul's vision of Jesus on the road to Damascus. Spiritualism, it is believed, furnishes the key that unlocks the biblical mysteries. The Transfiguration and the Resurrection are examples of materialization; Jesus walking on the water is an example of levitation; Pentecost was "the greatest séance in history."

During medieval times, the church acknowledged the reality of witches and wizards and forbade commerce with them. It was in the 19th century that Spiritualism as a self-conscious, organized religion began. In 1848 two girls, Margaret and Kate Fox, in Hydesville, a village in western New York, heard strange knockings in the cottage where they lived. They spoke to the spirit who produced the knockings, suggested a code, and began to receive messages. In the years just before these events, another resident of New York state, Andrew Jackson Davis, had written a book he called *The Principles of Nature, Her Divine Revelations*, based on his experiences with the spirits. The Fox sisters and Davis provided the impetus that brought Spiritualism into organized form.

In the years following, many persons both in America and abroad discovered their mediumistic powers and became centres around which groups formed. These groups grew into local churches, and the churches grew into over a dozen congregationally organized regional and national bodies. The *Yearbook of American Churches* lists three: the National Spiritualist Association of Churches founded in 1893 (the NSAC), the International General Assembly of Spiritualists, and the National Spiritual Alliance of the U.S.A. Most of the organizations are small and of short duration; moreover, they keep poor records. The *Yearbook*, whose figures are based on reports by the organizations themselves, shows somewhat less than 200,000 members in the United States in 1956, a figure that is probably much too high for actual formal membership and much too low for the number of persons interested in spiritualism.

Spiritualist organizations

Spiritualists claim that Spiritualist organizations have been formed in over 40 nations. Many of these national groups have joined to form the English based International Spiritualist Federation. Founded in 1923, it reports 38 affiliated societies in 29 countries; 9 of these countries have 2 affiliated societies, and Japan has 3.

Individuals who are members of the more traditional denominations and who have Spiritualist interests and yet do not wish to change their religious affiliation have formed the Spiritual Frontiers Fellowship (American) and the Churches' Fellowship for Psychical and Spiritual Study (British).

The interest in spiritist phenomena also took another organized form. In 1882 the Society for Psychical Research was founded in England. Its purpose was objective investigation, not worship or the giving of spiritual solace. Membership in the society did not, and does not, imply a necessary acceptance of the Spiritualist explanation for psychic phenomena. Among its members have been found such notable figures as Sir William Crookes, Sir Arthur Conan Doyle, F.W.H. Myers, and Sir Oliver Lodge. A parallel organization, the American Society for Psychic Research, was founded in 1885. Some of its well-known members have been William James, J.H. Hyslop, J.B. Rhine, and Gardner Murphy. In France, Camille Flammarion, Henri Bergson, and Gabriel Marcel were notable for their interest in psychic phenomena. Famous mediums whose manifestations have received serious investigation include D.D. Home, Mrs. Leonore E. Piper, Mrs. Gladys Osborne Leonard, and Mrs. Eileen Garrett.

BELIEFS, WORSHIP, ORGANIZATION

Beliefs. The NSAC, with its publication *Spiritualist Manual*, is the best organized of the Spiritualist bodies; the *Manual* is revised periodically. The NSAC encourages freedom of belief and also emphasizes its "Declaration of Principles."

In many of their emphases, these principles are similar to beliefs held by most religious persons in the West. Leaders in the NSAC, without depreciating in any way the importance of the teachings about the next life, urge that the member churches emphasize cultivation of the moral and spiritual virtues that enrich this life. Outside observers often note, however, that the major concern of the average Spiritualist seems to be contact with the spirits and the reception of messages from them.

Man has three elements, according to Spiritualist views: spirit, soul, and body. The body encases the soul and the spirit. The soul encases the spirit. At death the body dies, but the soul and the spirit live on. The soul apparently is made up of elements more refined than those of the body, and it serves the spirit in the next life as the body and soul serve the spirit in this life. Animals are believed to have souls but not spirits.

Spiritualists believe that the human spirit exists eternally as a part of the Universal Spirit and thus antedates the creation of the body and the soul. The NSAC holds that the soul enters the body at conception; some other Spiritualists hold that this entry happens a few months after conception. The soul, like the body, is said to be manufactured out of material elements, such as food, water, and air, but these are more refined than the bodily elements. The soul is the duplicate of the body in structure. At death the soul and the spirit are believed to go to one of the seven spheres that surround the earth. The lowest sphere is occupied by men who on earth were wicked, barbaric, and dominated by material desires. At the death of the body, most men go to the Third Sphere, often called the Summerland. There the life is very like the life of earth, except that it is made "finer" and the evils and sorrows are gone. The spirits in the Summerland are said to wear clothing, live in houses, have animals for pets. Each spirit finds and marries his soul mate (to whom he had been joined before his earthly birth), but the nature of the marriage relationship in the heavenly spheres is wholly platonic.

Beyond the seven earth spheres are higher and more universal spheres; they surround planets, suns, and even solar systems. Farther still are spheres still more tremendous. God, defined as Infinite Intelligence, is at the centre of the universe.

The souls of children who have died are believed to assume adulthood over a period of time. Persons in the lower spheres eventually attain higher spheres, where life is much happier than anything known on earth. Thus Spiritualists are universalists; that is, they believe in salvation for all. There is no hell in the traditional sense, though persons who have lived essentially evil lives must suffer before they can progress to the higher spheres. The law of progression applies to all. Whatever sphere a person is in, he is working to go higher. But he moves higher only through his own efforts, with of course the counsel of spirits wiser than himself.

These propositions are statements of fact, the Spiritualists assert. They claim "indubitable proof" that man lives after death. Their evidence comes from the spirits through mediums. Every human being, according to the theory, is a medium, but most are undeveloped. The simpler mediumistic powers can be cultivated if a person will centre his thoughts on spiritual concerns. *Mental phenomena* are thus said to be within the reach of many people; that is, spiritualists believe that people can and do receive ideas and information from their departed loved ones. Mental phenomena include the dramatic messages that come when a medium goes into trance. Through mediums in trance, prophecies have been made, books written, instructions given to surgeons on how to perform difficult operations, and intricate experiments proposed for proving the non-earthly source of the communications.

The function of mediums

Physical phenomena also are said to be produced by mediums. Apports are reported to occur; for example, when Jesus after the Resurrection stood among his disciples, "the doors being shut" (John 20:19). It is claimed that voices sometimes speak independently of the medium's voice, often magnified by a trumpet. Healings are said to be consummated—healings of physical as well as of psychosomatic illness. It is alleged that objects are levitated and float in the air without physical support. Spirits are said to materialize and thus can be seen with the physical eye; the appearances of angels recorded in the Bible are claimed as examples.

From the bodies of mediums a substance called ectoplasm is said to exude. It is believed to be a rarefied form of matter that the spirits can use to effect physical phenomena. Ectoplasmic hands may, for example, cause levitation and apports.

God is thought to be an impersonal being, the infinite,

eternal, distant Oversoul. Yet the possibility of intimate personal fellowship with superhuman powers is by no means absent from Spiritualism; what other faiths would call the "nearness of God" is for Spiritualists the nearness of our spirit guides, of souls who have gone on ahead of us and who are much concerned for our comfort, health, and safety.

Jesus is not thought to be God nor the second person of the Trinity, nor did he give his life as a ransom for mankind. Jesus was a man like all the rest of us. He was divine as are all men. But he did differ from us in his attainments. According to Spiritualists, he was the greatest medium who ever lived.

The doctrine of reincarnation is held by a minority but is not approved by the majority. The existence of poltergeists (boisterous ghosts) is considered so real that "few physical phenomena [are] better established." The possibility of "spirit possession" (of a low spirit entering the body of a living person) is also considered to be real. Possession is by ignorant and undeveloped spirits. But they are not evil; no being is basically evil. Each person can choose what spirits he will associate with, and they can harm him only in the sense that "unfitting companions" can harm him.

Spiritual healing is a strong emphasis. The value of drugs and surgery is freely acknowledged; but since man is so much more than just the body, the seat of disease is to be found largely in wrong mental and spiritual action.

Mediumship is not considered to be a supernatural gift, nor are spirits considered to be supernatural beings. There are no miracles. Everything happens according to natural law, both on earth and in the world to come. Throughout eternity man must strive to discover nature's laws, which are immutable, and to attune his life harmoniously with them.

Spirit messages are not thought to be infallible. No one should expect them to be infallible in view of their origin and the complex nature of the line of communication that they follow to reach the séance room. They originate with a spirit, who of course is not infallible. Then, in getting a message to earth, a spirit must contact a medium, or more probably first a medium's control. Neither the control nor the medium is thought to be infallible. Thus, the average message has several chances of being in error. Most of the messages deal with moral and spiritual concerns, and they conform to principles that are well known both on earth and in the spirit world. Spiritualists believe that messages that deal with "worldly affairs" are more likely to prove disappointing and misleading.

The conduct of séances

Worship. Séances and church services are the customary forms of worship for Spiritualists. Séances form around mediums, and the number of persons attending varies from one to a score, seldom more. The setting is usually a living room, the medium may or may not go into trance, and the room is usually darkened. Nonbelievers often criticize Spiritualists for conducting most of their séances in the dark. Spiritualists reply that the spirits find it easier to work in the dark than in the light, a fact considered no more remarkable than the greater strength of radio messages at night than in the daytime. Spiritualists believe that in the séance room the persons attending can communicate with the spirits simply by speaking to them. The spirits in turn, it is alleged, use different methods depending on the powers of the medium, such as rappings, table tippings, the voice of the medium, independent voices, automatic writings, the ouija board.

Church services are patterned after Protestant worship. They include prayers, hymns, musical selections, readings, collections, sermons, spirit messages, and benedictions. The readings make little use of the Bible; they are chiefly recent compositions dealing with spiritual subjects. The spirit messages, given by a message bearer (medium), are of the mental type of phenomena, and the bearer receives them in an undarkened room. The sermons often deal with the moral and spiritual concerns that are shared by most religious persons in the West.

Faith in the spirit origin of the communications seems to be largely confined to persons who have themselves served as mediums or else have received messages through mediums whom they trust. Some Spiritualist leaders acknowledge that fraud in the séance room is a recurring problem for the movement.

The Spiritualist, though he does not claim infallibility, believes that he is dealing with facts; he insists that his faith is constructed to conform to the facts, as contrasted with a change of facts to conform to faith.

BIBLIOGRAPHY. Books and pamphlets issued by the National Spiritualist Association of Churches include the *Spiritualist Manual* (rev. 1967); and the *Spiritualist Lyceum Manual* (rev. 1962), designed for youth. Publications of the Stow Memorial Foundation (an NSAC agency) include: B.F. AUSTIN, *The A.B.C. of Spiritualism* (1920); A. CAMPBELL HOLMES, *The Fundamental Facts of Spiritualism* (n.d.); and B. MORSE, *The "Whys" of Spiritualism* (1933).

Books by non-Spiritualists: G. LAWTON, *The Drama of Life After Death* (1932), is the best book-length study, definitive to its date. C.S. BRADEN, *These Also Believe* (1949), and J. STILLSON JUDAH, *The History and Philosophy of the Metaphysical Movements in America* (1967), contain excellent chapters. The latter describes some of the Spiritualist groups.

Books attempting to appraise Spiritualist and psychic phenomena: C.E.M. HANSEL, *ESP: A Scientific Evaluation*, with an introduction by E.G. BORING (1966), contains a chapter on Spiritualism. Hansel finds Spiritualist phenomena to be largely fraudulent. G. MURPHY, *Challenge of Psychical Research: A Primer of Parapsychology* (1961), does not accept the survival hypothesis, but acknowledges the existence of unexplained phenomena that have the appearance of communication with the dead. H.N. HART, *The Enigma of Survival: The Case For and Against an After Life* (1959), is a closely reasoned review of the evidence. The author reaches a conclusion favourable to survival. H. WESTWOOD, *There Is a Psychic World* (1949), is written by a clergyman who reaches conclusions favourable to the reality of spiritist phenomena.

(J.P.W.)

Sports, Amateur and Professional

A long-standing distinction in sports has been that between the professional and the amateur, and controversies over amateurism and professionalism, by no means new, have been growing in pace with growth of interest in sports as evinced by the increasing numbers of participants and spectators. Contributing factors in the controversies include the desires of manufacturers of sporting goods and accessories to promote their products; of promoters of sporting events to ensure the participation of topflight competitors; and of professional teams or sponsors to obtain the services of outstanding amateurs. Complicating factors include the application of different standards in different sports, and differences of standards and of interpretations of general standards among various countries. Questions of amateurism as they apply to the Olympic Games are dealt with in ATHLETIC GAMES AND CONTESTS. Pertinent information in relation to particular sports in which professionalism is or has been an issue will be found in the articles on those sports; *e.g.*, BASKETBALL; GOLF; TENNIS AND ALLIED COURT GAMES; and ASSOCIATION FOOTBALL (SOCCER).

The conflict between amateur and professional sports has often been compared to the distinction between love and money. The word amateur derives from the Latin *amator*, or "lover," and the amateur athlete by simplistic definition is one who competes for love of sport alone, while the professional competes for money. The problem with these definitions, as athletes have been discovering for centuries, is that it is hard to do the first without the second.

History. The burdens of pure amateurism were evident not long after the founding of the Olympic Games (about 776 BC). At first, when the city-states of ancient Greece sent only their aristocrats to the games, the Olympics were conducted in a spirit of amateurism. The aristocrats possessed both the money and leisure time to practice for their strenuous events. But as the influence of the merchant class began to pervade the games, the city-states took to subsidizing their best athletes. Money, food, homes, tax exemptions, and even army deferments were the awards athletes sought for winning their events (the original Olympians received only laurel wreaths).

The spirit of amateurism

Ultimately, the principles of amateurism were stretched

to intolerable limits. One Eupolus of Thessaly was convicted of bribing three boxing opponents in the 98th Olympiad. When the Romans conquered the Greeks, the Olympics fell into disrepute. The emperor Nero introduced such events as music for his own pleasure and managed to intimidate opponents wherever he chose to compete. Once, though he fell off a chariot, he was still declared winner of the race. When the Olympics were discontinued in AD 394, there were few cries of alarm.

Origins of concepts. The word *amator* was revived in 1788 and given the anglicized spelling "amateur" to separate "Gentleman Jack" Jackson, an aristocrat and a boxer, from the heavyweight bareknuckle fighters, who competed in Great Britain for side bets or small purses.

The word "professional" was first applied in the modern era to English jockeys who rode other men's horses for a fee. The nobles themselves had begun racing for awards in 1512, but soon hired hands were competing in their place.

First professionals

The concept of an amateur as a cultivated man of leisure and the professional as a money-seeking ruffian was standardized in 19th-century England. Beginning in 1812, an annual sports day was held at the Royal Military Academy in Sandhurst. Sports soon found their way into the curriculums of Rugby and other British schools, universities, and officer training schools. In 1866 the Amateur Athletic Club was founded

> . . . to afford as completely as possible to all classes of gentlemen amateurs the means of practising and competing against one another, without being compelled to mix with professional runners.

There came to be three categories of athlete: the professional, the gentleman amateur, and the amateur, who was distinguished from the gentleman amateur because he worked for a living. Democracy was introduced into the world of the gentleman amateur in the mid-1870s when W.J. Morgan, a walking champion and wage earner, had his entry accepted for a meeting organized by the prestigious London Athletic Club—over the objections of numerous members.

In spite of historical differences, the United States followed a similar concept of amateurism. As in England, the first amateur athletes were gentlemen learning their skills at New England boarding schools and Ivy League colleges. A professional might instruct a gentleman's son, but he left by the servant's door. Amateurs were rarely labelled as such, and many people, in fact, had no idea of the meaning of the word. Rather, professionals were defined as men who engaged in such sports as boxing, horseback riding, and sprinting for wages and cash prizes. Participants in such sports as lawn bowling and sailing were not paid and therefore were assumed to be amateurs. The first time the question of amateurism was raised in America was in 1872, when about half the oarsmen entered in a regatta on the Schuylkill River near Philadelphia were banned because they were "not amateurs." The nominations committee explained that anyone who competed for money or hoped to benefit financially through racing was not an amateur. The definition of an amateur as one who competes for pleasure and the professional as one who competes for money was accepted at a general meeting of rowing clubs in 1873.

The masses did take up sports on both sides of the Atlantic, but at first as second-class citizens. A late victim of the gentleman's code was a 1920 U.S. Olympic champion rower, John Kelly, who was banned from the Henley regatta because he was a bricklayer, not a gentleman.

When Baron Pierre de Coubertin revived the Olympics in 1896, he re-affirmed the traditional Olympic values. "The most important thing in the Olympic Games," he wrote, "is not to win but to take part, just as the most important thing in life is not the triumph but the struggle."

Amateur codes. An example of the general tenor of amateur codes is rule 26 of the Olympic regulations that defines an amateur as:

> one who participates and always has participated in sports solely for pleasure and for the physical, mental or social benefits he derives therefrom, and to whom participation in sport is nothing more than recreation without gain of any kind, direct or indirect. In addition he must comply with the rules of the International Federation concerned.

Differences between countries. The last sentence defines the major problem in the continuing conflict between amateurism and professionalism. The International Amateur Athletic Federation allows each country's national Olympic committee to define amateurism for its own athletes, resulting in widely diverging interpretations of the concept from country to country. The Soviet Union and other Communist nations, for instance, introduced a new category, the state amateur. Top amateurs are, in effect, subsidized by the state. There are no professionals as such, but leading amateurs are given jobs that allow them to train and travel at will. Many American athletes are also subsidized—at least insofar as they may receive full scholarships to college based primarily on their ability in a given sport. German athletes are aided by their Olympic association and French athletes by their government. An African Olympian may be a police lieutenant who spends most of his time training. Australian athletes on tour may be paid the same salaries they would be receiving at their regular jobs. This is known as "broken-time pay" a bone of contention over which the Rugby Football League in Great Britain broke away from the Rugby Football Union in the 1890s. In short, despite the restrictions of amateurism, most countries are able to send their leading athletes to international events.

The state amateur

Given the various interpretations of amateurism, it is not surprising that there have been disputes over alleged professionalism. At times, unmistakable professionalism has intervened. During the 1968 Olympics, for example, some athletes were paid, or offered payment, by track shoe manufacturers to wear their products.

Nationalism in the Olympics. To its defenders, amateurism connotes a state of mind as well as a condition. The Greek Olympics were originally conducted in a spirit of cooperation and brotherhood that crossed the political boundaries of the city-states. If professionalism erased that ideal in the original Olympics, the modern Olympics has never been able to avoid the nationalism that has affected competition. Point scores are unofficial (the contests are between individuals and teams, not nations), but the public, press, officials, and politicians follow them zealously. Since World War II the Olympics have been treated as an extension of the cold war between the United States and the Soviet Union. Before the 1968 Olympics Vice President Hubert H. Humphrey stressed the importance of American athletes winning more medals than the Russians.

Differences between sports. Just as amateurism has varied from country to country, amateurism has varied from sport to sport. Tennis amateurs are allowed to receive expenses but golfers are not. Yet amateur golfers are allowed to compete against professionals, while tennis players until recently were not. In track, there are at present practically no designated professionals. Track athletes receive expenses to compete in meets away from home, but under-the-table payments may often amount to something more like a living wage. A number of European ski amateurs have been suspended for instructing at clinics and accepting payment. "The only amateur I know," said one skier, "is the Agha Khan."

Critics and defenders. Critics of the amateur code contend somewhat facetiously that the only difference between amateurs and professionals is that amateurs demand payment in cash. The concept of a "gentleman amateur" is obsolete, they argue, when a modern athlete is nurtured under the belief that he will progress as far as his ability, not social origin, can take him; but the athlete who is not rich cannot progress without subsidization. Critics also contend that the amateur code is inconsistent and hypocritical. They wonder, for instance, if the famous American Indian Olympian Jim Thorpe, who lost his gold medals because he had played baseball for a reported $60 a month, was any more of a professional than the subsidized college amateur.

Defenders of the code maintain that without controls the differences between amateurs and professionals will

cease to exist. They point out that expensive catering to leading amateurs has weakened sport at the participant level. Frequently heard in many countries is the refrain, "We have become a nation of spectators." The traditionalists, who believe passionately in the original spirit of the ancient Olympics, contend moreover that the desired position of sports as a respite from the problems of mankind has been mitigated by professionalism and nationalism. "The ancient Greeks used to stop fighting to stage the games," a former president of the International Olympic Committee, Avery Brundage, once said. "Now we stop the games to stage our wars."

Signs of change. A number of developments, however, indicate that amateurism will continue to evolve from its roots. John Kelly, Jr., the son of the bricklaying rower and in 1971 the president of the Amateur Athletic Union (AAU) of the United States, took office saying,

> The old concept of only the gentry being allowed to compete and not the average citizen has to go. We are very hypocritical in this country.

Kelly also made peaceful overtures to the rival National Collegiate Athletic Association (NCAA). Agreement could yield a liberalized amateur code in America.

Athletes themselves have begun protesting restrictions, occasionally with success. Many European skiers have threatened to boycott the winter Olympics if amateur rules are not revised. Track athletes have asked permission to capitalize on their names in their professions and have suggested that track clubs be allowed to negotiate with private sponsors for financial backing. In international tennis, "open" competition between amateurs and professionals has been instituted, led by the British. Open tennis appears to be a sign of a similar trend in other sports. In golf, where amateurs are forbidden to accept expense money, changes that would allow for some expense money and free equipment have been suggested, especially since a United States amateur champion was suspended for three months in 1969 for accepting six dozen golf balls from a manufacturer.

De Coubertin himself saw the need to pay amateurs. In his memoirs, he contended that it was as unwise to declare an athlete a professional for accepting money as to call a sexton an infidel for being paid to take care of a church.

As a result of spiralling costs and insufficient income, many colleges in the United States are cutting back on athletic scholarships (or granting them solely on the basis of need, as other scholarships are granted). Some schools have relinquished athletic scholarships altogether and reverted to club sports, which have returned these colleges' programs to a level that is strictly amateur, both in finance and performance.

It appears that on one level the serious and talented amateur may be granted greater flexibility. On another level, the subsidized athlete of marginal ability may some day be eliminated. And on a third level, those of ordinary ability will continue to play sports for fun.

Problems of professionals. The professionals have not been spared their crises, either. Once regarded as a pastime, in which owners were considered kindly patriarchs and athletes willing gladiators, professional sport, for all practical purposes, existed outside the scope of the law.

Unions of athletes

Since World War II, professional sport has been regarded as a business; athletes have unionized and management has been taken to court for alleged antitrust violations; and labour laws in some countries—in Mexico, for example—have been enacted or revised to cover sports.

Expansion. Like many businesses, professional sport has undergone expansion most rapidly during the 1960s. In the United States and Canada, baseball, which once consisted of 2 leagues with 8 teams apiece, expanded into 4 divisions with 24 teams; the National Basketball Association expanded from 2 divisions of 4 teams apiece to 4 divisions with 17 teams, and a rival league, the American Basketball Association, was formed; ice hockey, too, expanded, from its once comfortable American-Canadian format of a single 6-team league to two 7-team divisions; and professional football merged its two competing associations and opened the 1970s with two 3-division associations under the control of the National Football League. There is also professional football in Canada.

At the same time, sports once regarded as minor are achieving national and international recognition at the professional level, thanks to widespread interest among fans and sponsors. Tennis, for the first time, has a regular series of tournaments—almost yearlong and touching down in such places as Iran, Italy, and Australia. Women's tennis has experienced similar growth. Purses are larger than ever in automobile racing, which continues its Grand Prix competition at the international level. Association football, or soccer, is finally recognized for what it always was—the world's most popular sport. The World Cup matches every four years attract widespread international attention. Even the only major sport with exclusively American roots—basketball—now has professional teams in Europe.

Manufacturers and sponsors. One reason professional sports have been able to expand has been the influence of manufacturers and sponsors. In golf and tennis, for instance, corporations finance the costs of and sometimes even assume the names of tournaments (Kemper Open golf, Virginia Slims tennis in the U.S.; Dunlop golf in Australia). Sponsored skiing competitions include the Evian World Cup. Auto races and racers are financed by automobile, tire, oil, and additive companies. Similarly, motorboat companies finance motorboat racers. Individual sponsors are prominent too, as in golf.

Television. But nothing has affected sports more dramatically than the ubiquitous eye of television. Television has sustained whole leagues, created franchise shifts, crippled some sports, sustained others, and even affected the rules of the games themselves. Television helped create expansion, too. In the United States, for example, there were only 42 professional sports franchises in 1960, mostly in the northeast quadrant of the country. Television exposure created instant interest elsewhere, and by 1970 there were 87 franchises throughout the nation.

Thanks to television exposure, athletes have become as marketable as the sports they play. Now some players, Brazil's Pele perhaps being the prime example, make as much money through promotions and business income as they do playing games.

The games themselves are affected by commercial timeouts, which have the potential to affect "momentum"—and scores. Once universally condemned, the commercials are now widely ignored. Even the formats of important sports events have been altered to suit the convenience of television. Games have been played at unusual hours to accommodate prime-time broadcasting. The U.S. Professional Golfers Association tournament was switched from its long tradition of man-to-man play to 72-hole medal competition in part because the latter format better suited television. Tie breakers for tennis sets have been adopted for the same reasons.

Television's impact on sports has not been entirely beneficial. Overexposure almost killed professional boxing. Minor league baseball in the U.S. dried up when television and major league expansion usurped its crowds and income.

Conclusions. Analysis of the available evidence indicates that professional sports will continue to grow; it is likely that recompense to athletes of national- and world-class calibre will be liberalized; and continuing and perhaps increased efforts may be required on the part of those interested in amateur sports for the masses—*i.e.*, those who may enjoy and benefit from taking part in sports as a leisure-time activity.

BIBLIOGRAPHY. LEONARD SHECTER, *The Jocks* (1969), raises some interesting points on the abuses of amateurism; PAUL WEISS, *Sport: A Philosophic Inquiry* (1969), discusses the ambiguities of amateurism and professionalism under their traditional definitions. See also F.G. MENKE (ed.), *The Encyclopedia of Sports*, 4th rev. ed. (1969), esp. the chapter, "The Amateur Athlete"; H.W. MORTON, *Soviet Sport* (1963); MELVYN WATMAN, *History of British Athletics* (1968); and WILLIAM JOHNSON, "TV Made It All a New Game," *Sports Illustrated*, 5 pt. (Dec. 22, 1969–Jan. 26, 1970), on the influence of television on sports.

(J.L.K.)

Springs and Wells

Springs and wells are generally openings for discharge of water from underground sources. A spring is a natural discharge with a perceptible current at the land surface or in the bed of a stream, lake, or sea; water that emerges at the surface without a perceptible current is called a seep. Springs and seeps represent water reappearing from the groundwater phase of the hydrologic, or water, cycle. A well, on the other hand, is a man-made hole excavated in the earth either to bring water or other fluids to the surface or to store fluids underground. From ancient times people have dug wells or pits to gain access to shallow groundwater, and in many places they have constructed wells from which water flows at the surface as from a spring.

Excavations for wells have attained depths exceeding two kilometres (1.2 miles) for water and to seven kilometres (4.4 miles) for oil and gas. Wells are also used for the extraction of valuable products such as oil, tar, sulfur, or soluble minerals; for the temporary storage of gas or other fluids or the permanent disposal of fluid wastes underground; and for the artificial replenishment of groundwater. Although wells have been constructed for a variety of purposes, the wells most numerous and widely distributed are those that yield water for use in homes, agriculture, and industry. Springs and seeps are widely distributed over the continents and some islands. Most wells withdraw water that has percolated down into groundwater reservoirs called aquifers, or water-bearers.

PHYSICAL CHARACTERISTICS

Water supply. Water in springs and wells generally originates as precipitation from the atmosphere. In humid regions the average annual precipitation exceeds the loss by evaporation from soil and vegetation and the surplus flows away as surface water or sinks as groundwater or both. In arid regions the loss to the atmosphere may exceed the average precipitation, and the surface water and groundwater are less abundant.

Some of the water from rain or snow becomes groundwater when the conditions of climate and geology are favourable. Rocks and soils that absorb water from storms faster than the soil can retain it are favourable for groundwater replenishment. Less favourable are soils that absorb water but retain most of it for subsequent evaporation or use by vegetation. Least favourable are land surfaces so impermeable that all rain either runs off immediately or collects in shallow pools and evaporates rapidly.

Technically, only water that percolates downward to reach a zone in which all rock pores are saturated is called groundwater. Throughout the saturated zone, the movement of goundwater is directed by geological conditions of the ground layers. Some saturated rock layers, or beds, are permeable enough to transmit water readily to springs and wells. The best of these aquifers are composed of gravel, sand, limestone, sandstone, or basalt. While other types of rocks may yield some water to wells and springs under favourable conditions, most of them are not good water-bearers. Some are too dense to allow water flow, and others have pores so small that little water can move through them. In regions having good aquifers, denser materials, such as clay or shale, may serve to confine water in the aquifers. In some places water so confined is under pressure sufficient to rise through an overlying layer to the land surface to form an artesian spring or artesian well.

The best water-bearing layers

Waterflow and temperatures. The aquifers that receive and discharge the largest quantities of groundwater are the unconsolidated materials (including gravel and sand) that nearly everywhere mantle the surface of the earth. The natural discharge of groundwater from such materials is partly by springs and largely by evaporation and seepage. The great majority of wells obtain their water from gravel and sand aquifers, especially those underlying valleys, lowlands, and plains.

Limestone and dolomite are sedimentary rocks composed predominantly of calcium and magnesium carbonates, which are soluble in water. Some of these layers are dense, others permeable; the permeability of many layers has been increased by the dissolving and subsequent removal of some of the original rock, leaving caverns and channels through which water may flow. The natural discharge from carbonate rocks forms many springs, including some of the largest on earth. The limestone spring area with greatest reported discharge is the Ra's al 'Ayn, a cape along the Khābūr River. A tributary of the Euphrates, its surface flows begin in Turkey and Syria and have a combined discharge averaging 42 cubic metres, or about 11,000 gallons, per second. (In the U.S., flow is generally given in gallons per minute, which would be 60 times the gallon per second figure, used throughout this article for consistency with metric units.)

In some limestone rock formations, rainwater disappears rapidly into holes or other openings and drainage is chiefly by subterranean channels. Some springs are actually underground rivers. One of these, the Trebišnica River in Yugoslavia, emerges five times in an otherwise subterranean course of 100 kilometres (about 60 miles). Such springs undergo marked increases in discharge and, perhaps, turbidity in response to rainstorms. The Fontaine-de-Vaucluse, the largest limestone spring in France, ranges from eight to 150 cubic metres, or about 2,100 to 39,600 gallons, per second, averaging 23 cubic metres, or about 6,100 gallons, per second. Submarine springs occur in several seas, generally issuing from limestone aquifers.

Aquifers in volcanic rock

Basalt, a volcanic rock that may form extensive plains, has about the same range of permeability as limestone. It is derived from lava flows, some of which are quite dense, in which solidification and subsequent weathering have produced very permeable aquifers. Several large springs rise in basalt along both flanks of the Cascade Range in the northwestern United States, and rivers draining these mountains have spring-fed tributaries with remarkably uniform flow. Large and spectacular springs discharge from the basalt aquifer underlying the Snake River plain in southern Idaho, which is exceptionally permeable and noted for wells of large yield.

The active volcanoes of Hawaii—Mauna Loa and Kilauea—are made up of basalt so permeable that even in the heaviest rainfall there is little water runoff. Impermeable materials hold fresh water in subterranean reservoirs on most of the Hawaiian Islands. Along the southern coast of Oahu, for example, a layer of clay and limestone rests above the basalt and confines freshwater under artesian pressure. For decades wells southeast of Oahu have withdrawn more than ten cubic metres, or about 2,600 gallons, per second from this aquifer, and nearly four cubic metres, or about 1,000 gallons, per second is discharged naturally by springs above sea level in the Pearl Harbor area.

Sandstone aquifers, which are sources of water in many springs and wells, give rise to very large springs only under exceptional conditions. A buried channel and prominent jointing are factors in discharges averaging 18 cubic metres, or about 4,750 gallons, per second of the Giant Springs along the Missouri River in Montana, called the largest fountain in the United States. The Lewistown Big Springs, whose four-cubic-metre-, or about 1,000-gallon-, per-second discharge provides the water supply of Lewistown, Montana, issue from sandstone that was dammed by a fault, or earth fold.

Springs, in fact, can be classified on the basis of their rate of flow from first magnitude, average discharge exceeding 100 cubic feet per second (2.83 cubic metres), or about 750 gallons, per second, down to eighth magnitude, emitting less than a pint a minute. At most, there are only a few hundred first-magnitude springs in the world, and the great majority of them flow from limestone or lava, mostly from cavern-like openings. Free and direct communication exists between the points of intake and of discharge, and the water has generally not reached depths greater than a few hundred metres (or yards) below the surface.

Springs are often classified on the basis of their water temperature. A thermal spring has a temperature significantly higher than the average air temperature of its surface surroundings. Among the largest thermal springs

are Warm Springs in Montana and Ana River Springs at Summer Lake, Oregon, both of which are first-magnitude springs that discharge water at about 20° C (68° F), significantly warmer than the air. Thermal springs are common in areas where there was volcanic activity in prehistoric times and in areas where rock layers have been broken and folded in geologically recent time. A cold spring has a temperature below the average for the region, as exemplified by Clear Lake in Oregon, where one spring discharges five cubic metres per second at 4° C (1,300 gallons per second at 39° F), which is 6° below the air temperature.

Quality of groundwater. Factors that determine the quality of water yielded by a spring or well include: the type of aquifer, or host rock, that yields the water, as well as the rock materials through which the water has passed; the temperatures along the route of subterranean travel; and the volume of water circulating, past and present.

Contributions of the aquifer

Water from precipitation is least modified underground where it travels for only short distances, shallowly and through permeable materials depleted of soluble minerals by previous water flow. For this reason, many gravel and sand aquifers, particularly in humid regions, yield water of excellent quality. In arid regions, on the other hand, spring and seepage areas may accumulate undesirable mineral residues as water evaporates. Shallow groundwater is especially vulnerable to degradation by chemical fertilizers and detergents, salt used in snow removal, waste from septic tanks, refuse from garbage dumps and land fills, and waste from industrial pits and ponds.

The quality of water in sedimentary rocks, most of which are of marine origin, depends upon the extent of freshwater flushing through such rocks. After the seawater has been washed away, many carbonate rocks and sandstones yield freshwater of good quality, though it may be hard because of its calcium and magnesium content. Many sedimentary rocks, however, are buried too deeply—and most shales are too impermeable—to be reached by freshwater from precipitation, and these still contain seawater. Of the numerous springs yielding saltwater from sedimentary rocks, one of the largest is Salt Springs in Florida, which discharges five cubic metres, or about 1,300 gallons, per second, containing about 5,400 milligrams of dissolved salt per litre (0.2 ounce per quart), chiefly sodium chloride, or common table salt (compare with seawater, at 35,000 milligrams per litre).

Many spring waters issuing from deep sedimentary rocks are mineralized: some are only slightly salty, whereas others are several times as saline as seawater.

In deep wells and mines the temperature generally increases 1° C for each 30 to 60 metres (100 to 200 feet) of depth. In the Australian Great Artesian Basin, however, the temperature rises about 1° per 20 metres (65 feet) of depth; wells 1,400 metres deep (about 4,600 feet) yield water at nearly 100° C (212° F), which must be cooled to be used for watering stock. The water has about 1,400 milligrams per litre of dissolved solids, chiefly the so-called black alkali, sodium bicarbonate, undesirable for irrigation. The natural discharge from this basin was by springs that built up large mounds of calcareous deposits.

Thermal springs are common in nonvolcanic regions where there has been faulting or intense folding of the rocks. One of the largest such hot springs in the world, the Big Horn Hot Springs at Thermopolis, Wyoming, discharges one cubic metre (265 gallons) per second of water at 56° C (135° F), and with 2,600 milligrams per litre dissolved solids, chiefly calcium bicarbonate, it forms quite beautiful and extensive travertine deposits. At the edge of the Sahara, springs at Nefta and Tozeur, in Tunisia, are of equal magnitude, but the temperature is less than 30° C (86° F), and the water, with less than 400 milligrams per litre dissolved solids, is excellent for irrigation.

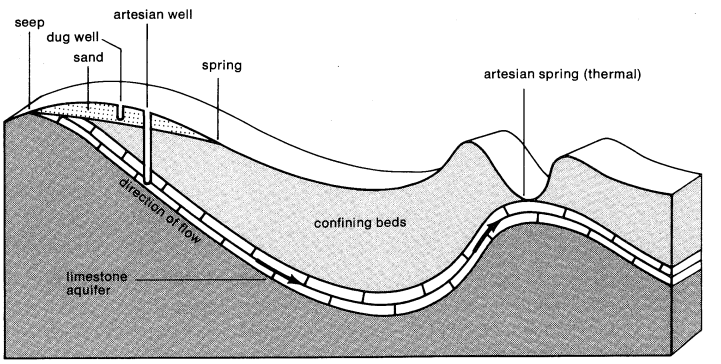
Volcanic hot springs

Most thermal springs, however, are in volcanic regions where molten rock is near enough to the surface to be a formidable source of heat. Thousands of such springs occur in Iceland. The largest volcanic spring in the world, at Deildartunguhver, Iceland, is a third-magnitude spring that discharges 250 litres, or about 65 gallons, per second. The combined discharge of thermal waters in Yellowstone Park, Wyoming, is about three cubic metres, or almost 800 gallons, per second, barely enough, if concentrated in one outlet, to form a first-magnitude spring.

Volcanic eruptions commonly generate steam and more than 90 percent of the gases expelled from surface fissures or holes is water vapour. These gases presumably come from molten rock, as indicated by high temperatures at the surface, up to 700° C (about 1,300° F), and high geothermal gradients to considerable depths. In near-boiling springs, carbon dioxide is dominant, and hydrogen sulfide and nitrogen are common among dissolved gases. A porous siliceous crust is a common deposit at the surface. The spring waters also include dissolved compounds of boron, fluorine, lithium, and arsenic, probably from molten rock material within the earth. Spring gases, high surface temperature, and geothermal gradient are all conducive to development of geysers, of which there are about 200 in Yellowstone Park, 30 in Iceland, and others in Alaska, Kamchatka, and New Zealand. The waters in geyser areas are generally alkaline and rich in sodium chloride; in other volcanic areas they are acidic and rich in sulfates. Some hot springs are radioactive with radon, stinking with hydrogen sulfide, or poisonous with arsenic.

Most of the thermal water discharged at the surface in volcanic regions is probably derived from groundwater from local rainfall, the contribution of water from the molten rocks being no more than 5 percent. Many springs in volcanic regions are potable and usable for irrigation. The El Gato Spring, south of San Luis Potosí in Mexico, flows about five cubic metres (about 1,300 gallons) per second of radioactive water, at 41° C (106° F), with 500 milligrams per litre dissolved minerals. Other waters from volcanic rocks are far more mineralized: in Israel the springs along the shore of the Sea of Galilee, which served the Roman baths at Emmaus, flow from basalt at 60° C (140° F) and contain 28,000 milligrams per litre of dissolved solids, chiefly sodium and calcium chlorides; and in Turkey near the Mediterranean Sea, the Gayzer suyu wells up from volcanic rock at 100° C (212° F) with 63,000 milligrams per litre of dissolved solids, chiefly sodium chloride.

Groundwaters of highest salinity are associated with salt deposits. Some saline thermal springs discharge water that has passed through salt deposits at considerable depth, exemplified by the Glenwood Springs in the canyon of the Colorado River in Colorado, which flow 200 litres, or about 50 gallons, per second of sodium chloride brine at 52° C (126° F). The Aqua Salada de Timban in Argentina yields tepid water with 317,000 milligrams per litre of dissolved salts. Wells in several countries obtain brines of similar concentration. (H.E.T.)



Ground water flow to springs and wells.
Dug wells tap unconfined water and artesian wells tap confined water. Springs occur at the contact of aquifer and less permeable rocks. This artesian spring is thermal because of the water depth. A seep is groundwater that appears at the surface.

LIFE IN SPRINGS

The environmental setting. Compared to other surface waters, springs offer a remarkably uniform habitat. The temperature varies only a few degrees between summer

Constancy of conditions

and winter, and the chemical composition and volume of flow are almost constant. Springs arising from deep layers exhibit the greatest uniformity. Seasonal change in light intensity is the greatest single factor influencing spring communities.

Many spring organisms are restricted to a small area at or just downstream from the point of emergence of the water, where conditions are most constant. Farther downstream the water increasingly resembles that of drainage streams. Except for hot springs, many springs are at the average temperature of the region in which they occur; spring water is colder than normal surface waters in summer but warmer in winter. The emerging groundwater of many springs is often deficient in oxygen but rich in carbon dioxide and other gases until aerated at the surface. If the oxygen-deficient water contains iron in solution, ferric hydroxide (ochre) will form as the water becomes oxygenated and filamentous bacteria build the sheaths that contain the iron precipitates.

The physical and chemical zonation in springs is accompanied by a corresponding zonation of plants, which, in turn, determines the animal life of the spring. Depending on the flow and the local climate, typical spring–brook communities extend from 10 to 500 metres (33 to 1,640 feet) downstream. Thus, springs are important ecological study areas because of their relative constancy near the source and their downstream gradation.

The biotic component. Spring communities typically have few species but many individuals. Three kinds of springs can be distinguished on the basis of the type of water emergence: rheocrenes, in which the water flows rapidly away from the source; limnocrenes, in which the water issues at the bottom of a pool before overflowing as a spring brook; and helocrenes, in which water seeps slowly from the ground. Rheocrenes resemble mountain brooks; they have rocky bottoms, and their organisms are adapted to living in swiftly moving water. Limnocrenes and helocrenes contain organic-rich deposits of mud or sand in which plants thrive; their communities resemble those of ponds or stream pools. The animal life of springs is continuous with that of the groundwater runoff. In limestone areas, cave animals frequently appear in springs; most such occurrences are accidental following flooding, but some animals regularly migrate between the two habitats.

Springs vary in biological productivity. Large mineral-rich springs are usually very productive; they are termed autotrophic, which means they support autotrophs, or producers, such as algae and rooted plants, which, in turn, become food for many small aquatic animals. Small soft-water springs are generally of low productivity; they are called heterotrophic because they support primarily heterotrophs, organisms that derive their energy by consuming other organisms; in this case, they subsist on leaves and other terrestrial material that falls into the spring. In hot springs, only heterotrophic bacteria occur above about 75° C (167° F); below 75° C, autotrophic blue-green algae form conspicuous mats that support bacteria and, below about 50° C (122° F), a few specialized animals.

Spring organisms characteristically are small. Even in wide-ranging species, individuals inhabiting springs are smaller than members of their species in other habitats. Size reduction in spring organisms may be a response to limited food and space. Spring representatives of cave species have degenerate eyes, loss of pigment, and well-developed tactile organs. Cave animals that occur in springs include flatworms, annelid worms, snails, isopods, amphipods, crayfishes, and fishes.

The spring as sanctuary

Springs often provide refuge for aquatic organisms following climatic changes. When extensive lakes have evaporated, isolated populations of fishes and other organisms have often managed to survive in their springs. Isolation gave rise to unique species such as a minnow represented by only a few hundred individuals in a single spring hole in Death Valley, California.

Hot-spring organisms are able to withstand both high temperatures and severe chemical conditions. The most conspicuous representatives are species of blue-green algae, whose slimy or matted growth is often tinted orange or yellow when exposed to high light intensities in the shallow outflow channels. The algal mats can accumulate because of the scarcity of decomposer bacteria and the total absence of grazing organisms in waters hotter than 45° C (113° F).

Although periods of growth and reproduction are extended in the uniform spring environment, biological activity often responds to seasonal changes in light. In Icelandic hot springs, for example, the low light intensity in winter inhibits algal growth even though water temperatures remain favourable. In the constant flow of Silver Springs, Florida, some of the most abundant invertebrates and fishes exhibit breeding cycles that are regulated by seasonal changes in daily light duration. In the spring, photosynthetic production is two to three times greater than in winter because of light differences.

While many organisms can tolerate the constant high temperatures of hot springs for varying periods, only a few can carry on an active life. The highest temperatures at which growth has been demonstrated for representative hot spring organisms is: bacteria, 95° C (203° F); blue-green algae, 74° C (165° F); protozoa, 50° to 58° C (122° to 136° F); insects, 50° C; fish, 50° C.

As already indicated, communities in autotrophic springs depend on the photosynthetic production of aquatic plants and algae. In contrast, the communities of heterotrophic springs depend on photosynthetic production of the surrounding terrestrial vegetation. Many springs, like the large limestone springs, are in a natural balance, with rapid growth of organisms and rapid consumption, resulting in a constant standing crop. Silver Springs, Florida, for example, is one of the most productive ecosystems known, its rate of total primary production by plants comparing with that of estuaries, coral reefs, moist forests, and intensive agriculture. It greatly exceeds the productivity of lakes, the oceans, coastal seas, and ordinary cultivated crops.

The high productivity of autotrophic springs is based on a constant favourable temperature, continuous supply of nutrients and removal of wastes, and the extreme clarity of water, which allows maximum light penetration. The small number of species, and consequent lack of diversity however, make spring communities especially vulnerable to disruption by introduced exotic species. (K.V.S.)

UTILIZATION

In order to obtain the benefits of a continuous flow of potable water, many human settlements have developed within easy access of springs. Many other locations remote from springs, streams, or lakes are underlain by aquifers that yield potable water to wells.

Historical uses. Springs and wells are especially valued in arid regions, and from ancient times life in deserts has centred about such watering places, or oases. Al-Khārijah, in the western desert of Egypt, probably the City of Oasis mentioned by Herodotus, was occupied by Stone Age peoples drawn there by the springs; now it has spring mounds but no longer any springs, and artesian wells provide water for the present inhabitants and their crops. In most of the so-called Fertile Crescent, between the Mediterranean Sea and Tigris River, water supplies are variable and precarious, but several localities have long had a consistent supply from springs. One of these is Damascus, Syria, reputed to be one of the oldest cities.

The tapping of springs

From ancient times cities have constructed long aqueducts from springs to meet their needs. Between 312 BC and AD 226, Rome completed 11 aqueducts, mostly underground but partly on arches, from spring sources 15 to 90 kilometres (9 to 56 miles) from the city; these water lines had a combined capacity of nearly ten cubic metres, or 2,640 gallons, per second. In North Africa during the 2nd century, Emperor Hadrian constructed a 70-kilometre (44-mile) aqueduct from the Zaghouan Springs to Carthage; the same springs now furnish part of the municipal supply of modern Tunis, Tunisia. After AD 330, rain-filled cisterns and local streams in Constantinople became inadequate for the city's increasing population; aqueducts were completed in 368 and 396 to deliver wa-

ter from springs in the Belgrade Forest, and others were constructed centuries later. Some of these aqueducts are still in use, but the city of Istanbul, as it is now called, currently relies also on surface reservoirs and wells.

Contemporary uses. Until 1930 Tehrān, the capital of Iran, obtained its entire water supply from 12 wells or *qanāt*s, which discharged 800 litres, or about 212 gallons, per second for its 200,000 inhabitants; reservoirs now serve the city. A *qanāt*, or karez, consists of a shaft dug to the water table, perhaps as much as 100 metres (330 feet) deep, in an upland area, and a tunnel with gradient sufficient for water to flow and long enough (commonly several kilometres) to bring the water to the surface on lower land. More than 2,400 years ago Darius encouraged the construction of *qanāt*s by exempting from tax for five generations the revenue produced by new *qanāt*s. Currently an estimated 40,000 *qanāt*s in Iran discharge about 600 to 700 cubic metres, or 158,000 to 185,000 gallons, per second, chiefly for irrigation.

Millions of wells supply water for drinking and domestic use, particularly in sparsely populated areas. Individual farms, villages, and towns depend chiefly on wells for water supplies. Even some cities, with populations ranging from 300,000 to 500,000 inhabitants, rely upon wells.

Most of the water pumped from wells is used for irrigation, especially in arid regions where evaporation leaves seasonal or year-round deficiencies in soil moisture and a short supply of surface water. Irrigation permits spacing of wells to distribute water over much of the area underlain by a productive aquifer. Drilled wells with power-operated pumps commonly yield 20 to 300 litres, or about five to 80 gallons, per second.

Some wells and springs are utilized as sources of energy in plants that utilize the flow from springs to generate electricity. The water from hot springs and wells in many resort areas is used for heating buildings. Geothermal heat supplies only a minor fraction of present world need for heat energy.

The value of dissolved matter

Some springs contain dissolved products of economic value. Radioactivity and mineral content contribute to the alleged therapeutic properties of some hot springs, and elaborate systems of storing and using the water are often built for such purposes. Other thermal springs are related to ore deposits. Some mercury and manganese deposits are clearly of hot-spring origin, and several ore deposits have been formed by solutions related to volcanism. Many wells have been drilled to obtain mineral products including common salt and potassium chloride. Sulfur also is mined by utilizing wells to inject water hot enough to melt the sulfur and force it to the surface.

(H.E.T.)

BIBLIOGRAPHY

Physical features: S.N. DAVIS and R.J.M. DE WIEST, *Hydrogeology* (1966), a college-level textbook of groundwater geology; D.K. TODD (ed.), *The Water Encyclopedia* (1970), a brief booklet on worldwide water facts and tabulations; G.A. WARING, "Thermal Springs of the United States and Other Countries of the World," Prof. Pap. *U.S. Geol. Surv. 492* (1965), a detailed summary and catalogue of springs, with references; EDWARD E. JOHNSON, INC., *Ground Water and Wells* (1966), a general treatise, especially valuable in the construction and operation of wells.

Biological features: G.A. COLE, "Desert Limnology," in G.W. BROWN, JR. (ed.), DESERT BIOLOGY, vol. 1 (1968), a technical review of the literature of biological, chemical, and physical interactions in desert waters; G.K. REID, *Ecology of Inland Waters and Estuaries* (1961), a readable discussion of these aquatic environments, including physical, chemical, and biological conditions; FRANZ RUTTNER, *Grundriss der Limnologie*, 3rd ed. (1962; Eng. trans., *Fundamentals of Limnology*, 1963), a most understandable textbook on physical, chemical, and biological aspects of inland waters; R.L. USINGER, *The Life of Rivers and Streams* (1967), a highly readable, beautifully illustrated account of streams from source to mouth.

H.E.T./K.V.S.

Sri Lanka (Ceylon)

Sri Lanka (known as Ceylon until 1972) is an independent island state in the Indian Ocean lying off the southeast coast of peninsular India. The Republic of Sri Lanka (Ceylon) has a parliamentary form of government and is a republic within the Commonwealth of Nations. The island is pear-shaped, with a maximum length of about 272 miles and a maximum width of about 140 miles; it has an area of 25,332 square miles (65,610 square kilometres). A shallow sea, Palk Strait, separates Sri Lanka from India; Talaimannar, on Sri Lanka's Mannar Island, is only 22 miles away from Dhanushkodi on India's Pāmban Island. A line of shoals, or sandbanks, called Adams Bridge, runs between Sri Lanka and India immediately south of this point. The capital of Sri Lanka is the port city of Colombo.

Since gaining independence from the United Kingdom in 1948, Sri Lanka has been striving to secure economic stability—a problem accentuated by a rapid population growth. The population has doubled in the past 25 years to its present total of about 13,000,000 and is increasing. This population explosion is the cause of a severe food shortage (nearly one-third of the food consumed has to be imported) as well as of unemployment and social unrest. The economy is overwhelmingly dependent upon the export of tea (which represents more than 65 percent of all Sri Lanka's foreign exchange earnings), rubber, and coconut products and is thus subject to the price fluctuations that characterize the world commodity markets.

The Sinhalese and Tamil linguistic groups

Sri Lanka is inhabited by a varied mixture of peoples who have immigrated to the island at intervals since remote times. The population, while consisting of nationals of Sri Lanka, is divided into several groups. The two principal linguistic groups are the Sinhalese, who generally are Buddhists and number over 9,000,000, and the Tamils, who generally are Hindus, numbering more than 2,500,000. Muslims (misnamed Moors), who number over 850,000, also speak Tamil. Relations between Sinhalese and Tamils have deteriorated in recent years, more as a result of current economic maladies than for historic reasons. Linguistic and religious groups—as well as the caste groups into which both Sinhalese and Tamils are divided—vie with one another to secure economic and political advantage, using pseudo-historic race myths and concepts in their struggles. Sri Lanka's ethnic and linguistic plurality is reflected in the design of its flag, which shows a yellow lion—symbolizing the legendary origin of the Sinhalese from a lion—on a red background, with a saffron stripe added to represent the Tamils and a green stripe to represent the Muslim minority.

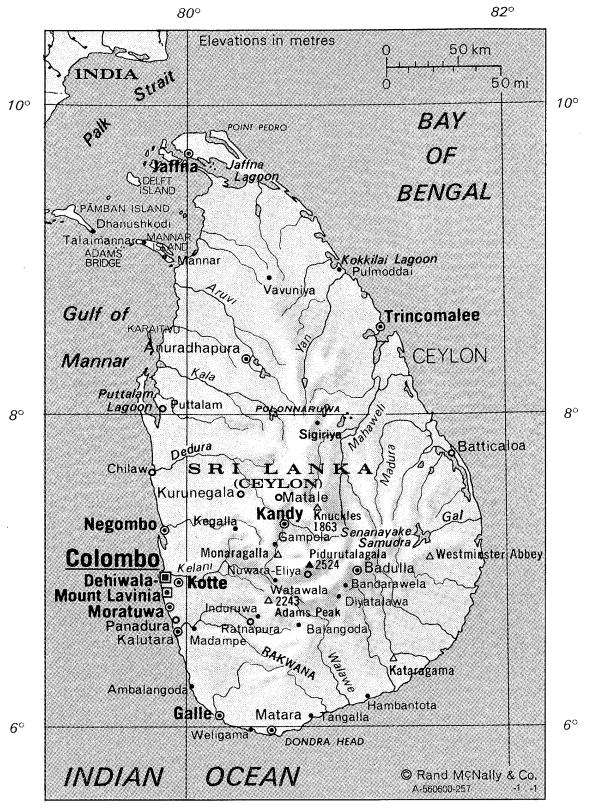
Sri Lanka is a country of great beauty, with scenery varying from the palm-fringed beaches of its coasts to the highlands of the interior, where the famous Buddhist Temple of the Tooth is located at Kandy, once the capital of an independent kingdom. Sri Lanka has also been traditionally associated with elephants, but their numbers have decreased substantially in recent years. The name Ceylon is an English word derived from the Sanskrit Siṃhaladvīpa, meaning "Island of the Sinhalese." The Sinhalese name for the country is Laka, the Pāli name is Laṅkā and the prefix Sri is an honorific. The Tamil name is Īḻam or Ilaṅkai. (See CEYLON, HISTORY OF for historical aspects; for more detailed information on cultural aspects, see SOUTH ASIAN PEOPLES, ARTS OF; see also the city article COLOMBO.)

THE LANDSCAPE

The natural landscape. *Relief.* Like peninsular India, of which it forms a continuation, Sri Lanka is a fragment of an ancient southern continent called Gondwanaland by geologists.

Both peninsular India and Sri Lanka stand on the same continental shelf, the edge of which does not extend beyond 12 miles from Sri Lanka's coastline. This shelf is shallow, lying at a depth of between 30 and 40 feet in the Palk Strait, and attaining a maximum depth of only about 210 feet. Beyond the edge of the continental shelf the depth increases to about 3,000 feet within two miles.

The relief of the island is dominated by a highland massif (mountainous mass) situated in the southern central part of the island. This is surrounded by an inter-



SRI LANKA (CEYLON)

MAP INDEX

Cities and towns

Ambalangoda.... 6·14n 80·03e
Anuradhapura... 8·21n 80·23e
Badulla..... 6·59n 81·03e
Balangoda..... 6·39n 80·42e
Bandarawela.... 6·50n 80·59e
Batticaloa..... 7·43n 81·42e
Chilaw..... 7·34n 79·47e
Colombo..... 6·56n 79·51e
Dehiwala-Mount Lavinia..... 6·51n 79·52e
Diyatalawa..... 6·48n 80·57e
Galle..... 6·02n 80·13e
Gampola..... 7·10n 80·34e
Hambantota..... 6·07n 81·07e
Induruwa..... 6·46n 80·26e
Jaffna..... 9·40n 80·00e
Kalutara..... 6·35n 79·58e
Kandy..... 7·18n 80·38e
Kegalla..... 7·15n 80·21e
Kotte..... 6·54n 79·54e
Kurunegala..... 7·29n 80·22e
Madampe..... 6·41n 80·02e
Mannar..... 8·59n 79·54e
Matale..... 7·28n 80·37e
Matara..... 5·56n 80·33e
Moratuwa..... 6·46n 79·53e
Negombo..... 7·13n 79·50e
Nuwara-Eliya.... 6·58n 80·46e
Panadura..... 6·43n 79·54e
Pulmoddai..... 8·56n 80·59e
Puttalam..... 8·02n 79·49e
Ratnapura..... 6·41n 80·24e
Sigiriya..... 7·57n 80·45e
Talaimannar..... 9·05n 79·44e
Tangalla..... 6·01n 80·48e
Trincomalee..... 8·34n 81·14e
Vavuniya..... 8·45n 80·30e
Watawala..... 6·57n 80·32e
Weligama..... 5·58n 80·25e

Physical features and points of interest

Adams Peak..... 6·48n 80·30e
Adams Bridge... 9·04n 79·37e
Aruvi, *river*..... 8·49n 79·55e
Bengal, Bay of... 9·30n 81·40e
Dedura, *river*.... 7·36n 79·48e
Delft Island..... 9·30n 79·42e
Dondra Head *headland*..... 5·55n 80·35e
Gal, *river*..... 7·18n 81·46e
Indian Ocean.... 5·45n 80·40e
Jaffna Lagoon... 9·35n 80·15e
Kala, *river*..... 8·18n 79·50e
Karaitivu, *island*. 8·23n 79·47e
Kataragama, *inselberg*..... 6·23n 81·20e
Kelani, *river*..... 6·58n 79·52e
Knuckles, *mountain*..... 7·24n 80·48e
Kokkilai Lagoon.. 9·00n 80·56e
Maduru, *river*.... 7·52n 81·31e
Mahaweli, *river*.. 8·27n 81·13e
Mannar, Gulf of.. 8·30n 79·30e
Mannar Island... 9·03n 79·50e
Monaragalla, *inselberg*..... 7·06n 80·38e
Palk Bay..... 9·30n 79·15e
Palk Strait..... 9·30n 79·30e
Pedro, Point..... 9·50n 80·14e
Pidurutalagala, *mountain*..... 7·00n 80·46e
Polonnaruwa, *ruins*..... 7·56n 81·00e
Puttalam Lagoon. 8·07n 79·47e
Rakwana, *mountains*..... 6·30n 80·35e
Sangamankanda Point..... 7·01n 81·52e
Senanayake Samudra, *reservoir*..... 7·11n 81·29e
Walawe, *river*.... 6·06n 81·01e
Westminster Abbey, *inselberg*..... 7·02n 81·32e
Yan, *river*..... 8·55n 81·01e

mediate zone consisting of upland ridges and valleys, which are separated from the massif by well-defined scarps. The intermediate zone, in turn, is surrounded by an outer or lower coastal lowland zone, except in the southwest, where the land abuts directly onto the sea in the form of cliffs. Elsewhere there is a coastal fringe consisting of a series of lagoons, sandbars, peninsulas, dunes, and marshes. Five-sixths of the country lies at an altitude of less than 1,000 feet.

The four physiographic regions

For convenience Sri Lanka may be divided into four regions—the central highlands, the southwest, the east, and the northern lowlands. The central highlands, which were formed by faulting, have some fine scenery. The highest point is Pidurutalagala, a rounded peak 8,281 feet (2,524 metres) high; more renowned, however, is the pyramidical Adams Peak, 7,360 feet (2,243 metres) high, which lies to the southwest of Pidurutalagala and is a centre for pilgrimages by Buddhists, Hindus, Muslims, and even Christians.

The southern edge of the central massif consists of a rim of steep scarp running east to west and curving southward in the centre; at one point, called the World's End, the scarp wall forms an almost sheer drop of about 5,000 feet. To the north of the central highlands lies Knuckles Mountain, which is separated from the highlands by the valley of the eastward-flowing Mahaweli Ganga, Sri Lanka's largest river.

The southwest region is also mountainous and includes the Rakwana, or Bulutota, Range, a continuation of the central massif. Its topography consists of elongated ridges, parallel valleys, and a trellis (crisscross) drainage pattern.

The east consists of an undulating plain, dotted with isolated hills or inselbergs (steep-sided hills that remained after the remainder of the landscape had been scoured by erosion). Among these hills are Westminster Abbey (named for its shape), Monaragalla, and Kataragama.

The northern lowlands region consists of rolling countryside intersected by rivers and marked by occasional ridges and isolated hills. In the extreme north, the Jaffna Peninsula, which is made of limestone, is underlain by a coral reef. Several offshore islands have the same formation.

Drainage. The rivers of Sri Lanka are noted for their number rather than for their size. They radiate from the central highlands like the spokes of a wheel, except for the upper course of the Mahaweli Ganga, which has developed different characteristics from the fault movements that have shaped the terrain. Most of the rivers, however, tumble down in a series of waterfalls and cascades before reaching the plain. The Mahaweli Ganga is only 206 miles long and Aruvi Aru 104 miles long; the other rivers, of which there are more than 100, are all less than 100 miles in length. The mouths of the rivers are choked with sand.

The rivers

Geology. Forming part of a very old landmass, Sri Lanka has experienced a long series of geological events, including folding movements as well as uplifting movements that have resulted in the fracturing and faulting of rocks. These movements, in addition to providing differences in rock composition, account for the scarps that separate the different elevation levels. Climate has also exerted an influence upon the sculpturing of the land surface.

Nearly nine-tenths of the island is floored by rocks of Precambrian age (570,000,000 to 4,600,000,000 years old), consisting of schists (rocks changed by heat and pressure at depth and having a banded and foliated structure) and gneisses (imperfectly foliated rocks changed by heat and pressure, with alternating light and dark bands differing in mineral composition, and more coarsely grained than schists), overlain by sediments changed by heat and pressure. Among the sedimentary rocks, only a few outcrops of shale and sandstone of Jurassic age (from 136,000,000 to 190,000,000 years old) are to be found. These occur in the Tabbowa and Andigama areas; some coal has been found in them. The north and northwest areas are floored by sedimentary limestone of Miocene age (from 7,000,000 to 26,000,000 years old). Gravel deposits adjoining this sedimentary limestone belt may have been laid down in periods when the area lay under a shallow sea during the last 2,500,000 years. Sand dunes and alluvium are commonly deposited in

coastal areas. Some coral formations occur at Mathagal in the north and in the Madampe-Ambalangoda area in the southwest.

Climate. Sri Lanka lies between 6° and 10° north of the Equator, and consequently might be expected to experience an equatorial type of climate. Its position in relation to the Indian subcontinent and to the neighbouring seas, however, has resulted in modification of the climate, with a regime of monsoons (rain-bearing winds) being superimposed upon the equatorial regime. Further modifications are caused by relief.

Rainfall is the most important element in the climate. Although Sri Lanka is a small island, the distribution of rainfall is not uniform, varying from more than 200 inches to less than 40 inches a year in different parts. Further climatic variations in different places are the result of the seasonal incidence of rainfall. Because of elevation and aspect, the central and southwestern parts receive the most rain. Some strips along the extreme northwest and southeast coasts are described as arid because they receive less than 50 inches of rainfall a year; while this is a relatively large amount, much of the moisture is lost through evaporation, in addition to which there are several rainless months. Much of the rain is also lost because it runs off the surface. Cultivation in these areas is restricted to the short rainy season, except when irrigation is resorted to.

The southwest monsoon blows from May to October, and from December to March the northeast trade winds —locally known as the northeast monsoon—prevail. The monsoons are often irregular, beginning or ending earlier or later than expected, thus adding to the hazards of agriculture. Between the monsoons—from approximately October to December and again from March to May—normal equatorial conditions prevail.

The five climatic zones

While the north and east of the island are generally drier, and the south and west wetter, five climatic zones may be distinguished in terms of rainfall: the wet lowlands and the wet highlands in the west; the dry lowlands and the dry highlands in the east; and the arid zones, which, as already mentioned, are in the extreme northwest and southeast. At Colombo in the wet lowlands, annual rainfall amounts to 90 inches, and the average annual temperature is 81° F (27° C); at Watawala in the wet highlands, the annual rainfall amounts to 218 inches, and the average temperature is 65° F (18° C); at Trincomalee in the dry lowlands, annual rainfall averages 64 inches, and the temperature 82° F (28° C); at Mannar, on the island of Mannar in the arid zone, rainfall averages 40 inches a year, and the temperature averages 82° F (28° C). Nuwara Eliya in the wet highlands experiences some night frost but no snow. In Sri Lanka in general, temperatures vary little at different seasons; the annual and diurnal (daily) temperature variation does not exceed about 8° F (4° C). Most places in Sri Lanka have a high relative humidity.

Soils. In the wet zone, climate is the dominant factor in the formation and characterization of soils. Lateritic loams (reddish leached iron-bearing soils, composed of a mixture of clay, silt, sand, and organic matter) are the most common. They are generally well-drained, friable when dry, but poor. In the dry zones, the same types of rocks result in semilateritic and nonlateritic earths. In the sedimentary limestone areas, thin mantles of red loams, rich in lime, are common. Along the lower courses of the rivers, and on their floodplains, rich rice soils occur. Sandy soils are prevalent on the coastal dunes and sandbars.

Vegetation and animal life. As Sri Lanka has been occupied by man for many centuries, the natural vegetation cover has been altered by cultivation. The open, rolling grasslands, known as *patanas*, in the eastern highlands of the Uva region are said to have resulted from the destruction of the forests that previously grew there.

Nearly 70 percent of the island, however, is still covered with forest, natural grassland, and wasteland. The best developed evergreen forests are to be found in the wet zone; the larger part of the dry zone forests are of poor composition.

Animal life, once abundant, is now becoming depleted. The elephant population has dwindled to a few hundred. The common forest animals are deer, leopards, bears, jackals, monkeys, wild boars, cheetahs, and wild buffalo.

Traditional regions. The present-day pattern of population distribution bears little resemblance to that of the period before the advent of the Europeans, which ended with the arrival of the Portuguese in 1505. Earlier Sinhalese and Tamil settlements were confined to the dry lowlands of the north and were focussed on Anuradhapura. This environment resembled that of the settlers' original homelands in South India, where rice cultivation was based on irrigation from water tanks. The forested wet zone of southern Sri Lanka remained almost uninhabited until plantation agriculture was introduced by the British in the early 19th century.

The landscape under human settlement. *Rural settlement.* Individual villages are surrounded with cultivated plots and orchards, from which vegetables, fruits, and spices are obtained. Elsewhere, temporary (shifting) cultivation is practiced; after being used for from three to five years, the fields are allowed to revert to forest again.

Urban settlement. In 1971, nearly one-fourth of the total population was urban, and nearly half this number lived in towns in the Colombo region alone; the population of the city of Colombo is over 562,000. The urban population in other districts in relatively small. In the Kalutara district, to the south of Colombo, it is 161,000; in Galle, in the extreme southwest, it is 158,000; in Jaffna, in the extreme north, it is 236,000; at Matara, in the extreme south, it is 66,000; and at Batticaloa, in the east, south of Trincomalee, it is 71,000. It is less than 55,000 in each of the other 16 districts.

The urban population

PEOPLE AND POPULATION

Groups historically associated with the contemporary country. *Ethnic and linguistic groups.* As already mentioned, the two principal linguistic groups in Sri Lanka are the Sinhalese and the Tamils. The Sinhalese number over 9,000,000, of which over 5,000,000 are low-country Sinhalese and about 4,000,000 up-country, or Kandyan, Sinhalese. The Tamils number more than 2,500,000; they are divided about equally into Sri Lanka Tamils and Tamils of Indian origin. The Muslims, who as mentioned are sometimes miscalled Moors, are partly descended from Arab seafarers; they number over 850,000, and their mother tongue is Tamil. Other groups are the Burghers, who are of mixed Dutch and Eurasian origin, and Eurasians, who together number about 44,000; Malays, who number about 42,000; and a few small groups, numbering about 14,000 altogether, of Europeans who have adopted Sri Lanka nationality, Indians, and others. The Veddas, who are believed to be the aboriginal inhabitants of the island, are very few in number and are now practically indistinguishable from other groups, with whom they have intermarried. The Tamils of Indian origin are treated as stateless and have no civil rights. By the terms of an agreement with the Indian government in 1964, more than 500,000 of them are to be repatriated to India within 15 years, while 300,000 of them will be allowed to apply for citizenship in Sri Lanka.

Religious groups. The Sinhalese are generally Buddhists and the Tamils usually Hindus. Christians of all denominations number almost 987,000, while Muslims number approximately 910,000.

Archaeology. Geographically and culturally Sri Lanka is closely related to India, with which its sequences of prehistoric culture have much in common. Artifacts, probably relating to Middle Paleolithic times, have been found in different parts of the country, although their authentication must await discovery in well-stratified contexts. Data on later cultures are most numerous.

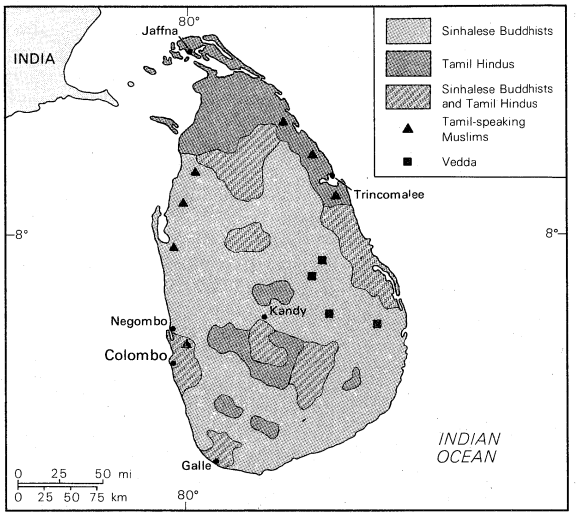
The outstanding feature of this culture is its highly developed stone industry, which shows traces of a Middle Stone Age tradition, such as is found on the mainland of India, combined with Late Stone Age techniques.

In the topmost layer of many sites, iron tools and pottery are found. Their exact age has yet to be determined,

but it is unlikely that they date to a time much earlier than the early Buddhist period.

The connection between these cultures and that of the modern Veddas has yet to be determined. The rock paintings in the caves, with depictions of elephants and riders, would be part of these stone cultures. It is equally possible that they are the work of the Veddas in more recent times.

No remains of the Neolithic or Chalcolithic cultures



General ethnic composition of Sri Lanka.

that spread through peninsular India in the 2nd millennium BC, with their distinctive painted pottery and ground stone axes, had been found in Sri Lanka by the 1970s.

That there was an extension of the South Indian Iron Age into the island is attested to by the finds of megalithic remains, stone cist graves, and circles, mainly in the north, and of terra-cotta elephant-shaped sarcophagi and of urn burials in the North Western Province. There, typical black and red pottery, sometimes with scratched owners' marks, has been found. These marks included some that resemble letters of the early Brāhmī inscriptions of India and Sri Lanka (from the 3rd century BC), and the discovery of similar black and red pottery in the lowest levels of Anuradhapura suggests that this invasion, probably of Tamil-speakers, may not have been earlier than the popularly believed colonization of Sri Lanka by North Indian settlers.

The problem of population growth

Demography. The high potential for growth of the population is indicated by the fact that over 40 percent of it is under 14 years old. In 1971 there were about 6,500,000 males and 6,200,000 females. The population density per square mile, which was about 140 in 1900, had risen to about 400 in the early 1960s and was nearly 502 in 1971. Life expectancy, which was about 42 years in 1946, had risen to about 58 by 1953 and to 66 in 1967. It is estimated that the 1971 population of about 13,000,000 will have increased to about 18,000,000 by 1983, and that it may increase to approximately 26,000,000 by the year 2000.

The birth rate remained approximately constant at 32 per thousand per annum during the 1960s, but decreased to 29 per thousand in the early 1970s. The death rate, as a result of better health conditions, has decreased from about 22 per thousand in 1945 to about 8 per thousand in 1970. The rate of increase of the population is thus about 2.1 percent a year.

THE NATIONAL ECONOMY

A comparative view. Sri Lanka's economy is still precariously dependent upon the exports of its plantation products, which are subject to severe fluctuations in the world market prices as well as to the influence of other factors beyond national control, such as the development of substitutes, changes in people's tastes abroad, changes in manufacturing practices, or the conclusion of multilateral trade pacts. The necessity of diversifying the economy in the agricultural sector by reducing dependence upon imports, especially of food crops, and encouraging the domestic production of substitutes, and achieving a greater degree of industrialization has been partially realized in recent years. The country remains dependent, however, upon imports of rice, sugar, and other foods, as well as of clothing. The diversification of agriculture has been rendered inevitable by falling prices in the world market for tea and rubber—two of the mainstays of the export sector. Tea, rubber, and coconuts together earn about 90 percent of Sri Lanka's foreign exchange, while tea alone, as mentioned, accounts for more than 55 percent. The output of rice has been increased through achieving higher yields, and it is hoped that enough may soon be grown to make Sri Lanka self-sufficient in this respect. The Mahaweli Ganga scheme alone is adding nearly 1,000,000 acres to the area devoted to rice production. Yields are also being increased by irrigation, flood control, the use of artificial fertilizers, intensive cultivation, and the use of better seeds.

Mineral deposits

Natural resources. Mineral deposits include iron ore, ilmenite (iron titanate, a common black mineral), graphite, and gemstones. Iron ore deposits of high quality, suitable for electrosmelting and found at a shallow depth, have been located in the Ratnapura, Balangoda, Kalutara-Batigama, Matara-Akuressa, Matale, and Chilaw areas, all in central or southern Sri Lanka; they total about 7,000,000 tons. Ilmenite deposits in the form of beach deposits are located north of Trincomalee at Pulmoddai; they amount to about 4,000,000 tons. Additional deposits exist at Turrukovil, Pamunugama, and elsewhere. Deposits of monazite (the principal ore of thorium, used in alloys) are found with ilmenite in beach sands along the west coast. There is an extensive deposit of ilmenite at Pulmoddai and important concentrations of monazite at Beruwela and Induruwa. Rutile (titanium dioxide, used to coat welding rods), zircon (zirconium silicate, used to line furnaces), and garnet sand (used as an abrasive) also occur as beach deposits. Sri Lanka is famous for its almost inexhaustible deposits of high-grade graphite. It also possesses both a quantity and a variety of gemstones, including rubies, sapphires, star-stones, and cat's-eyes (gemstones that change in lustre or colour, like the eyes of a cat). Other mineral deposits include large quantities of kaolin (china clay); other clays used for making tiles, bricks, and pottery; an inferior quality of mica; quartz, which is used for glass-making; and limestone suitable for cement manufacture.

Sri Lanka, Area and Population

	area*		population	
	sq mi	sq km	1963 census†	1971 census‡
Districts				
Amparai	1,775	4,598	212,000	273,000
Anuradhapura	2,809	7,274	280,000	389,000
Badulla	1,090	2,822	522,000	616,000
Batticaloa	1,017	2,633	196,000	258,000
Colombo	808	2,093	2,207,000	2,673,000
Galle	625	1,689	641,000	737,000
Hambantota	1,013	2,623	274,000	341,000
Jaffna	999	2,586	613,000	704,000
Kalutara	624	1,616	631,000	732,000
Kandy	914	2,367	1,044,000	1,187,000
Kegalla	642	1,663	579,000	652,000
Kurunegala	1,844	4,776	853,000	1,028,000
Mannar	964	2,497	60,000	78,000
Matale	770	1,995	256,000	316,000
Matara	481	1,246	515,000	588,000
Monaragalla	2,188	5,666	132,000	192,000
Nuwara Eliya	474	1,228	398,000	453,000
Polonnaruwa	1,332	3,449	114,000	164,000
Puttalam	1,172	3,036	303,000	380,000
Ratnapura	1,250	3,239	546,000	662,000
Trincomalee	1,048	2,714	139,000	192,000
Vavuniya	1,467	3,799	69,000	96,000
Total Sri Lanka	25,332§	65,610	10,582,000§	12,711,000

*Including 335 sq mi (868 sq km) of inland water. †Excluding adjustment for underenumeration. ‡Preliminary. §Figures do not add to total given because of rounding.
Source: Official government figures; UN.

Sri Lanka's lagoons and other coastal features are favourable for manufacturing salt from seawater by solar evaporation during the rainless period. Prospecting for oil is in progress, but so far without result.

In addition to the climate and soil, which favour such crops as tea and rubber, the forests constitute a substantial resource, although pure stands of timber are rare. Fish and other edible forms of marine life abound in the seas around the island.

More than 50 waterfalls, some perennial and of considerable size, are found in the hill country; in addition, major irrigation projects, such as those on the Gal Oya, Mahaweli, and Walawe rivers, are also capable of yielding much hydroelectric power.

Sources of national income. *Agriculture, forestry, and fisheries.* As mentioned, tea, rubber, and coconut products are the principal sources of foreign exchange earnings. Most of the people are subsistence farmers; rice is their staple food, although only two-thirds of the country's needs are met by domestic production. Agricultural production is hampered by the fact that farms are often fragmented and scattered. The government, however, is taking measures to consolidate holdings and is seeking to encourage production by means of guaranteed prices. About 40 percent of the tea-producing lands, 70 percent of the rubber-growing areas, and most of the coconut lands are in national hands; the remainder is owned by British companies. Some livestock is raised. There are about 1,600,000 cattle, about 720,000 buffalo, 585,000 sheep and goats, 110,000 pigs, and 6,900,000 chickens on the island. The government is sponsoring a reafforestation program. Local timber production consists mostly of logs and firewood for domestic consumption. Fisheries employ about 40,000 people. Sea fishing is the principal source of the catch, although some fish are caught in rivers. Pearl fishing, once of importance, has been unsuccessful in recent years.

Livestock

Mining. Graphite is the principal mineral export. Most of the ilmenite produced is exported to Japan. Stone quarrying employs about 9,000 people. The cement industry is based on the deposits of limestone in the northern part of the country. The production of salt is a government monopoly.

Manufacturing. Industries are still of minor importance, accounting for only about 10 percent of the gross national income. Several large-scale industries are engaged in the processing of agricultural produce. Cement, textile, leather, and ceramics industries are being developed. The paper industry uses domestic paddy straw, wastepaper, and imported pulp. Steel-rolling, plywood, and tire manufacturing industries are thriving. Industrial estates are being established to promote light industries. Foreign aid is being used to establish sugar-refining, petroleum-refining, chemical, and textile industries.

Energy. About 785,800,000 kilowatt-hours of electricity are generated, both by hydroelectric plants and by thermal plants, which are fuelled by imported oil. Distribution is confined to the cities—Colombo, Nuwara Eliya, Diyatalawa, Norton Bridge, Bandarawela—and some few villages.

Financial services. The Central Bank administers and regulates the country's banking and monetary system and issues the national currency. Commercial banks include three national banks and nine branches of foreign banks. Government-sponsored credit institutions include the State Mortgage Bank, the Agricultural and Industrial Credit Corporation, and the Development Finance Corporation. Savings institutions are the Post Office Savings Bank, the Savings Bank, the Insurance Corporation, and the Employees' Provident Fund.

Foreign trade. Sri Lanka's unfavourable balance of trade is exacerbated by its dependence upon agricultural products, which are subject to fluctuations in market value. Imports are largely rice and other foodstuffs, manufactured goods, raw materials, and machinery. Exports are usually destined for the United Kingdom, China, the United States, Australia, and South Africa. Imports are mainly from the United Kingdom, China, the United States, India, Australia, Japan, and West Germany.

Management of the economy. *The public sector.* The state sponsors a wide variety of industrial corporations, among them a cement works, a chemical factory, an industrial estates corporation that provides ready-made factory units to small industry, a pulp and paper mill, a mineral sands processing plant, an oil refinery, and a steel-rolling mill. The government also sponsors numerous small manufacturing plants as a means of increasing employment.

The private sector. Private industrial development is encouraged by such measures as a five-year tax exemption, duty-free raw materials, and tax relief on investment. Foreign investment is also encouraged; it occurs mainly in the manufacture of textiles, electric appliances, transport equipment, metal, and glass. More than one-third of private industry is engaged in the production of food, beverages, and tobacco products.

Taxation. The rate of taxation is high and continues to rise. Taxes levied include an income tax on individuals and companies, a tax on wealth, a capital gains tax, and various customs and excise duties. Deficit budgeting has become a permanent feature, due to increased governmental expenditure on welfare services. The deficit is partially made up by market and administrative borrowing and foreign aid in the form of grants and loans. The principal sources of aid are the United States and Canada; the largest loans have been made by the Soviet Union, the United States, and West Germany.

Foreign aid and loans

Trade unions and employer associations. There are over 1,400 trade unions, more than half of which are unions for public servants. Sri Lanka is represented in the International Labour Organization. Industrial disputes are settled by 12 labour tribunals. Employer associations include those for planters and manufacturers.

Tourism. A Ministry of Shipping and Tourism has been established and is helping to promote tourism from abroad, in cooperation with the private sector and with the participation of international airlines and foreign governments. A number of hotels have been established in Colombo, seaside resorts developed, and additional transport facilities supplied. The travel trade is already begining to show profits.

Transportation. *Road networks.* Sri Lanka has a good road system, although the roads are often uneven in quality. There are about 23,800 miles of surfaced roads, most of which are in the western lowland, the southern central highlands, and the Jaffna Peninsula in the north. There is also a ferry service from Mannar Island to Pāmban Island, India; it runs parallel to Adams Bridge.

Railways. There are 954 miles of railways, of which about 850 miles have a gauge of five feet six inches; the remainder are narrow gauge. The railways are owned and controlled by the government. They serve the flat, low country, the central hill areas, and the Kelani Valley in the southwest.

Port facilities. Colombo, the capital, possesses the principal port. Other ports are Trincomalee on the east coast and Galle on the south coast. Galle handles the overflow of commercial shipping from Colombo. Trincomalee is one of the finest and largest natural harbours in the world. It was formerly a British naval base and now handles some of the island's export trade.

Air transport. There is an international airport at Colombo. Air Ceylon maintains both international and domestic services, as well as a regular air service between Tiruchchirāpalli, India and the Jaffna Peninsula. The Indian Airlines Corporation operates regular flights to Colombo from Madras, Tiruchchirāpalli, and Trivandrum.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* Sri Lanka is an independent republic within the Commonwealth of Nations. The former 30-member Senate was abolished in September 1971, and under the terms of the 1972 constitution the House of Representatives was renamed the National State Assembly. There is a Cabinet, headed by the prime minister. There is a well-established public service that administers through various government departments and corporations. The head

The constitution of 1972

of the republic is represented by the president (formerly governor general). The basic constitutional document was, until it was superseded by the 1972 constitution, the Ceylon Independence Order in Council, which was issued in 1947.

Regional and local government. Sri Lanka is divided into nine administrative provinces; these in turn are subdivided into 22 districts, which constitute the principal rural administrative units. Local authorities consist of 10 municipal village councils, 76 town councils, 36 urban councils, and 510 village councils.

The political process. *Elections.* The present National State Assembly was elected in 1970 and consists of 151 popularly elected members and six members who were appointed by the governor general (now the president). Under the terms of the 1972 constitution, the elections of 1977 will begin a five-year election cycle and the appointed membership will be eliminated.

Parties. The principal parties are the Sri Lanka Freedom Party (SLFP), the United National Party (UNP), the Trotskyist Lanka Sama Samaja Party (LSSP), and the Ceylon Communist Party (CCP).

De facto political developments. In 1970 a left-wing coalition government, supported by the SLFP, the LSSP, and the CCP and headed by Mme Sirimavo Bandaranaike of the SLFP as prime minister, came to power. A Constituent Assembly to frame a republican form of constitution was subsequently convened.

The Supreme Court

Justice. The judicial system is elaborate, comprising the Supreme Court, district courts, magistrate's courts, courts of requests, and rural courts. The Supreme Court consists of the chief justice and ten puisne (junior) justices. The justices of the Supreme Court are appointed by the governor general, other judicial officers by the independent Judicial Service Commission. The common law of Sri Lanka consists of Roman–Dutch law, but aspects of English law have also been adopted, particularly in commercial matters. Customary law frequently operates in matters concerning the personal status of Kandyans (those living in the former kingdom of Kandy), Tamils, and Muslims. In civil and criminal matters, English procedure generally is followed; the penal code is based on that of India.

The armed forces. The armed services comprise the army, navy, and air force. The army and navy each have a regular force, a regular reserve, a volunteer force, and a volunteer reserve. Service is entirely voluntary; there is no conscription. The navy possesses a frigate, a number of small patrol boats, and smaller craft. The air force has a strength of about 1,700 men and possesses a number of aircraft, including jets, light transports, helicopters, and trainers.

Administration. *Education.* For purposes of education the island is divided into 14 regions, for each of which a director of education is responsible. Education is free, from kindergarten to university level. Almost all children between the ages of 6 and 11 attend school, and a high proportion continue until the age of 16 or 18. There are more than 9,600 schools, with a total enrollment of about 3,000,000 students. There are several technical and teacher training schools and colleges. The island's four universities are the University of Ceylon and the University of Colombo, both in Colombo; Vidyodaya University in Gangodawila; and Vidyalankara University in Kelaniya. In 1963 almost 80 percent of the males and 64 percent of the females were literate.

Health and welfare. Public health and medical care services are mostly planned and financed by the Ministry of Health, under which is the Department of Health, which is responsible for providing these services. There are more than 800 medical institutions, including about 330 hospitals and 332 central dispensaries. Campaigns are conducted periodically against tuberculosis, cancer, venereal diseases, malaria, and leprosy, but they have not been notably successful. In addition to Western medical facilities, Sri Lanka in recent years has revived the practice of traditional medicine, which is conducted under government auspices.

The Ministry of Social Services

The Ministry of Social Services is responsible for public assistance, the relief of distress, the care and welfare of the aged and infirm, the administration of workman's compensation, and the rehabilitation of antisocial persons. Rural development activities include training in the performance of community services and self-help programs. There is a vigorous cooperative movement that is especially active in agriculture, industry, and fisheries.

Housing. The Ministry of Housing and Construction is responsible for the national housing program and for town and country planning. The housing shortage is most acute in and around Colombo. The ministry has constructed over 1,000 units in the city and promotes private construction by the granting of loans and the acquisition of land for housing development. Housing plans are directed toward the provision of low-income units, the extension of sanitation facilities, and the clearance of city slums.

Police services. The police and prisons services are efficient and adequate. There are about 260 police stations, and the police service consists of almost 10,000 persons. Village headmen assist the police. There are 14 prisons.

Social conditions. *Wages and cost of living.* The Employees' Provident Fund and wages boards have been established to regulate labour conditions, including conditions of employment, wage rates, industrial safety, and related matters. Wage scales are reviewed periodically in relation to the cost of living index. Unemployment and rising living costs have, nevertheless, caused much labour and student unrest; the increase in the population has been the principal cause of these problems.

Health conditions. Tuberculosis, malaria, and filariasis (parasitic infestations of the blood or tissues) are serious health problems. Leprosy, venereal diseases, and intestinal disorders also occur with frequency. In an attempt to lower the high birth rate, family planning has become an integral part of maternal and child care programs.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. Sri Lanka has a rich cultural heritage reaching back to the 3rd century BC. Its artistic tradition centres largely on Buddhist and Hindu religious art and the dance-drama. During the British administration the arts suffered an eclipse, and since independence there has been a tendency to imitate Indian and Western art forms. The government, however, is aware of the importance of preserving traditional forms.

The state of the arts. There are three main forms of the national dance-drama. The *kōlam* (a Tamil word for costume) play is a masked dance-drama that is performed annually. The devil dance, an expression of pre-Buddhist beliefs in spirits, is a masked dance performed to alleviate misfortune or illness. Most important is Kandyan dance, the sophisticated national dance. It is performed by elaborately costumed dancers and deals with mythological, historical, and naturalistic themes. It is taught at the National Academy of Dance in Kandy.

The national dance-drama

Since independence there has been a movement to adapt traditional dramatic forms to the modern stage. Translations of foreign plays are also performed.

Examples of national sculpture, painting, and architecture are seen in the remains of ancient works. These include colossal Buddha images, temples (especially around Kandy), and temple paintings. Other surviving art traditions include ivory carving, metalwork, lacquered woodwork, and basketry.

Cultural institutions. The Arts Council and the Government College of Fine Arts play an important role in the revival of folk arts and crafts. The three national museums in Colombo, Kandy, and Ratnapura contain cultural and artistic collections. The Government Archives are in Colombo. Archaeological reserves, distributed throughout the island, include temples and cities.

Press and broadcasting. The press is free and independent. There are four main newspaper groups, which publish in Sinhala, Tamil, and English. The government publishes a gazette as well as its own weekly newspapers.

The Ceylon Broadcasting Corporation provides national, commercial, and educational broadcasting services; it is responsible to the Ministry of Information and

Broadcasting. Domestic programs are broadcast in Sinhala, Tamil, and English. There are also broadcasts to Southeast Asia and Europe. There is as yet no television.

PROSPECTS FOR THE FUTURE

The increase in population and the decline in living standards are major problems in contemporary Sri Lanka. In addition, the government confronts the task of fostering and strengthening the sense of national unity, a problem it is attempting to solve by instilling in the various communities a spirit of good-will and of mutual accommodation.

Attempts at population redistribution

Among the economic and social measures taken by the government have been attempts to relieve pressure on the land by a redistribution of population, to increase the area of land under cultivation, to increase the yield of fisheries, and to sponsor more education. At present, for example, the southwest part of the island and the northern peninsula, which comprise together only one-third of Sri Lanka's area, contain 80 percent of its population. Attempts are being made to move families out of this area and into the sparsely populated dry zone by providing irrigation, as in the Mahaweli Ganga scheme. Agricultural and land development schemes together account for nearly 30 percent of all current capital expenditure. Increased investment in fishing is likewise expected to step up production to over 200,000 tons of fish per year, thus reducing the present rate of expenditure upon imports of fish and fish products. The use of deep-sea trawling methods and of mechanized fishing craft has been increased, and fishery harbours and fish freezing plants have been established.

Although 14 percent of the government's budget is spent upon education, much of the education has been of the academic type, with emphasis on the humanities; as a result, a large body of educated unemployed have come into existence. Educated and unemployed youths have helped to contribute to disruptions of normal life from time to time, of which one of the most serious has been the revolutionary uprising that took place in 1971. Much will depend upon the emergence of a type of nationalism that, in facing the challenges of contemporary national problems, will draw upon the past rather than seek to return to it.

BIBLIOGRAPHY. The administration reports of the various government departments and corporations in Sri Lanka contain much information, as do the publications of the national Association for the Advancement of Science, the Medical Association, and such specialized publications as *Tropical Agriculturist, Industrial Ceylon,* and the *Ceylon Journal of Science.* E.K. COOK, *Ceylon,* rev. by K. KULARATNAM (1951), is a good general account, but information on the economy is dated. *A Concise Atlas Geography of Ceylon* (1971), is a helpful adjunct. Relevant chapters by B.H. FARMER in O.H.K. SPATE and A.T.A. LEARMONTH, *India and Pakistan,* 3rd ed. rev. (1967); and by K. KULARATNAM in S.P. CHATTERJEE (ed.), *Developing Countries of the World* (1968), also provide complementary material. H.A.J. HULUGALLE, *Ceylon* (1942), is a delightful general account in concise form. P.G. COORAY, *The Geology of Ceylon* (1969), is a comprehensive work. M.D. RAGHAVAN, *The Karāva of Ceylon* (1961); N.D. WIJESEKERA, *The People of Ceylon,* 2nd ed. (1965); and K. KULARATNAM, *Tamil Place Names in Ceylon Outside the Northern and Eastern Provinces* (1966), throw an objective light on myths and legends about the people and their origins. Recent and reliable information on the economy may be found in: the INDUSTRIAL DEVELOPMENT BOARD, COLOMBO, *Statistical Guide* (1960) and *Research and Industry* (1970); *Statistical Abstracts of Ceylon 1967–68* (1970); and the *Statistical Pocket Book of Ceylon* (1969).

(K.Ku.)

Ssu-ma Ch'ien

Ssu-ma Ch'ien (in Pin-yin romanization Si-ma Qien), astronomer, calendar expert, and the first great Chinese historian, was the author of the *Shih-chi* ("Historical Records"), the most important history of China down to the end of the 2nd century BC. He was the son of Ssu-ma T'an, the "grand historian" (sometimes translated as "astronomer royal") at the Han court during the period 140–110 BC. The office of grand historian combined responsibility for astronomical observations and for the regulation of the calendar with the duties of keeping a daily record of state events and court ceremonies.

Early career

Ssu-ma Ch'ien was born at Lung-men. His date of birth is unknown. In view of his later achievements, the year 145 BC seems most likely. After travelling extensively in his youth, he entered court service. In 111 BC he accompanied a military expedition into the southwest of China, and in 110 BC he was a member of the Emperor's entourage when the latter visited Mt. T'ai to conduct the sacrifices symbolizing the dynasty's authority. In the same year, his father died, and after the mandatory period of mourning he was appointed in 108 BC to succeed him in the post of grand historian.

In 105 he was among those responsible for a complete reform of the Chinese calendar, a reform prompted by the Han emperor Wu Ti's inauguration of what was to be a "new beginning" to the Han dynasty. At about the same time, Ssu-ma Ch'ien began to undertake the unfulfilled ambition of his father to write a definitive history of the Chinese past, an ambition strengthened by his belief that under Wu Ti the Han had reached a peak of achievement that deserved to be recorded for posterity. Before his history was completed, however, Ssu-ma Ch'ien deeply offended the Emperor by coming to the defense of a disgraced general. Ch'ien was arraigned for "defaming the Emperor," a capital crime. Either because the Emperor felt him too valuable a man to lose or because Ch'ien himself requested a reprieve so that he could complete his history, he was castrated instead of executed.

By courtesy of the National Palace Museum, Taipei



Ssu-ma Ch'ien, ink and colour on silk by an unknown artist.

Wu Ti later relented, and Ssu-ma Ch'ien again rose in the Imperial favour, becoming "palace secretary" (*chung-shu-ling*). But he remained bitterly conscious of the shame he had suffered and lived a retiring life, devoting himself to the completion of his great masterpiece. He died *c.* 85 BC.

Structure and content of *Shih-chi*

The *Shih-chi* is his great claim to fame. There had, of course, been many histories before Ssu-ma Ch'ien's time. The keeping of court chronicles was already an established practice under the earlier dynasties. One such work, the *Ch'un Ch'iu* ("Spring and Autumn [Annals]") of the petty state of Lu, was said to have been the work of Confucius. It had achieved status as a canonical book largely because of its ethical judgments on the events which it recounts.

Ssu-ma Ch'ien denied that his work was in any way comparable to this great Classic. He termed himself not a "maker," like Confucius, but merely a "transmitter" of

past events. His great successor as historian of the Han, Pan Ku (AD 31–92), took him to task for his haphazard use of ideas from the various schools of philosophy and for his devotion to Taoism. But the ethical standards of Han Confucianism that Pan Ku and his contemporaries could take for granted had not in Ssu-ma Ch'ien's lifetime achieved the authority they had by the 1st century AD. Like most of his contemporaries, Ch'ien was an eclectic, employed at a court at which magic and the supernatural were still potent forces and where the state religious cult and accepted moral and political standards were still in a fluid state. Ssu-ma Ch'ien's moral judgments are, thus, not in accordance with any consistent theories.

His main achievement was that he reduced to an orderly narrative the complex events of the past, recorded in often contradictory sources deriving from the many independent states, each of which employed its own chronology. He organized these facts not, as in previous histories, simply as a chronologically ordered record but according to a new five-part plan. The "Basic Annals" gave a dated chronological outline centred around events at the court considered to have been the paramount power at the time. The following section consisted of chronological tables in which he attempted to clarify the confusion of the history of the various independent feudal kingdoms and to enable his reader to see at a glance what was happening in each of the states at any given time. The detailed accounts of each state were given in chapters entitled "The Hereditary Houses." A number of monographs dealt with various crucial aspects of government. These sections show Ssu-ma Ch'ien to have favoured the practical reformist statesmen who, in his own time, were formulating new policies for the increasingly centralized state, rather than the proponents of Confucian moral theories. The work ends with a collection of "Biographies" that deal with a variety of famous individuals, who are selected as exemplars of various types of conduct, and also with the affairs of the various foreign peoples, relations with whom were beginning to become increasingly important during the reign of Wu Ti.

Ssu-ma Ch'ien's aims and methods

Although the *Shih-chi* provided a model for the later dynastic histories, it differs from them in many ways. Its time span is far longer: such attempts to encompass the whole of human history were rare among later Chinese historians. Its source material, too, was far more varied. It incorporated not only the court annals of the Ch'in and the Han but also various earlier histories, parts of court chronicles of various feudal states, material from the canonical books and from the philosophical writings of all the schools, even historical romances. Neither is his subject matter exclusively court centred and "political," as were the later histories; it includes a far wider range of society, including businessmen and merchants, condottieri and bandits, actors and court favourites, good officials and bad.

He did not attempt to compose "objective" history but rather belonged unmistakably to the didactic Chinese tradition of history. He makes moral judgments on his characters. He also attempts to characterize them in types, recording an individual's exemplary deeds in one chapter and his misdeeds elsewhere. But the lessons he derives from history are varied and often mutually incompatible. He is much more notable for the critical attention he devotes to his sources. His acute critical comments are appended at the end of each chapter.

Finally, Ssu-ma Ch'ien is important not only as a historian but also as a master of racy, flexible Chinese prose. He has exerted a potent influence on later writers, particularly upon the early writers of narrative prose and fiction. Since Ssu-ma Ch'ien's time, his history has been acknowledged as the great historical masterpiece in Chinese, a standard against which all later histories would be measured and a model for large-scale historical composition, not only in China but in all Far Eastern countries influenced by the Chinese literary tradition.

BIBLIOGRAPHY. The standard study of Ssu-ma Ch'ien is the extremely perceptive work of BURTON WATSON, *Ssu-ma Ch'ien, Grand Historian of China* (1958), which gives a full critical analysis of Ssu-ma Ch'ien's works and also translates all the relevant biographical material from early sources. The historiographical methods and ideas of Ssu-ma Ch'ien are discussed in FRANK A. KIERMAN, *Ssu-ma Ch'ien's Historiographical Attitude As Reflected in Four Late Warring States Biographies* (1962). Further critical discussions are given in DERK BODDE, *Statesman, Patriot, and General in Ancient China: Three Shih chi Biographies of the Ch'in Dynasty, 255–206 B.C.*, rev. ed. (1967); and in the magisterial classical study of EDOUARD CHAVANNES, *Les Mémoires historiques de Sema-Ts'ien*, vol. 1 (1895). The latter work, in five volumes (1895–1905) translated the first 47 chapters of the *Shih-chi* with copious scholarly annotation. A sixth volume (1969) includes further translations by Chavannes and a bibliography of Western-language translations. Much of the remaining portion is translated in BURTON WATSON, *The Records of the Grand Historian of China*, 2 vol. (1961), which deals entirely with sections on the Han period.

(D.C.T.)

Stadium

The stadium is a unique form of structure with an exceptionally significant role in 20th-century construction technology. Basically, it is an enclosure that combines broad space for athletic games and other exhibitions with large seating capacity for spectators. The name derives from the Greek unit of measurement, the stade, the distance covered in the original Greek footraces (about 600 feet).

Classical stadiums. The first Greek stadiums were long and narrow and in the shape of a horseshoe, sometimes cut into the side of a hill, as at Thebes; Epidaurus; and Olympia, site of the Olympic Games, begun there in the 8th century BC. The Greeks also built hippodrome stadiums similar in layout but broad enough to accommodate four-horse chariot races, a feature of the Olympic Games as early as the 7th century BC.

Hippodrome stadiums

The design of the Greek stadium was taken over and improved on by the Romans, who built two types of stadiums: the circus and the amphitheatre. The circus was the Roman version of the hippodrome, a hairpin-shaped structure designed for chariot races. Down the middle of the arena ran a dividing wall, the spina, decorated with statues and obelisks; the racecourse ran around the spina. The largest, and doubtless the finest ever built, was the Circus Maximus in Rome, which in its principal form dated from the 1st century BC. Situated between the Palatine and Aventine hills, it measured 2,000 by 600 feet (610 by 190 metres) and is reputed to have held from 150,000 to 250,000 spectators. Rebuilt several times, it was ultimately demolished.

In contrast to the circus, the amphitheatre, one of the most characteristic of all Roman buildings, was completely enclosed on all sides. Intended for gladiatorial combats and other contests in which the precise dimensions of the field were of slight importance, the amphitheatre was designed to afford maximum seating capacity and optimum visual facility for spectators. The focus of the amphitheatre was on the relatively small arena, which drew its name from the Latin word for sand. The Romans built amphitheatres in all parts of their empire, but the classic example remains the Flavian Amphitheatre, or Colosseum, built in Rome in the latter half of the 1st century AD. Their possession of concrete enabled the Roman engineers to build this remarkable building as a self-supporting structure. To reinforce their joints they employed iron clamps, whose pillage in the Middle Ages at one time threatened the structure's stability.

Other notable Roman amphitheatres include those at Pompeii, Italy, and at Arles, France, the latter remaining the best preserved example of the form.

After the fall of Rome, amphitheatres survived for some time until the decline of the cities themselves terminated the spectacles they had held. Nearly two millennia passed before the form was revived, much smaller, usually temporary stands sufficing for such intervening spectatorial events as the knightly tournaments of the Middle Ages.

The renewal of large-scale spectator sports came about gradually in the 19th century with the growth of cities and the revival of interest in athletic contests; symbolically, the great impetus for the design of modern stadiums came from the revival of Olympic competition in 1896.



Maracaña Stadium, the municipal stadium in Rio de Janeiro, one of the world's largest, with a total capacity of 200,000 persons. The Maracañazinho, the smaller domed stadium behind it, can hold up to 40,000.
Manchete Press Agency—Pictorial Parade

Stadiums for modern Olympics. The stadium for the first modern Olympiad in Athens was a reconstruction of the ancient marble stadium built by Herodes Atticus on the site of an even earlier stadium. The Olympics since that time have provided a major focal point for the development of the modern stadium concept. In each four-year period since the first modern Olympiad, the host country has called on its most talented designers to erect a permanent stadium to mark the event. The first stadium of the modern genre was for the IV Olympiad in 1908 at Shepherd's Bush in London. The stands were partly roofed and the stadium seated over 50,000 people. Other Olympic stadiums of architectural note built before World War II include those at Stockholm (1912); Colombes, outside Paris (1924); Amsterdam (1927); and Berlin (1936). The Helsinki stadium, which was built for the XII Olympiad (1940, which was cancelled by the outbreak of World War II), was the scene for the XV Olympiad in 1952.

Other modern stadiums. Stadiums rivaling the size of those of ancient Rome were built in several cities in the first half of the 20th century, notably the vast Strahov Stadium, in Prague, completed in 1934 for the Sokol gymnastics exhibition and seating over 240,000 people. The municipal stadium in Rio de Janeiro has accommodated 200,000 people for football (soccer) games. Other stadiums seating over 100,000 include Hamden Park, in Glasgow (149,500); Wembley, in North London (126,000); Melbourne Cricket Ground in Melbourne, Australia (116,000); Tsentraly (Lenin) Stadium, in Moscow (103,000); Odsal Stadium in Bradford, England (102,500); Aztec Stadium, in Mexico City (100,000); and Soldier Field, Chicago (110,000).

Football stadiums

American football inspired a new type of stadium design, the elliptical bowl, first employed in the Yale Bowl at New Haven, Connecticut, in 1914, and repeated in several other stadiums, including the Rose Bowl, Pasadena, California, and the University of Michigan Stadium in Ann Arbor, Michigan, both seating 100,000. Because the bowl is entirely unsuited to the other principal U.S. sport, baseball, another type of American stadium has evolved, in which the aim is to supply maximum roofed seating capacity to protect spectators against the sun. A notable pioneer in this trend was triple-tiered Yankee Stadium, New York, built in 1923.

Design innovations. A basic difficulty of the roofed stadium was the interference with visibility by the columns supporting the roof. Dodger Stadium in Los Angeles was the first tiered stadium to provide column-free views from all seats (1959), followed by Shea Stadium in New York, which added field seat sections that rotate around the stadium to permit conversion from a baseball to a football arrangement (1964). Several other outstanding stadiums have been built for baseball, for example, Robert F. Kennedy Stadium, Washington, D.C.; Busch Memorial Stadium, St. Louis, Mo.; and Atlanta Stadium, Atlanta, Georgia.

The most significant development, however, was the Astrodome, built in Houston, Texas, the first major fully roofed stadium. Indoor stadiums of considerable size had been built previously, notably Madison Square Garden in New York (1890), and Chicago Stadium, Chicago (1928), capable of seating 21,000. For the 1960 Olympics in Rome, the leading Italian architect-engineer Luigi Nervi created the Roman sports complex, which included the Palazzetto dello Sport with a ribbed-dome roof. The Astrodome, however, was, by comparison, gigantic, with a seating capacity of 66,000 people around a large playing field. The dome, of transparent plastic panels supported by steel lattice, has a span of 642 feet and rises 208 feet above the playing field. The entire interior is air-conditioned. Under design or construction are several new stadiums (*e.g.*, at Kansas City, Missouri, and at Philadelphia) with roofs that can be slid out of the way in sunny weather.

An important new development was the application of flexible steel cables to span large roof dimensions. Cables contributed significantly to speed of construction, to lightness of roof, and to economy of construction cost in covered stadiums. In 1956 Lev Zetlin, a U.S. engineer, developed a rigid cable system since used on many arenas, the first constructed in Utica, N.Y., in 1959. A modern stadium with this system is being designed for Saudi Arabia and is also projected for designs for the Minnesota Vikings professional football team. The Saudi Arabian structure spans 720 feet and is covered with plastic bags under pressure. The Minnesota Viking design is saddle-shaped and has a 760-foot span.

Such cable systems can span large distances; a concept for a 200,000-capacity, covered baseball stadium has been developed by Zetlin. The lower portion of the stadium would contain a small city with all required services, including rapid transit trains that would move on ramps attached to the outside walls of the stadium. The cable roof would hold the walls together and relieve weight from spaces underneath.

The concept of an enclosed stadium and the desire for greater capacities have led toward the search for improved materials and construction techniques. Modern technology, including that developed in the aerospace

industry, already possesses techniques for covering spans of thousands of feet. Such is the potential of the modern stadium design concept that proposals to span all of Manhattan Island have been seriously discussed. It is by no means unrealistic to see today's roofed stadiums, with their arrays of restaurants, exhibition halls, and other auxiliaries, as the forerunners of completely roofed, weather-controlled cities.

BIBLIOGRAPHY. For further information on stadiums, see BODO CICHY, *Erziehung zur Kunst: Architektur and Baustile* (1959; Eng. trans., *The Great Ages of Architecture: From Ancient Greece to the Present Day*, 1964); THOMAS C. ATWOOD, "The Amphitheatre, Theatre, and Stadium: Ancient and Modern," *Am. City Mag.*, 16:1–9 (1917); "For the Olympics: Flaminio Stadium, Rome, Sports Palace, Rome," *Archit. Rec.*, 124:108–111, 114–117 (1958); "Record-Span Dome Roofs Air Conditioned Stadium," *Engng. News Rec.*, 172:26–32 (1964); "Stadiums: Latest Crop of Sports Arenas from all over the World," *Archit. Forum*, 119:96–103 (1963); LAWRENCE LESSING, "Suspension Structures," *Archit. Forum*, 107:134–141 (1957).

(L.Z.)

Staël, Madame de

The brilliant personality of Madame de Staël, woman of letters, political propagandist, and conversationalist, epitomizes European thought and culture in her time, which spanned the final years of Louis XVI, the French Revolution, the Napoleonic era, and the Bourbon Restoration. Her writings include novels, plays, moral and political essays, literary criticism, history, autobiographical memoirs, and even a number of poems translated or imitated from the English 18th-century elegiacs, Goethe, and Schiller. Her chief literary importance was as a theorist of Romanticism. Chateaubriand saw in her *De la littérature* a "prospectus of Romanticism."

Giraudon



Madame de Staël, portrait by Jean-Baptiste Isabey, 1810. In the Louvre, Paris.

She was born Anne-Louise-Germaine Necker in Paris on April 22, 1766, of Swiss parents: her father was Jacques Necker, the Genevan banker who became Louis XVI's finance minister; her mother, Suzanne Curchod, the daughter of a French Swiss pastor, helped her husband in his career by establishing a brilliant literary and political salon in Paris.

The young Germaine Necker early gained a reputation for lively wit, if not for beauty. While still a child, she was to be seen in her mother's salon, listening to, and even taking part in, the conversation with that lively intellectual curiosity that was to remain her most attractive quality. When she was 16, her marriage began to be considered. William Pitt the Younger was regarded as a possible husband, but she disliked the idea of living in England. She was married in 1786 to the Swedish ambassador in Paris, Baron Eric de Staël-Holstein. It was a marriage of convenience and ended in 1797 in formal separation. There were three children: Auguste (b. 1790), who edited her complete works; Albert (b. 1792); and Albertine (b. 1796), who was allegedly fathered by Benjamin Constant.

Before she was 21, Germaine de Staël had written a romantic drama, *Sophie, ou les sentiments secrets* (1786), and a tragedy inspired by Nicholas Rowe, *Jane Gray* (1790). But it was her *Lettres sur les ouvrages et le caractère de J.-J. Rousseau* (1788) that made her known. There is in her thought an unusual and irreconcilable mixture of Rousseau's enthusiasm and Montesquieu's rationalism. Under the influence of her father, an admirer of Montesquieu, she adopted political views based on the English parliamentary monarchy. Favouring the Revolution, she acquired a reputation for Jacobinism. Under the Convention, the elected body that abolished the monarchy, the moderate Girondin faction corresponded best to her ideas.

Political views

Protected by her husband's diplomatic status, she was in no danger in Paris until 1793, when she retreated to Coppet, Switzerland, the family residence near Geneva. It was here that she gained fame by establishing a meeting place for some of the leading intellectuals of western Europe. Since 1789 she had been the mistress of Louis de Narbonne, one of Louis XVI's last ministers. He took refuge in England in 1792, where she joined him in 1793. She stayed at Juniper Hall, near Mickleham in Surrey, a mansion that had been rented since 1792 by French *émigrés*. There she met Fanny Burney (later Mme d'Arblay), but their friendship was cut short because Mme de Staël's politics and morals were considered undesirable by good society in England.

She returned to France, via Coppet, at the end of the Terror in 1794. A brilliant period of her career then began. Her salon flourished, and she published several political and literary essays, notably *De l'influence des passions sur le bonheur des individus et des nations* (1796), a treatise on the influence of the passions on happiness, which became one of the important documents of European Romanticism. She began to study the new ideas that were being developed particularly in Germany. She read the elderly Swiss critic K.V. von Bonstetten; the German philologist Wilhelm von Humboldt; and, above all, the brothers August Wilhelm and Friedrich von Schlegel, who were among the most influential German Romanticists.

But it was her new lover, Benjamin Constant, the author and politician, who influenced her most directly in favour of German culture. Her fluctuating liaison with Constant started in 1794 and lasted 14 years, although after 1806 her affections found little response.

At about the beginning of 1800 the literary and political character of Mme de Staël's thought became defined. Her literary importance emerged in *De la littérature considérée dans ses rapports avec les institutions sociales* (1800). This complex work, though not perfect, is rich in new ideas and new perspectives—new, at least to France. The fundamental theory, which was to be restated and developed in the positivism of Hippolyte Taine, is that a work must express the moral and historical reality, the *Zeitgeist*, of the nation in which it is conceived. She also maintained that the Nordic and classical ideals were basically opposed and supported the Nordic, although her personal taste remained strongly classical. Her two novels, *Delphine* (1802) and *Corinne* (1807), to some extent illustrate her literary theories, the former being strongly sociological in outlook, while the latter shows the clash between Nordic and Southern mentalities.

De la littérature

She was also an important political figure and was regarded by contemporary Europe as the personal enemy of Napoleon. With Constant and his friends she formed the nucleus of a liberal resistance that so embarrassed Napoleon that in 1803 he had her banished to a distance of 40 miles (64 kilometres) from Paris. Thenceforward Coppet was her headquarters, and in 1804 she began what she called, in a work published posthumously in 1821, her *Dix Années d'exil*. From December 1803 to April 1804 she made a journey through Germany, culminating in a visit to Weimar, already established as the shrine of

Goethe and Schiller. In Berlin she met August Wilhelm von Schlegel, who was to become, after 1804, her frequent companion and counsellor. Her guide in Germany, however, was a young Englishman, Henry Crabb Robinson, who was studying at Jena. The journey was interrupted in 1804 by news of the death of her father, whom she had always greatly admired. His death affected her deeply, but in 1805 she set out for Italy, accompanied by Schegel and Simonde de Sismondi, the Genevan economist who was her guide on the journey. Returning in June 1805, she spent the next seven years of her exile from Paris mostly at Coppet.

De l'Allemagne

While *Corinne* can be considered the result of her Italian journey, the fruits of her visit to Germany are contained in her most important work, *De l'Allemagne* (1810). This is a serious study of German manners, literature and art, philosophy and morals, and religion in which she made known to her contemporaries the Germany of the *Sturm und Drang* movement (1770 to 1780). Its only fault is the distorted picture it gives, ignoring, for example, the violently nationalistic aspect of German Romanticism. Napoleon took it for an anti-French work, and the French edition of 1810 (10,000 copies) was seized and destroyed. It was finally published in England in 1813.

Meanwhile Mme de Staël, persecuted by the police, fled from Napoleon's Europe. Having married, in 1811, a young Swiss officer, "John" Rocca, in May 1812 she went to Austria and, after visiting Russia, Finland, and Sweden, arrived, in June 1813, in England. She was received with enthusiasm, although reproached by such liberals as Lord Byron for being more anti-Napoleonic than liberal and by the Tories for being too liberal. Her guide in England was Sir James Mackintosh, the Scottish publicist. She collected documents for, but never wrote, a *De l'Angleterre:* (the material for it can be found in the *Considérations sur la Révolution française* [1818], which represents a return to Necker's ideas and holds up the English political system as a model for France).

On the Bourbon Restoration in 1814, Mme de Staël returned to Paris but was deeply disillusioned: the fall of Napoleon had been followed by foreign occupation and had in no way reestablished liberty in France. During the Hundred Days she escaped to Coppet and in September 1815 set out again for Italy. In 1816 she returned to spend the summer at Coppet, where she was joined by Byron, in flight from England after his unhappy matrimonial experience. A strong friendship developed between the two writers.

Mme de Staël's health was declining. After Byron's departure she went to Paris for the winter. Though poorly received by the returned *émigrés* and suspected by the government, she held her salon throughout the winter and part of the spring, but after April she was an invalid. She died in Paris on July 14, 1817.

Assessment

Mme de Staël's purely literary importance is far exceeded by her importance in the history of ideas. Her novels and plays are now largely forgotten, but the value of her critical and historical work is undeniable. Though careless of detail, she had a clear vision of wider issues and of the achievements of civilization. Her involvement in, and understanding of, the events and tendencies of her time gave her an unusual position: it may be said that she helped the dawning 19th century to take stock of itself.

MAJOR WORKS

Sophie, ou les sentiments secrets (1786); *Lettres sur les ouvrages et le caractère de J.-J. Rousseau* (1788; *Letters on the Works and Character of J.-J. Rousseau*, 1789); *Courte réplique à l'auteur d'une longue réponse* (1789); *Jane Gray* (1790); *Réflexions sur le procès de la Reine* (1793); *Réflexions sur la paix, adressées à M. Pitt* and *Réflexions sur la paix intérieure* (both 1795); *Recueil de morceaux détachés* (1795), including the first edition of the *Essai sur les fictions; De l'influence des passions sur le bonheur des individus et des nations* (1796; *A Treatise on the Influence of the Passions upon the Happiness of Individuals and of Nations*, 1798); *De la littérature considérée dans ses rapports avec les institutions sociales*, 2 vol. (1800; *A Treatise of Ancient and Modern Literature*, 1803, and *The Influence of Literature upon Society*, 2 vol., 1812); *Delphine*, 4 vol. (1802; *Delphine*, 3 vol., 1803); *Mémoires sur la vie privée de mon père*, first as preface to the *Manuscrits de M. Necker* (1804; later as separate book, 1818; *Memoirs of the Private Life of My Father*, 1818); *Corinne, ou l'Italie*, 2 vol. or 3 vol. (1807; *Corinna, or Italy*, 3 vol. or 5 vol., 1807); *De l'Allemagne*, 3 vol. (first edition suppressed 1810, republished 1813 and 1814; *Germany*, 3 vol., 1813); *Réflexions sur le suicide* (1813; *Reflections on Suicide*, 1813); *Zulma et trois nouvelles, precede d'un Essai sur les fictions* (1813; *Zulma and Other Tales . . .* , 1813); *Considérations sur les principaux événemens de la Révolution française*, 3 vol. (1818; *Considerations on the Principal Events of the French Revolution*, 1818); *La Doctrine des ultra-royalistes réfutée . . . ou Nouvelle réponse à la note secrète exposant les prétextes et le but de la dernière conspiration* (1818); *Dix Années d'exil* (1821; *Ten Years' Exile*, 1821); *Essais dramatiques* (1821), comprising *Agar dans le désert, Geneviève de Brabant, La Sunamite, Le Capitaine Kernadec, La Signora Fantastici, Le Mannequin,* and *Sapho; Des circonstances actuelles qui peuvent terminer la Révolution, et des principes qui doivent fonder la République en France* (1906). The first collected edition of the *Oeuvres complètes de Mme. la Baronne de Staël*, 17 vol. (1820–21), was supplemented by a collection of *Oeuvres inédites . . .* , 3 vol. (1821), and followed by another edition of the *Oeuvres complètes*, 3 vol. (1836). An edition of Madame de Staël's *Correspondance générale* was begun in 1960.

BIBLIOGRAPHY. LADY BLENNERHASSETT, *Frau von Staël, ihre Freunde und ihre Bedeutung in Politik und Literatur*, 3 vol. (1887–89; Eng. trans. by J.E.G. CUMMING, 1889; French trans. by A. DIETRICH, 1890), the basic biography of Madame de Staël although its documentation is still incomplete; D.G. LARG, *Madame de Staël, la vie dans l'oeuvre (1766–1800)* (1924; Eng. trans. by V. LUCAS, 1926), and *Madame de Staël, la seconde vie (1800–1807)* (1928), the first comprehensive view of Madame de Staël both as a writer and as a historic figure; J.C. HEROLD, *Mistress to an Age: A Life of Madame de Staël* (1958), the best English-language introduction to Staëlian studies; A. LANG, *Une Vie d'orages: Germaine de Staël* (1958), a controversial and controverted biography, well worth reading; R. ESCARPIT, *L'Angleterre dans l'oeuvre de Madame de Staël* (1954), an analysis of Madame de Staël's description and analysis of English history, institutions, and customs; COMTESSE J. DE PANGE, *Le Dernier amour de Madame de Staël* (1944), and V. DE PANGE, *Madame de Staël et le Duc de Wellington* (1962), two studies evolved from family documents (the Countess of Pange is the great-granddaughter of Madame de Staël) that are essential to Staëlian studies.

(R.Es.)

Staging and Stage Design

Staging is the total concrete shape of a dramatic event; it embraces the arrangement of words, performers, dance, music, setting, costume, makeup, lighting, and properties for maximum theatrical effectiveness.

Role of staging and stage design in the performing arts

Through most of its history the theatre has encompassed all of the communal performing arts in a creative celebration of the richness of human activity and the aspirations of mankind. Each period has emphasized some forms and de-emphasized others, but the civilizations in which theatre has played a vital role in the lives of the people have marshalled the most varied tools of dramatic expression and did not separate text, music, and dance. In the healthiest of theatrical times, staging and design were not specialties handled by departmentalized experts; the direction of plays as a separate profession is a mid-19th-century innovation. Yet, Western audiences have become so thoroughly indoctrinated with the formulas of the literary drama of the late 19th and 20th centuries that most theatre of earlier periods and of the East are incomprehensible to them. The splintering of the components of theatre into serious drama, comedy, and farce or into the spoken word, dance, and music is a late development in theatrical history. This fragmentation is an outgrowth of the Age of Reason of 18th-century Europe, aggravated by the Industrial Revolution of the 19th century, and completed by the electronic age of the 20th century. It is difficult to project oneself into past times when theatre naturally embraced the word, dance, and music and dealt simultaneously with solemn and joyous human concerns. Because only dramatic texts survive, the prejudice in favour of the word is sanctified. Even the terminology in current use is conditioned by the small segment of theatrical history the vestiges of which are still prevalent. The word spectacle conjures up ostentatious display, gratui-

tous pageantry, and visual fireworks not proper to the dignified mounting of dramatic literature. Other labels attached to the nonliterary aspects of the theatrical event are equally revealing. "Divertissement" refers to a frivolous interpolation of music and dance in the interstices between the acts of weighty drama. "Decor" is the ocular icing. "Embellishments" refers to the settings and costumes, in fact, to all the physical appurtenances of the stage. "Staging" is the passive handmaiden, the interpretive servant, the midwife who tenderly deposits the dramatist's child on the boards of a public playhouse. Since contemporary theatre has shrunk to a marginal option in the life of the times and large integral functions of theatre have been siphoned off into the mass media, the mere semantics involved in explaining living theatre through the centuries becomes a problem.

In this article an attempt will be made to reverse the customary procedure of brushing off the primitive beginnings as quaint worship and childish play, of heralding the serious dramatics of the Greeks, and of progressing upward to the giants of modern realism. Instead, theatre will be considered as it has existed through most of its history—to wit, as a communal celebration. The article begins, therefore, with an extensive discussion of the history of staging practices, followed by theoretical analyses of shifting styles and other general aesthetic questions. Later, costuming and makeup and lighting receive separate attention. Technological aspects of sound are covered in the article ACOUSTICAL ENGINEERING. Though theatre housings are discussed insofar as they influence stage design, they are treated at greater length in another article, THEATRES AND STAGES. Related material will also be found in the articles THEATRICAL PRODUCTION; ACTING; and DIRECTING; as well as THEATRE, ART OF.

This article is organized as follows:

I. History of staging
Ancient Greece and Rome
Asian theatre
The Middle Ages
The Renaissance
The Baroque
The 19th century
The beginnings of modern theatre
Modern staging and design
II. Scenic design
Development of scenic design
Procedures
III. Stage lighting
Early forms of stage lighting
Electrification
Light sources
Lighting controls
Contemporary lighting techniques
IV. Stage machinery
V. Theatrical costume design
Historical development
Contemporary costume-design procedures
VI. Theatrical makeup
Early development
Contemporary practices

I. History of staging

Origins of dramatic representation

Since the anthropological data on primitive drama is meagre, fanciful theories abound. One major misconception looks upon early rites as the tribe's excuse for emotional festivities and furious orgies, but research does not support the image of the childlike savage on colourful holiday from his routine chores. Sympathetic magic was a deadly serious, practical instrument for controlling the environment. The primitive's tools for influencing nature were not the same as the tools of technology, and the contemporary notion of theatre as a leisure-time safety valve should not be imposed on primitive ritual.

There is an apparent uniformity in the earliest development of dramatic representation, even among groups geographically and ethnographically separated from each other. The origins of drama seem to be religious, and the beginnings of religious rites often are in magic. When man feels that there are elements in his environment that he cannot control, he tends at first to believe that his ancestors are the spirits who hold a magical control. An offering of the first fruits of harvest was not only a thanksgiving but also insurance that their ancestors would bring them further harvests. These first fruits were offered ritually to the spirits of the dead, among the first nonreal forces man identifies. Corn and vegetable spirits and what are labelled "totemistic beliefs" are dependent on a belief in immortality, which may be the first collective idea shared by man. In the first stage, men pray to the dead and not to abstract spirits; in gratitude they make harvest offerings to the dead. It is only at a later stage that such generalizations as food gods appear, that man symbolizes his desires, transforming his ancestors into the shapes of universal spirits.

The earliest rituals were probably staged at regular intervals, often with the same members of a tribe playing the same roles. Men would imitate the behaviour of the spirits by wearing masks and costumes and dancing in certain patterns that were passed on from one generation to another. The second stage was reached when man passed from ritual incantation to developing his own myths about the spirits. In Western culture, the worshippers of Dionysus—the god of nature's procreative power and the ecstasy related to it—were the first group in ancient Greece to develop myths about their god. Whoever surrendered to the god forgot about himself and turned into another being. This belief in transfiguration explains why, from the beginning, the Dionysiac worshippers wore masks and disguises. It is commonly held that each year the worshippers became more complex in their celebration of Dionysus, until eventually they staged athletic contests, dances, and simple drama during their festivals.

Whatever the origin, the earliest phases of the theatre involved masks, costumes, and properties. Primitive societies tended to use the same piece of land for all rituals, although hunting tribes moved too often to have a particular sacred ground. A stage, and particularly scenery, came at a much later point in the development of drama. In all Western cultures, it seems that a permanent stage and scenery are introduced when drama starts to shift from a magical-religious institution to one that incorporates the lives of the members of the community.

ANCIENT GREECE AND ROME

Greece. During the earliest period of theatre in ancient Greece, when the poet Thespis—who is credited both with inventing tragedy and with being the first actor—came to Athens in 534 BC with his troupe on wagons, the performances were given in the agora (*i.e.*, the marketplace). The spectators sat on wooden stands until 498, when the stands collapsed and killed several members of the audience. Detailed literary accounts of theatre and scenery in ancient Greece can be found in *De architectura*, by the Roman writer Vitruvius, of the 1st century BC, and in the *Onomasticon*, of the next century, by the Greek scholar Julius Pollux. As they were writing several hundred years after classical theatre, however, it is questionable how accurate are their descriptions.

Little survives of the theatres in which the earliest plays were performed, but essential details have been reconstructed from the architectural evidence of the Theatre of Dionysus in Athens, which has been remodelled several times. The centre of the theatre was the original dancing place, the flat, circular space with the altar of Dionysus, called the *orchēstra*. In the centre stood a platform with steps (*bēmata*) leading to the altar (*thymelē*). Nearby was the temple out of which the holy image would be carried on festival days so the god could enjoy the plays.

The earliest theatres

Theatrical representations, coloured by religious sentiment, directed their appeal toward the total community, so that when these theatres were built, their first need was enough space for a large audience. In the beginning, entrance was free; after they started charging admission, poor citizens were given entrance money, since attendance was practically compulsory. It seems reasonable to assume that, because of the size of the theatres, there was a raised platform for actors to make them easier to see and to separate them from the chorus. In later times there was a high stage, with a marble frieze below and a short flight of steps up from the orchestra.

The earliest productions did not have a background building. The actors dressed in the *skēnē* (from which the word scene is derived), a small tent, and the chorus and actors entered together from the main approach, the *parodos*. The earliest properties, such as altars and rocks, could be set up at the edge of the terrace. This terrace, according to recent archaeological studies, was probably six feet (about two metres) above the level on which the earliest temple stood. The first extant drama for which a large building was necessary was Aeschylus's trilogy the *Oresteia*, produced in 458 BC. There has been a controversy among historians as to whether the *skēnē* was set up inside a segment of the orchestra or outside the edge of the orchestra. The *skēnē* was probably a long, simple building at the left of the orchestra terrace.

Origin of the scenic facade

In the first period of Greek drama, the principal element of the production was the chorus, the size of which appears to have varied considerably. In Aeschylus' *Suppliants*, there were 50 members of the chorus, but in his other plays there were only 12, and Sophocles increased it to 15. The size of the chorus, as the ritual element of drama diminished, became smaller in the 5th century. Since the number of actors increased as the chorus shrank, and the plots of the dramas became more complex, doubling of roles became necessary. On a completely open stage such substitutions were delayed, and the suspense of the drama was dissipated. Dramatic plausibility was also vitiated when gods and mortals, enemies and friends, always came from the same direction. The addition of a permanent scenic facade, with three doors, more than doubled the number of entrances and gave the playwright more freedom to build dramatic tension. About 425, a firm stone basis was laid for an elaborate building, called a *stoa*, consisting of a long front wall interrupted at the sides by projecting wings, *paraskēnia*. The spectators sat on wooden benches, which may have been arranged as a polygon or as a semicircle. The seats of honour were stone slabs with inscriptions on them assigning them to the priests. The upper rows were benches of movable planks supported by separate stones planted in the ground. The single wedge-shaped divisions were separated from each other by radiating aisles.

The background decoration was originally a temporary structure leaning against the front wall of the *stoa*. A wooden structure served as a skeleton, which was covered with movable screens. These screens were made of dried animal skins tinted red; it was not until Aeschylus that canvases in wooden frames were decorated according to the needs of a particular play. Aristotle credits Sophocles with the invention of scene painting, an innovation also ascribed by others to Aeschylus. It is notable that Aeschylus took an interest in staging and is credited with the classic costume design. Simple Greek scenery, however, was comparable with that of the 20th century; the impulse to visualize and particularize the background of the action became strong. Painted scenery was probably first used in production of the *Oresteia;* some 50 years afterward a second story was added to the wooden scene building. A wooden colonnade, or portico, the *proskēnion,* was placed before the lower story of the building. This colonnade, which was long and low, suggested the exterior of either a house, a palace, or a temple. Painted screens were set between the columns of the front of the *proskēnion,* and they indicated the locale rather than literally representing it.

The first painted scenery

In the beginning, scenery was probably altered slightly during the intermission between the four plays of a tetralogy or during the night between two of the festival days. By the latter part of the 5th century, scene changing was accomplished through the use of movable painted screens. Several of these screens could be put up behind one another so that when the first one was removed, the one immediately behind appeared. There are indications that in the theatre of Megalopolis, Greece, there were tracks coming out of a storeroom on the side, not only for movable screens but for a whole wooden scene building.

Soon after the introduction of the facade, plays were uniformly laid before a temple or a palace. To indicate a change of scene, the *periaktoi* were introduced. These were upright three-sided prisms set flush with the palace or temple wall on either side of the stage, mounted on wooden pegs, the sides painted to indicate different localities. There were several theatrical conventions that were observed in Greek theatre; one was that only the right *periaktos* was turned to indicate a different locality in the same town. The side entrances of the theatre, the *parodoi,* which were near the *periaktoi,* also took on conventions of significance. At the right one, the actors entered as if coming from the city or harbour; from the left, as if coming from the country.

The *periaktoi*

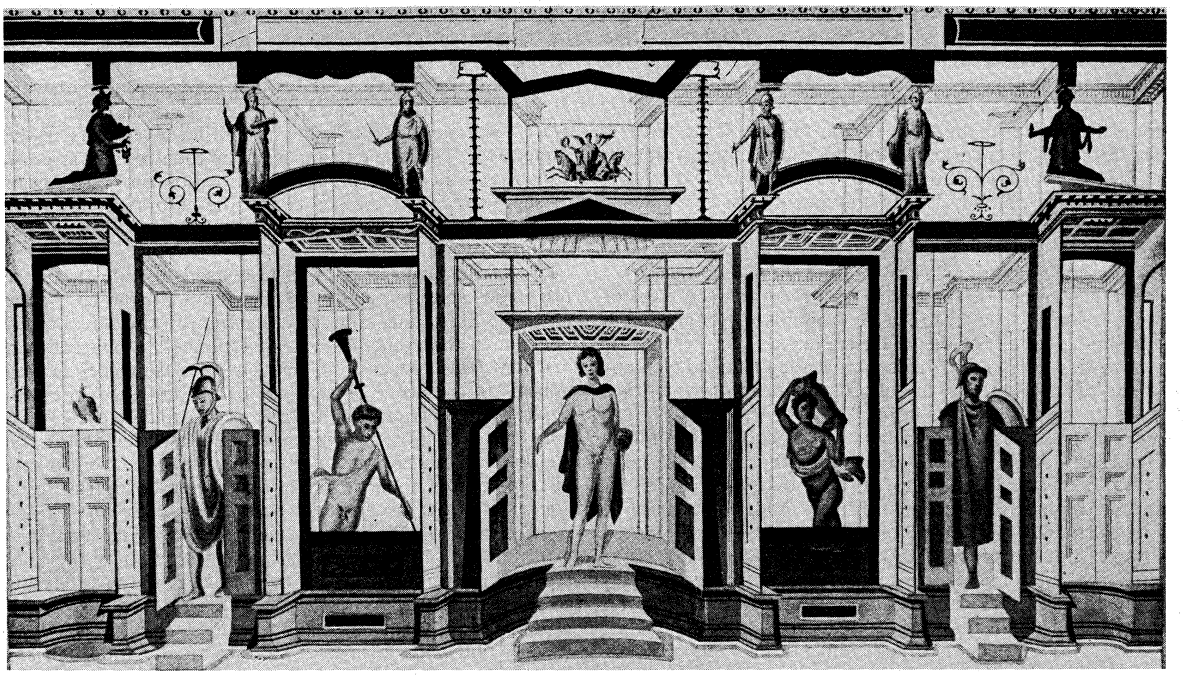
The permanent facade was also used to hide the stage properties and the machinery. Evidence for the use of the so-called flying machine, the *machina,* in the 5th century is given in the comedies of Aristophanes; a character in his play *Peace* ascends to heaven on a dung beetle and appeals to the scene shifter not to let him fall. The *machina* was a derrick. In the time of Euripides it was used conventionally for the epilogue. At that point the crane was used so that a god could descend from heaven to sort out the complications in the plot at the end of a tragedy, a convention that became known as a *deus ex machina*—"god from a machine." The lavish use of flying machines is testified to by the poet Antiphanes, who wrote that tragic playwrights lift up a machine as readily as they lift a finger when they do not have anything to say.

Wheeled platforms or wagons, called *ekkyklēma,* were used to show the results of actions in the interior of a house, such as wheeling out murdered people. It was, like the *periaktoi,* an expedient for open-air theatre, in which the possibilities for creating more realistic illusions were severely limited. Because of the tiered seats of the auditorium, a realistic picture of an interior scene under a roof could not be shown. So the Greeks, to represent the interior of a palace, for example, would wheel out a throne on a podium. These wheeled platforms were usually round, although recent archaeological evidence suggests that some may have been square. New machines were added in the Hellenistic period, by which time the theatre had almost completely lost its religious base. Among these machines was the *hēmikyklion,* a semicircle of canvas depicting a distant city, and a *stropheion,* a revolving machine, used to show heroes in heaven or battles at sea.

There was no one Greek theatre; the existing buildings, or those for which there is archaeological evidence, divide into four groups. The first is the pre-Aeschylean wooden theatres, which, though previously believed to have been circular, recent evidence suggests might have been trapezoidal. The second is the classical Athenian theatre of the 5th century BC. In all probability the classical wooden theatre with a stone foundation was of two stories. The upper story, the *episkēnion,* was primarily utilized for the stage machinery. The raised stage was probably called the *logeion,* the speaking place, although that word and *theologeion,* speaking place of divinities, have both been applied to the top of the *proskēnion* or first story of the scene building itself. The third group of theatres were the Hellenistic, built from the 4th century on, mostly in Greek territories. The Theatre of Dionysus in its present state is a product of 4th-century renovations, a rebuilding under the Roman emperor Nero in the 1st century AD, and alterations in the 3rd century AD. The last group was the Greco-Roman, found as Greek culture was fading and the Roman Empire spreading.

Rome. The development of the theatre, following the development of dramatic literature, was slower in Rome than in Athens. The essential distinction between Roman and Greek stage performances was that the Roman had no deep religious convictions. Despite the fact that the spectacles were technically connected with festivals in honour of the gods, the Roman audience went to the theatre for entertainment. The circus was the first permanent public building for spectacles, which included chariot races and gladiatorial fights. When Etruscan dancers and musicians were introduced in 364 BC, they performed in either the circus, the forum, or the sanctuaries in front of the temples. The players brought temporary wooden stands for the spectators. These stands developed into the Roman auditorium, built up freely from the level ground.

Distinctions between Roman and Greek theatre



Roman stage, or *scaenae frons*, with the presentation of a mime; drawing by G. von Cube, 1905, of a fresco in the Caseggiato dei Gladiatori, Pompeii.
By courtesy of the Deutsches Archaeologisches Institut, Rome; photograph, Art Reference Bureau

The most important feature of the Roman theatre, as distinct from the Greek theatre, was the raised stage on which all the performers appeared. As every seat had to have a view of the stage, the portion occupied by the seating (*cavea*) was limited to a semicircle. The *orchestra*, immediately in front of the stage, seated distinguished visitors. As in Greek theatre, the scene building behind the stage, the *scaenae frons*, was used both as the back scene and as the actors' dressing room. The *scaenae frons* was no longer painted in the Greek manner but tended to have architectural decorations combined with luxurious ornamentation. The audience sat on tiers of wooden benches, *spectacula*, supported by scaffolding. There was no curtain; the back scene, with its three doors, faced the audience.

The *phlyakes* stages

When the popular comedies or farces of southern Italy were introduced to Rome, they came with their own distinctive type of stage—the *phlyakes* stage. Comedies in Italy were mimes, usually parodies of well-known tragedies, and the actors were called *phlyakes*, the gossips. They used temporary stage buildings of three main forms. One was the primitive low stage, a rough platform with a wooden floor on three or four rectangular posts. The second was a stage supported by low posts, covered with drapery or tablets; sometimes steps led up to a platform and a door was indicated. The third type was a higher stage supported by columns, without steps but usually with a back wall. The stages often had a short flight of five to seven steps in the centre, leading to the podium. The forewall, covered with drapery, was often decorated, and the background wall usually had objects hanging from it. The rear wall sometimes had other columns, besides the ones set at the corners, as well as doors and, in several cases, windows to indicate an upper floor. The door was usually behind a heavily decorated porch, with a sloping or gabled roof supported by beams and cross struts. Among the furnishings there were usually trees, altars, chairs, thrones, a dining table, a money chest, and a tripod of Apollo (*i.e.*, an oracular seat). This catalog demonstrates that the stage properties were simple, probably because the troupes wandered from town to town. The stage was set up in the marketplace in the smaller towns and in the *orchēstra*s of Greek theatres in the larger cities.

Coincident with the development of the *phlyakes* stages, and under the inspiration of Hellenistic colonists, the Romans began to build stone theatre buildings. A Hellenistic theatre was converted into a Roman one by making the following three fundamental changes: a deep stage was erected; the rear wall was made higher and decorated in Roman style with pillars, statues, and entablatures; and vaulted passageways replaced the open *parodoi*.

Roman stage conventions

The actors entered the playing space from one of the house doors or the side entrances in the wings. Among the stage conventions, the side entrance on the audience's right denoted the near distance and the one on stage right the farther distance. If a scene took place in a town, for instance, an actor exiting audience right was supposed to be going to the forum; if he exited to their left, he might be going to the country or the harbour. *Periaktoi* at the side entrances indicated the scenery in the immediate neighbourhood. If the play required a character to move from one house to another without bringing him on the stage, which represented the street, the actor was supposed to use the back door and the *angiportum* (*i.e.*, an imaginary street running behind the houses). Since interior scenes could not be represented on level ground, all action took place in front of the houses shown in the background. If a banquet was to be depicted, the table and chairs would be brought on stage for one scene and removed at the end of the scene. Costumes were formalized, but real spears, torches, chariots, and horses were used.

In the 1st century BC, theatre buildings and scenery became more elaborate. The invention of the arch and the use of concrete for vaulting allowed immense stone structures to be built on level ground. In 55 BC the first stone theatre was built, in Rome, by the general Pompey; the Roman writer Pliny wrote that it held 40,000 spectators and was used for "bloody shows," with 500 lions and 20 elephants killed at the grand opening. To fulfill the government policy of keeping the people entertained, theatres were built all over the Roman Empire. Many of the theatres were on flat sites, and the tiers of seats were supported by masonry, through which passages, called *vomitoria*, led from outside to the *orchestra* or to a corridor at the top of the auditorium. From there, by means of steps, a member of the audience could find his way to his seat.

The addition of the roof over the stage

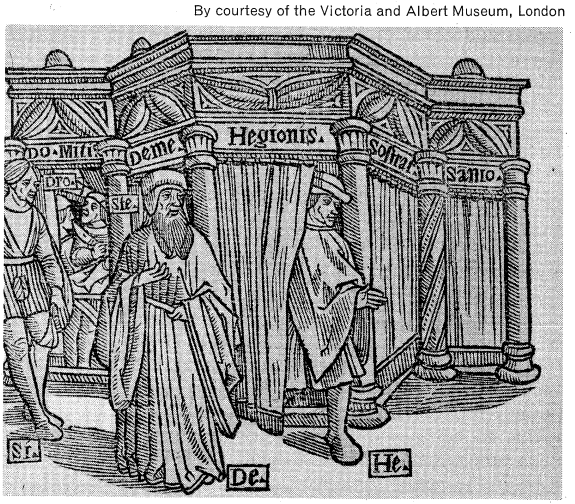
A wooden roof was added over the stage, protecting actors and scenery from the weather. The auditorium was also covered by coloured linen awnings attached to fixed masts in the enclosing wall. The presence of the roof in some theatres led to two particular innovations. The first was that as it was now possible to cover the entire theatre

with an awning, it was being brought reasonably close to the indoor theatre. The second was the front curtain, the *aulaeum*, which was probably on a fixed frame that was dropped into a recess in the stage floor during the performance. There was also a second type of curtain, the *siparium*, used to cover parts of the facade behind the actors.

Setbacks in the development of the theatre building did occur, as when a stone theatre was actually demolished in a period of austerity because seats were thought to be unworthy of the stern simplicity of the Roman citizens. Preserved in Rome is the stone theatre of Marcellus, begun by Julius Caesar and finished by Augustus in 13 BC. Most important provincial towns had at least one theatre, and often a small covered theatre for inclement weather was constructed near the large open theatre.

In general, the history of the Greek and Roman theatre is so rich and varied that each successive period of European culture has found some aspect of it to use as an inspiration or model. Even the periods that resented the ancient theatre and its pagan overtones found some aspect to explore and use for their own. In the Middle Ages, for instance, although the tragedies and the theatre buildings of the Greeks and Romans were little known, the 2nd-century-BC Latin comedies of the Roman Terence remained alive. And the architectural writings of the 1st century AD of the Roman Vitruvius became the model for the theatre building of the later Renaissance and early Baroque periods.

By courtesy of the Victoria and Albert Museum, London



Scenery for *Adelphi* (*The Brothers*) by Terence, from an illustration in the edition of Terence's plays published in Lyons, 1493. In the Victoria and Albert Museum, London.

ASIAN THEATRE

Although the emergence of Asian theatre was not simultaneous with that of ancient Greece and Rome, it merits discussion here rather than as an appendage to the history of Western drama. Asian theatre, to speak in generalities, is an appeal to both the senses and the inner eye of the audience. It is "magical" theatre in the most sophisticated sense of the word, exerting charm and power in a combination of realism and illusion. Perhaps the primary difference between Asian and Western drama stems from its attitude toward the spectator; Asian drama appeals to the varied sensibilities of its audience. It is the conjunction of emotional, intellectual, and spiritual influences and as such affects each spectator differently. Western drama hinges on finding a common denominator to manipulate its audience as a single body, whether it is composed of the Renaissance aristocracy or the 20th-century middle class. Asian drama, in contrast, allows a member of the audience to meet the theatre on an individual basis.

India. India is the point of origin for most of the theatre in Asia and had developed its dance and drama by the 8th century BC. According to Hindu holy books, the gods fought the demons before the world was created; and Brahmā asked the gods to re-enact the battle among themselves for their own entertainment. Once again the demons were defeated, this time by being beaten with a flagstaff by one of the gods. To protect theatre from demons in the future, a pavillion was built, and in many places in India today there are roofed platforms with a flagstaff next to the stage to mark the location of performances.

The myth of the development of dance drama was that Brahmā ordered the two elements combined; as a result, the words for "dance" and "drama" are the same in all Indian dialects. Early in Indian drama, however, dance began to dominate the theatre. By the beginning of the 20th century there were few performances of plays, though there were myriads of dance recitals. It was not until its political independence in 1947 that India started to redevelop the dramatic theatre.

In the 4th century a codification was written of the *śāstra*, or the staging conventions of the dance. It lists not only the costumes, makeup, gestures, and body positions but also any plots considered unsuitable, and it is the most complete document of stagecraft ever compiled. There is no scenery in Indian dance, although there are usually a few properties such as a three-foot-high brass lamp. A curtain is used, however, by troupes that dance *kathākali*, an ancient danced drama of southwestern India. The curtain itself is a cloth rectangle that is held between the stage and a large lamp by two stagehands.

The dancers perform a group of preliminary dances behind the curtain until an important entrance called "peering over the curtain." In this, a character fans the lamp by pulling the curtain in and out until the flames are spectacularly high. The dancer, still hiding his face, displays his hands and legs at the borders of the curtain. At the climactic moment the dancer pulls the curtain aside, displaying his awesome makeup, and the stagehands remove the curtain until the next dramatic entrance.

Classical Indian drama had as its elements poetry, music, and dance, with the sound of the words assuming more importance than the action or the narrative; therefore, staging was basically the enactment of poetry. The reason that the productions, in which scenes seem to follow in arbitrary order, seem formless to Westerners is that the playwright, even today, utilizes the symbolic language of simile and metaphor to fashion his drama. Because of the importance of the poetic line, a significant character is the storyteller or narrator, who is still found in most Asian drama. In Sanskrit drama the narrator was the *sūtra-dhāra*, "the string holder," who laid the scene, gave background information, and interpreted the actors' moods. Another function was performed by the narrator in regions in which the aristocratic vocabulary and syntax of the main characters, the gods and the nobles, was not understood by the majority of the audience. The narrator, first through the use of pantomime, later through comedy, made the story clear and offered a relief from the tension of the drama. In general, though, as everyone learned the stories after a number of years, the emphasis shifted to how passages were acted rather than on what was being portrayed.

The narrator's role

A new Indian theatre that started about 1800 was a direct result of the English colonization. With the addition of dance interludes and other Indian aesthetic values to Western theatre, India today is developing a national drama. Two examples of "new" theatre staging are the Prithvi Theatre and the Indian National Theatre. The Prithvi Theatre, a Hindī touring company founded in 1943, treats the stage as similar to a film. It utilizes dance sequences, incidental music, frequent set changes, and extravaganzas of movement and colour. The Indian National Theatre, founded in Bombay in the 1950s, takes an entirely different route to appeal to audiences; it gives performances all over India, in factories and on farms. Its themes usually involve a national problem, such as the lack of food, and the troupe's style is a mixture of pantomime and simple dialogue. They use a truck to haul properties, costumes, and actors; there is no scenery.

China. The most noticeable contrast between China and other Asian countries is that it produced virtually no dance. The classic theatre of the Chinese is called "opera"

because the dialogue is punctuated with arias and recitatives. Of the amazingly detailed written record of Chinese theatre, the first reference to opera was during the T'ang dynasty (618–907). The development of the opera style popular today took place during the Manchu rule of the 19th century. The Empress Dowager, the last hereditary ruler of China, was so enamoured of opera that she had a triple-deck stage constructed in the summer palace at Peking. Each level stood for either heaven, hell, or earth.

The most important individual in Chinese theatre of the 20th century, Mei Lan-fang, an actor and producer, was the first to apply to it academic principles of scholarship in reviving lost or forgotten masterpieces and opera forms. He restored historical accuracy and improved old costumes.

The Chinese theatrical performance

In general, Chinese theatrical performances start in the early evening and conclude after midnight. The performance itself consists of several plays and scenes from the best known dramas. The audience drinks tea, eats, and talks, and there are no intermissions. The stage itself has a curved apron, covered only by a square rug. On one side is a box for the orchestra, who play throughout the evening. There is neither a curtain nor any setting to speak of other than a simple, painted backdrop. This absence of scenery contrasts with the complicated and colourful costumes and makeup of the actors.

The behaviour of the audience differs from Western theatre in that its members stop talking to each other only at climactic moments. The actors are concerned with their movements only when they are at the centre of the stage; when they stand at the sides waiting for their parts, they drink tea and adjust their costumes in full view of the audience. An interesting interpretation of this behaviour was the view of the 20th-century German dramatist Bertolt Brecht that a Chinese actor constantly keeps a distance between himself, his character, and the spectator; his performance is always on a conscious level, with all his emotions transposed.

Property men walk around on stage setting up properties for the next play before the preceding one is finished. There are usually very few properties, only a table and a few chairs. A chair may act as a throne, a bench, a tower if an actor stands on it, a barrier if the heroine stands behind it, and so on. A curtain suspended in front of two chairs represents a bed. Doors and stairs are always suggested by pantomime; an actor always mimes opening a door and taking a high step when he "enters" a room.

There are a number of stage conventions; all entrances, for instance, are from a door stage left, and all exits through a door stage right. After a scene of combat, the man who is defeated exits first. Wind is symbolized by a man rushing across the stage carrying a small black flag. Clouds painted on boards are shown to the audience to represent either the outdoors or summer. Fire, however, is always represented realistically, either by the use of gunpowder or pyres of incense. The Chinese feel that the constant use of realism in Western drama atrophies the imagination.

China has few modern dramatists whose work is known abroad; from a designer's view there is one of note—T'ien Han, who after World War II wrote a 21-act tragedy called *Ballad of the Fair Women*. Its production involved the use of what the author named "synthetic propaganda," in which slides and motion pictures are combined with the drama.

Japan. Japan is the only country in Asia whose theatre has never undergone significant renovation in 1,300 years of continuous history. When an attempt is made to recreate the original production of a Greek tragedy or even a play by Shakespeare, its historical accuracy can only be an educated guess. In Japan, the traditions of stagecraft and costumes for both drama and dance have remained unaltered. Japanese staging is far more sophisticated than Western theatre; by the time of Shakespeare, for instance, the Japanese had already invented a revolving stage, trapdoors, and complex lighting effects.

Nō theatre

Although there are many kinds of theatre in Japan, the best known are the Nō and the Kabuki. Nō, the second oldest form of theatre, was developed by the late 1300s and first seen by Westerners in the 1850s. It developed from two sources: the first was the *dengaku*, a rice planting and harvesting ritual that was transformed into a courtly dance by the 14th century. The second source was the *sarugaku*, a popular entertainment involving acrobatics, mime, juggling, and music, which was later performed at religious festivals.

Two performers and adherents of Zen Buddhism in the late 1300s, Kan-ami and his son Zeami, combined the *sarugaku* elements with *kuse-mai*, a story dance that uses both movements and words. Soon *dengaku* elements were added, and the distinctive Nō style slowly emerged. Like the Zen ways of tea ceremony, ink drawing, and other arts, Nō suggests the essence of an event or an experience within a carefully delineated art form. There are scores of Nō theatres in Japan today, even though the design of a Nō theatre is so stylized that it is not usable for other types of performances. The Nō stage is a platform completely covered by a curving temple roof. The audience sits on three sides of the stage and is separated from the stage by a wide garden of gravel, plants, and live pine trees.

Masks are used, though they are restricted to the principal dancer and his companions. The male roles are made up in brilliant stiff brocades and damasks, perfect for the grandiose posturing of the actors. The female roles are played in bright flowered brocades. The outer robes of both sexes are of a fine-woven gauze, light and suitable for the gliding dances when sleeves and fans float in the air. Mask carving is an important art in Japan, and Nō masks add considerable beauty to the gorgeous traditional robes. Most costumes are based on the classic court hunting dress of the Heian (794–857) and Kamakura (1192–1333) periods.

A more detailed discussion of Nō staging practices will be found in the article NO THEATRE.

Kabuki staging techniques

Kabuki troupes, originally composed only of women, developed in the early 17th century. By the 1680s Kabuki had become an established art form, and curtains and scenery were introduced. Kabuki was first seen in western Europe during the latter part of the 19th century, but it was not until the 1920s that it was accepted there as something more than quaint. The Russian film director Sergey Eisenstein, after seeing the Kabuki troupe that toured the Soviet Union in 1928, interpreted it as a brilliant example of effects that could be achieved on film. Sound, movement, space, and voice functioned as elements of equal significance, not as accompaniments of each other. Kabuki staging devices were tried out in theatres in the Soviet Union, France, and Germany, and one Kabuki actor, in turn, brought back Russian techniques that influenced the Japanese theatre.

One Kabuki theatre in Tokyo is the largest legitimate theatre in the world, with a 91-foot- (28-metre-) wide stage and seating for 2,599 people. Running through the audience and connecting the stage with the rear of the auditorium is the platform runway, called the *hanamichi*. It is utilized for significant entrances and exits, processions, and dance sequences. Its purpose is to unite the actor and audience by moving the actor out of the decorative background. Originally there were two runways, with a connecting bridge at the rear of the auditorium. Because of economic pressure to seat more people and the influence of Western architecture, the second *hanamichi* was removed in the early 1900s. The runway cannot be compared to European or American central staging or use of theatre aisles because the Kabuki platform allows the actor to remain separate from the spectators.

The scenery for Kabuki may be as elaborate and complex as that found anywhere; the stage, for instance, may be a house, a forest, and a river simultaneously. Some settings are triple-level palaces, with the actors using all levels at once; others have only a simple backdrop similar to the Nō.

All of its costumes are of the period when Kabuki reached the peak of its perfection, during Edo times (1603–1868). Wigs and makeup carefully conform to classical tradition, enabling habitual playgoers to recognize the type of play and characters at a glance. Many of

the costumes are much exaggerated; all are designed to accentuate dramatic movement. Courtesans and heroes, for instance, wear stilts that raise them several inches off the ground.

Kabuki staging practices are discussed at greater length in the article KABUKI THEATRE.

Bali. Balinese theatre is included here as representative of theatre in the smaller nations of Asia, such as Thailand, Cambodia, Laos, Malaya, the Philippines, and Vietnam, in all of which drama consists almost exclusively of dance. Balinese dancing may take place anywhere; usually it is executed in front of a temple or a pavillion used for community meetings. The audience sits on three sides of the performers; if they are in a clearing, the audience sits in the round. The musicians, called the gamelan, sit on one side of the stage area. There are neither settings nor visible indications of scene changes; location is suggested by the dialogue or the facial expressions and gestures of the actors. A scenic "device" is employed at the beginning of each section of the dance; for instance, the dancer makes a gesture called "opening the curtain." The hands, palms out, are in front of the face; they separate on a diagonal line to reveal the figure, stopping only when the full, formal posture of Balinese dance is reached.

Balinese scenic "devices"

THE MIDDLE AGES

In terms of performances and theatres, Roman drama reached its height in the 4th century AD. Although it continued as a government-supported institution for 200 more years, it had already encountered difficulties that would lead to its demise. One of the two main sources of opposition was the growing Christian Church. From about AD 300 on, churches tried to dissuade Christians from going to the theatre, and in AD 401 the fifth Council of Carthage decreed excommunication for anyone who attended performances on holy days. Actors were forbidden the sacraments of the church unless they gave up their profession, a decree not rescinded in many places until the 18th century. An important contradiction, though, can be seen in an edict of Charlemagne (*c.* 814) stating that no actor could put on a priest's robe except under penalty of punishment or banishment. This indicates that mimes either were continuing their secular ridicule of the church or were endeavouring to pacify religious sentiment by the performance of "godly" plays.

The other main source of opposition that the Roman theatre could not stave off was the invasion of the barbarians from the north and east. Although by 476 Rome had been sacked twice, some of the theatres were rebuilt. The last definite record of a performance in Rome was in 533. Archaeological evidence suggests that the theatre did not survive the Lombard invasion of 568, after which state recognition and support of the theatre was abandoned. Theatre did, however, continue for a while in the Eastern Roman Empire, the capital of which was Constantinople. But by 692 the Trullan Council of the church passed a resolution forbidding all mimes, theatres, and other spectacles. Although the effectiveness of the decree has been questioned, historians, until recently, used it to signify the end of the ancient theatre.

The assumption now is that although official recognition and support of performances were withdrawn and theatres were not used, some remnants of at least the mime tradition were carried on throughout the Dark Ages. Frequently during this period, the writings of Christians suggest that performers were familiar figures. For instance, two popular sayings were "It is better to please God than the actors" and "It is better to feed paupers at your table than actors." Apart from the mime tradition, one Roman playwright, Terence, retained his reputation through the early Middle Ages, probably because of his literary style. From a period between the 7th and 10th centuries comes a dialogue, "Persona Delusoris," between Terence and another character. This was probably a prologue either for a performance of one of Terence's plays or a comedy based on his work.

The beginnings of medieval European drama were in the tropes—embellishments of the liturgy that were sung in churches on Christmas and Easter; appearing by 1000, they consisted of short dialogues in Latin, originally sung antiphonally (*i.e.*, alternately between priest and choir). Later they were elaborated with more words, extra characters, and more acting. It is a reasonable assumption that the religious drama of the Middle Ages came into being through the desire of the church to make the gospel story more real to an illiterate congregation.

A number of staging conventions that evolved in the church were to continue throughout the Middle Ages. The playing area had two components; small scenic structures, called *mansions, sedes, loci,* or *domi;* and a general acting area, called a *platea, playne,* or *place.* The methods of staging from these first liturgical dramas to the 16th-century interludes can be divided into six main types. The first involved the use of the church building as a theatre. In the beginning, for Easter tropes, a tomb was set up in the north aisle. As dialogue was added, the entire nave of the church was used by the actors, and within this space different localities were indicated by *mansions.* A few *mansions* housed elaborate and numerous properties, particularly those for the Last Supper. Some *mansions* had curtains so that characters or objects might be revealed at the right moment or concealed at the end of an episode. Sometimes the choir loft was used to represent heaven and the crypt to represent hell.

The six types of medieval staging

The second type of staging evolved by the 12th century, as drama began to outgrow the capacity of the church to contain it. As long as the action was confined to the central theme, it could be played in an arrangement of *mansions* down the length of the nave. But as the subject matter extended to include both Old and New Testament history, the action was transferred to a stage outside the west door of the church. In 1227 Pope Gregory decreed the removal of what had become a show from holy ground to the marketplace or an open field. Simultaneously, there had begun a corresponding transition in the language in which the plays were presented: from Latin to the vernacular. When drama was first taken outdoors, the crucifix was placed at one end, no doubt where it would have appeared above the altar in the church. *Mansions* were placed alongside each other, their positions usually determined by memories of performances within churches.

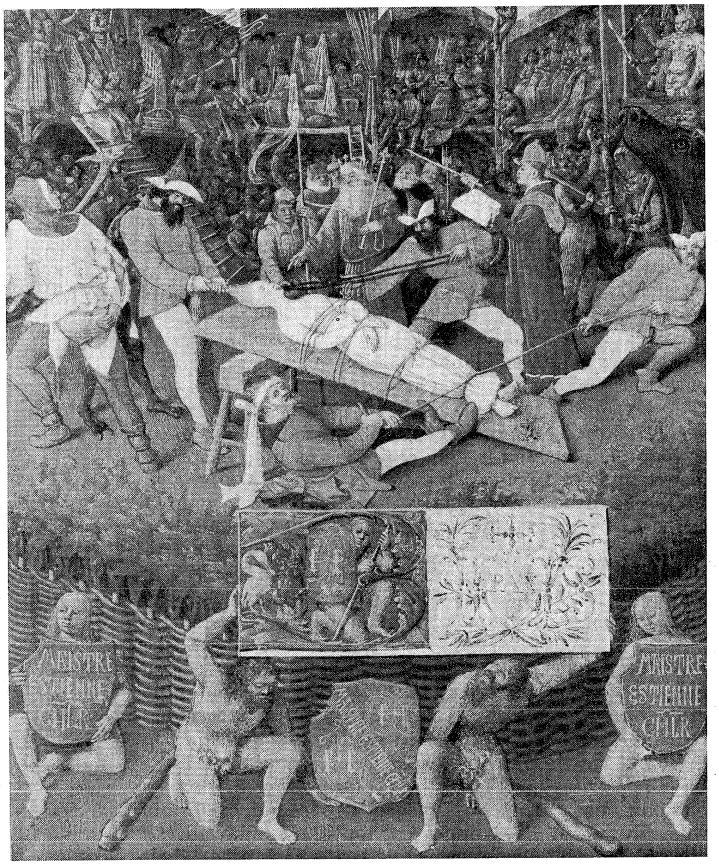
A third type of staging was the so-called stationary setting, found outside of England, in which the audience accepted three conventions. One was the symbolic representation of localities by the *mansions;* the second was the placing of the *mansions* near each other; and the third was the use for acting purposes of such actual ground as was enclosed by or in front of each *mansion.* The ground-level area was the area for the actors, and the *mansions* were placed either in a straight or a slightly curved line. The fixed settings were probably impressive, as all of the scenery was visible simultaneously. Because of their scope, many of the plays were divided into parts separated by intermissions ranging from one to 24 hours. During the intermissions, *mansions* were changed. Also, some *mansions* might represent more than one location; the identity of the *mansions* was announced by the director before each segment of a play. It is difficult to know exactly how many *mansions* were used; in a play at Lucerne, Switzerland, in 1583, for instance, 70 different locations were indicated, though only about 32 mansions were actually used.

Mansions of heaven and hell

The two *mansions* almost always present were heaven and hell, set at opposite ends of the playing area. The earthly scenes were set in the middle, and the two poles were supposed to represent man's dual nature and the choices that faced him. In the 15th and 16th centuries, heaven was usually raised above the level of the other *mansions.* Sometimes heaven had a series of intricate turning spheres, with concealed torches, so that it would look as though golden light was emanating from it. As hell was supposed to be terrifying, it was built to be the complete opposite of heaven; some portions of it, for instance, were even below stage level. Sometimes hell was made to look like a fortified town, which was effective when Jesus Christ forced open the gates to free the cap-

tive souls. The entrance to hell was usually shaped like a monster's head and was called a hell mouth, from which came fire, smoke, noises, and the cries of the damned.

By courtesy of the Musee Conde, Chantilly, France; photograph, Giraudon



Le Martyre de Sainte Apolline, a medieval religious drama using *mansions;* miniature by Jean Fouquet, *c.* 1460. In the Musée Condé, Chantilly, France.

The fourth type of staging was in the round. In France and England particularly, still-standing Roman playhouses were used for drama, and the *mansions* were probably placed in a circle. The fifth type involved the use of the pageant, which was like the *mansion* that was used in fixed settings except that it had wheels. The upper level was the main acting area, and the pageant was often supplemented by a movable open platform set along side it. "Processional" staging was particularly popular in Spain, and continued until 1765. The wagons, called *carros,* on which the scenery was mounted were pulled next to platforms that had been erected in every town. The last type of staging, and the one of which the least is known, was the curtained platform. Itinerant professional actors who performed interludes toward the end of the Middle Ages required only a curtain behind them for staging.

Machinery for special effects

Special effects, which were very popular, became so complex and numerous by the 15th century that many scenes were added to take advantage of the machinists' abilities. Many effects involved flying; fixed settings were often placed against buildings on the top of which were pulleys and windlasses. Additional flying machinery was also hidden inside the heaven *mansion.* Angels flew between heaven and earth, souls released from limbo ascended, and devils and fire-spitting monsters flapped around. The machinery became so complex that 17 people were needed simply to operate the hell scenes at a theatre at Mons, Belgium. Trapdoors were used for sudden appearances, disappearances, and substitutions of dummies for actors in scenes of violence. In a production involving Barnabas' burning at the stake, the dummy was filled with bones and entrails to give off a realistic smell.

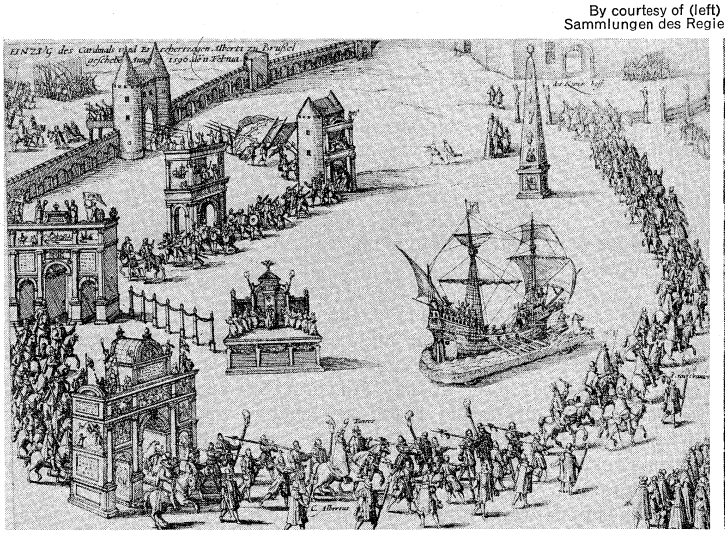
The morality plays, which first appeared in the 14th century, involved no scenery or complex properties. Although some were performed indoors, most were done outside on a stage that anticipated the English Elizabethan public theatre. A fixed facade was built at the rear of a large platform, and there were three openings at stage level that could be used to show interior scenes. A second level included similar openings, and a third level had a throne for the figure being honoured by the morality.

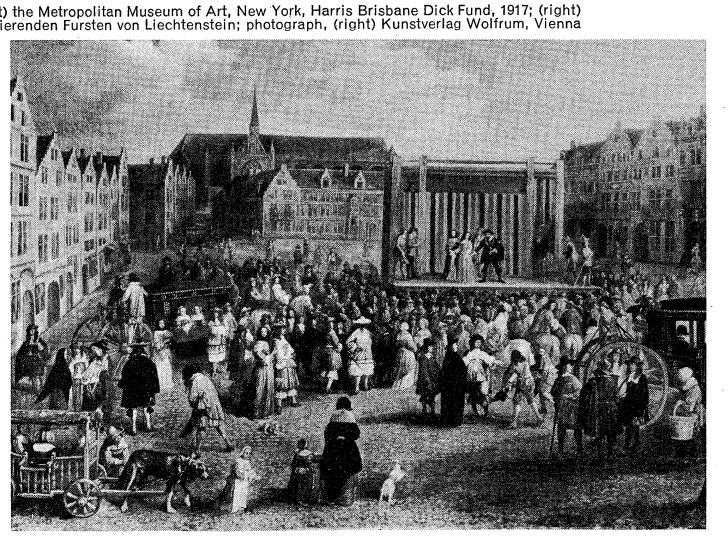
THE RENAISSANCE

Renaissance Italy. Just before 1500, Italian amateur actors were performing classical comedies on stages with no decoration except a row of curtained booths. By 1589, productions involving complex painted scenery and scene changes were being offered in Florence. And by 1650, Italy had developed staging practices that would dominate European theatre for the following 150 years.

The impact of the beginning of the Renaissance was felt differently in each country. The two forces moving Italian medieval theatre to change were the desire of the mercantile city-states to imitate the glories of ancient Rome and Greece (and in the process to escape liturgical constrictions) and the fact that Renaissance painters and architects, having discovered perspective, found that in

By courtesy of (left) the Metropolitan Museum of Art, New York, Harris Brisbane Dick Fund, 1917; (right) Sammlungen des Regierenden Fursten von Liechtenstein; photograph, (right) Kunstverlag Wolfrum, Vienna





Staging techniques originating in the Middle Ages.
(Left) Staged processional with pageant carts, or *carros,* honouring Cardinal Archduke Albert's entry into Brussels, 1596; etching, 16th-century German school. In the Metropolitan Museum of Art, New York. (Right) Use of the curtained platform stage shown in "Market Theatre, Brussels" by F. van der Maulen (1632–90). In Schloss Vaduz, Liechtenstein.

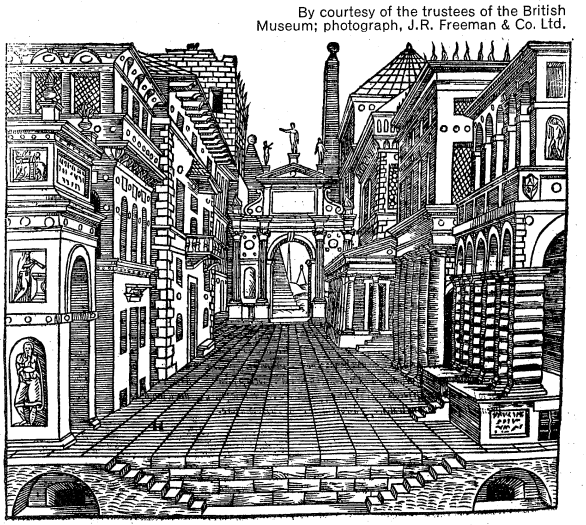
the theatre they could experiment beyond what had been possible in their studios. Toward the middle of the 15th century, scholars discovered the manuscripts of the 1st century AD of the Roman writer Vitruvius; one of these scholars, the architect and humanist Leon Battista Alberti, wrote *De re aedificatoria* (1452; first printed in 1485), which stimulated the desire to build something in the style of the classical stage. From 1537 to 1547 the architect Sebastiano Serlio published a work that concentrated entirely on the practical stage of the early 16th century.

In the beginning of the Renaissance, there were two distinct kinds of theatrical productions. The first was of the type presented by the Humanist Julius Laetus at the Roman Academy, a semisecret society he founded in the mid-15th century to revive classical ideals. In terms of staging, medieval-type *mansions* were placed next to each other to form a single large unit. There were, however, two elements not found previously: one was that the *mansions* were probably framed by decorative columns. This was the first movement toward the framework that would develop into the proscenium arch—the arch that encloses the curtain and frames the stage from the viewpoint of the audience. (The first permanent proscenium was built in the Teatro Farnese at Parma, Italy, in 1618–19, a temporary one having been constructed by Francesco Salviati 50 years earlier.) The second new element was that the *mansions*, by being linked, were treated as components of a general city street. In 1508 at Ferrara a background painted according to the rules of perspective was substituted for the mansions; the scene included houses, churches, towers, and gardens. In 1513 a production in Urbino had both a setting of a city and three-dimensional structures on the stage. There were towers on each side, in which the musicians played.

Origin of the proscenium arch

Serlio's treatise on the theatre had three especially significant items. The first was a plan for an auditorium and stage that assumed a rectangular hall, with spectators arranged in the same pattern as in the Roman cavea (*i.e.*, the tiered semicircular seating area of a Roman theatre), the difference being that the semicircle of the audience was cut short by the sidewalls. Second, his three types of stage designs, tragic, comic, and satiric, were the same as Vitruvius' classifications. Third, for the stage, he started with a Roman acting platform; but instead of the *scaenae frons*, he introduced a raked platform, slanted upward toward the rear, on which the perspective setting of a street was made up of painted canvases and three-dimensional houses. Since the perspective required that the houses rapidly diminish in size with distance, the actors were able to use only the front houses. Serlio used three types of scenes, all with the same basic floor plan: each required four sets of wings (*i.e.*, the pieces of scenery at the side of the stage), the first three angled and the fourth flat, and a perspective backdrop.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



"Tragic Scene" by Sebastiano Serlio, a background in perspective for players who performed in front of this scene. Woodcut from *Architettura*, English edition, 1611.

Also in the latter half of the 16th century, intermezzi, entertainments totally unrelated to the play, were inserted between the acts. These grew out of Neoclassic divertissements for masked balls and banquets. The special effects and spectacle galore were the principal appeal of intermezzi, which grew until they virtually absorbed the full attention of the audience. Now the need to change settings rapidly for the alternating segments of plays and intermezzi encouraged the development of new devices for shifting scenery. The first solution to scene shifting adopted for intermezzi was derived from discussions of *periaktoi* found in Vitruvius. Nicola Sabbatini's "Manual for Constructing Theatrical Scenes and Machines," published in 1638, listed three main methods of changing scenery: one used *periaktoi;* the second manoeuvred new wings around those already there; and the third pulled painted canvas around the wings to conceal the previously visible surfaces. In addition, the author explains how to change the flat wings near the back of the stage by sliding them in grooves or turning them like pages in a book. All of Sabbatini's devices indicate a considerable simplification of Serlio's wings, for painted details have replaced the three-dimensional ones.

Development of wings

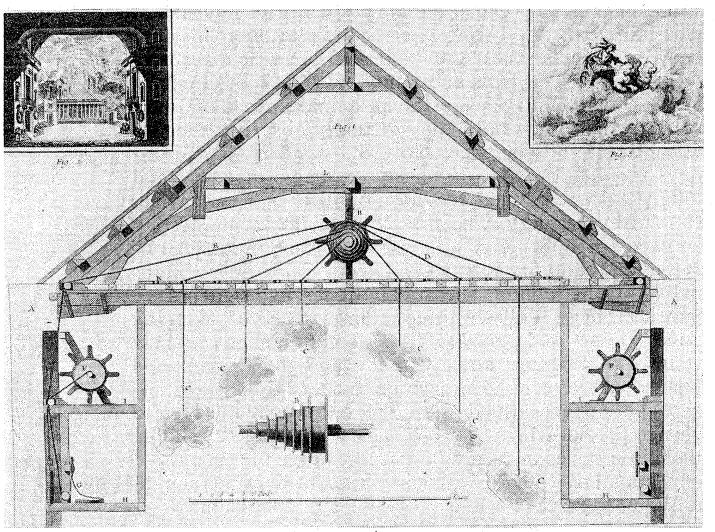
The final solution to scene changing required that flat wings replace the angled ones. The problem of transferring a perspective picture successfully to a series of flat wings was not solved until 1600. By 1650 angled wings were completely outmoded; at each wing position as many flats were set up, one behind the other, as there were settings. Scenery was changed by removing the visible wings to reveal the set behind. Grooves were made in the stage floor to support the flats and facilitate their movement. The background was painted on two flats, called shutters, which met at the centre of the stage; and cloths that could be rolled up were occasionally used.

The final step in scene shifting was introduced by Giacomo Torelli in 1641, when he perfected the chariot-and-pole system of scene changing. Slots were cut in the stage floor to support uprights. These poles, on which flats were mounted, were attached below the stage to chariots mounted on casters that ran in tracks parallel to the front of the stage. As the chariots rolled to the centre of the stage, they carried flats into view. An elaborate system of winches, ropes, and pulleys made it possible to change every part of a setting by turning a single winch. This invention, adopted by every European country except England and Holland, was the standard method of shifting scenery until the end of the 19th century.

The Italians were particularly taken with special effects and delighted in elaborating on medieval practices. Most of the special effects were associated with the intermezzi and involved the replacement of biblical characters of the Middle Ages by pagan deities. Flying machines with from one to 50 characters might be flying in any particular scene. According to Serlio, moving mechanical figures representing men, animals, and objects were cut out of pasteboard and drawn across a scene by invisible wires. In many productions clouds that engulfed the stage would hide the activity of lowering painted cloths and flats. The front curtain was used to conceal the scenery and increase amazement at the beginning of a performance. At first the curtain was dropped, but as this was hazardous, the roll curtain was soon adopted.

In spite of the interest taken in classical drama by the Italian aristocracy (with the overlay of exotica brought back by the merchant fleets), the movement was by no means a popular one, which explains why so few permanent theatres were built during the Renaissance. Audiences for the plays were comparatively small and could be accommodated comfortably in the large rooms of private palaces. One theatre of this time has survived and may be considered an expression of the architectural ideals of the period. It is the Teatro Olimpico at Vicenza, Italy, built from plans drawn by the influential architect Andrea Palladio and completed in 1584.

The theatre has a semi-elliptical auditorium that allowed a good view of the stage to all members of the audience and seated about 3,000 people. The lower floor of the auditorium was arranged as a series of curved



Stage grids and clouds used in Renaissance and Baroque theatre. Cross-section diagram from *Dictionnaire des Sciences* shows method of operation and insets illustrate the effect created.
By courtesy of the Harvard Theatre Collection

steps, encircled at the rear by a colonnade, the space between the columns forming boxes. Stairs led from the rear of the auditorium to the gallery. A semicircular acting space opened directly before the front row of seats, and beyond this forestage was the stage proper. The stage was set with the typical Renaissance scene; rows of houses forming three streets in forced perspective, the bare forestage serving as a public square.

Commedia dell'arte staging

Simultaneous with the theatre of the aristocracy was another that entertained the majority of the Italian population, the commedia dell'arte. The commedia became popular in the early 1500s and in the following 150 years swept over western Europe. It may have arisen from the Atellanae (*i.e.*, ancient popular farces played in Atella, a city near Rome). How it remained dormant for 1,000 years has never been adequately explained, particularly since the characters in the 16th century were the same as the earlier ones. There are no records of these masked, improvised farces being performed during the Middle Ages, although they probably were. At any rate, the persistence of the commedia is presumably due to the populace's love of the raw, earthy, satirical gang of scoundrels who made light of the court and their rarefied theatrics. Naples was a flourishing stand, where ridicule of the Spanish regime, which controlled the area, could be disguised by the vulgar doings of stock commedia characters. Each commedia troupe travelled constantly, either hiring large rooms, improvising stages outdoors, or performing on court stages. When perspective scenery and elaborate special effects were available, the troupes used them, but they were not a necessary ingredient of their productions (see also COMMEDIA DELL'ARTE).

Decades passed before the proscenium arch and decorated stage developed in Italy would influence the rest of Europe. In Renaissance Italy the existence of a small cultured aristocracy offered the ideal environment for the invention and growth of the new theatre. In England the Renaissance spirit discovered its characteristic medium in poetry rather than the visual arts; France, politically torn apart in the 16th century, had almost no cultural flowering; and Spain maintained its own national style.

The 16th and 17th centuries in France. In 16th-century France, most activity was associated with the Confrérie de la Passion, a Parisian organization set up for the performance of mystery cycles. In 1402 the company was granted permission to stage any mystery play, but by 1548 it had been forbidden to produce sacred mysteries, this satirical forum for the lower clergy having proved to be too much for the ecclesiastical authorities. They were granted, instead, complete control over secular drama, and they converted the Hôtel de Bourgogne into a theatre. The designs were all made up from sets of simulated houses painted on frames. What is unusual about these designs is the attempt to put so many houses representing different localities in the same stage area.

French *décor simultané* method of staging

In France, despite the number of critics seeking to hold to the classical "unities"—*i.e.*, the notion that a play should deal with only one action, set in one day, in one place—popular sentiment inspired plays involving many localities. Such a play could be presented by the use of curtains, by changing scenery, or by a third method, the *décor simultané*, which was utilized by Laurent and Mahelot, the designers for the Confrérie. Several localities were represented on the stage at the same time—each on a different portion of the stage—using angled mansions of wooden frames covered with canvas. When a play needed more *mansions* than could be fitted on stage at once, units were converted by removing the canvas coverings or by opening curtains to show an interior scene. Molière inherited a permanent town square set when he moved into the Petit-Bourbon in 1658.

The Elizabethan stage. During the early part of the 16th century, there were two distinct types of theatre in England. One was represented by small groups of professional actors who performed in halls, inns, or marketplaces. The location of a play was established by the words and gestures of the actors. As in the commedia dell'arte, these localities had little significance. The second type of theatre, found in the London area, was made up of amateurs, usually university students, performing for the royal court and assorted gentry. The audience and the actors were educated, acquainted with the classics, and knowledgeable about theatre in other countries, particularly France. The stage was probably set with buildings made of laths, covered with painted canvas, with cloud borders masking the upper part of the acting area.

The significant achievement of the Elizabethan stage was connected with the theatres of the professional acting groups, not the court theatre. During the second half of the 16th century, as they became successful, the troupes no longer needed to remain itinerant. In 1576 the first permanent public theatre, called simply the Theatre, was erected by the actor James Burbage. The building boom continued until the end of the century; the Globe, where Shakespeare's plays were first performed, was built in 1599 with lumber from the demolished Theatre.

The stage was a platform, as large as 40 feet square (more than 12 metres on each side), sticking out into the middle of the yard so that the spectators surrounded most of it. It was raised four to six feet (about one to two metres) and was sheltered by a roof, called "the shadow" or "the heavens." In most theatres the stage roof, supported by two pillars set midway at the sides of the stage, concealed an upper area from which objects could be raised or lowered. At the rear of the stage was a multilevelled facade with two large doors at stage level. There was also a space for "discoveries," probably between the doors. Historians disagree as to its function: either it was used only to locate scenes or scenes were actually performed there. Some scenes took place in a playing area on the second level of the facade, but again historians disagree as to which scenes they were.

Properties were occasionally carried onto the platform stage, but from extant lists it is obvious that they were few in number. Some properties were so cumbersome that they remained on stage throughout a performance. Smaller properties were probably revealed in the discovery space, and servants carried some properties on and off. It appears that the audience was not concerned by the scenic inconsistencies.

All of the theatre buildings were either round, square, or octagonal, with thatched roofs covering the sides. Spectators, depending on how much money they had, could either stand in the yard, which may have sloped toward the stage; sit on benches in the galleries that went around the greater part of the walls; sit in one of the private boxes; or sit on a stool on the stage proper.

The importance of this type of theatre was its flexibility. In some ways it was similar to earlier attempts to reconstruct the *scaenae frons* of the Romans; it had the facade

and the entrance doors. The Elizabethan theatre differed in that it had a platform that made movement possible in all directions instead of the long, narrow stage. The actors now had an inner stage to use, and even vertical action was possible by utilizing the upper stage level.

The Elizabethan playhouse, in its dynamic use of stage space, its predominantly formal approach to production, and its reflection of the new middle class ideology, represented a decided break with the classicism of the court-dominated spectacles; it broke the ground for the coming romanticism and naturalism. Shakespeare has always been considered the pivot for radical changes in theatrical thinking.

Spain's Golden Age. Religious drama developed during the Middle Ages only in northeastern Spain because the Moors occupied the remainder of the peninsula. During the 16th century, as Spain became the most powerful country in Europe, it started to develop a sophisticated theatre. Following an interest in classical drama and the introduction of printing came Juan del Encina, the founder of modern Spanish drama. Although the origin of professional players is obscure, actors were being paid as early as 1454. The popularity of the theatre mushroomed in the 1570s, and among the playwrights of this era were such masters as Lope de Vega, Cervantes, and Calderón de la Barca.

Staging based on the *corrales*

During this "Golden Age" in Spain, the theatre assumed a form more flexible than that of the Italian or French stages. The model was that of the *corrales*—courtyards enclosed by the backs of several houses—in which the earliest troupes had performed. The staging arrangements were almost identical to those in contemporary London. The stage itself was a raised platform, without a front curtain or a proscenium arch but with a permanent facade at the back. Stages were about 28 feet (nine metres) long and 23 to 30 feet (seven to nine metres) deep. The open platform was usually backed by a facade of two levels, with pillars dividing the lower level into three openings. The second level, basically a gallery, usually represented towers, city walls, or hills.

Spanish staging conventions, like those of the Elizabethan theatre, tended to be simple. To denote a change of location, an actor merely exited and re-entered. Three kinds of scenic background were utilized: the facade; the curtains concealing the facade, which were used when the location was not particularly important; and medieval-type *mansions,* which were sometimes erected on the main stage. As spectacle increased after 1650, painted flats with doors and windows were set into the facade in place of curtains. The *corrales* themselves, enclosing a square or rectangular courtyard, were unroofed until the 18th century, and when roofed a row of windows was added under the eaves. Benches were placed on the ground level and balconies set in the containing walls for wealthier patrons. A special boxlike gallery, called the *cazuela,* the "stewpot," was assigned to women spectators. And above the *cazuela* were galleries for members of the city government, the church, and the aristocracy.

THE BAROQUE

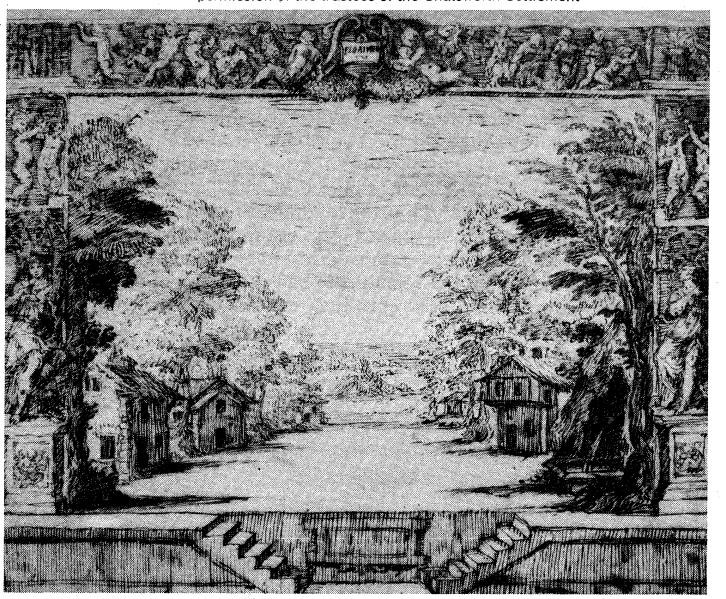
The combination of two innovations, the first opera in 1597 and the formulation of the laws of perspective in the 15th century, provided the foundation for the Baroque theatre, a period that would extend until the 19th century. During this era all countries were brought into the same orbit, although Italy remained the primary inspiration. The central emphasis of staging was on scenery for entertainment that differed from the classically inspired 16th-century dramas; now there were intermezzi, ballet, masques, and opera. New means of presenting spectacular visual effects were invented, and these encouraged the installation of more machinery in the theatre building. The result was that spectacle dominated all other aspects of production. The stage floor, which previously had extended only a few yards back from the proscenium arch, was now stretched farther from the auditorium.

England: the 17th and 18th centuries. The court masques served to introduce Italian staging to England and started the trend toward the proscenium-arch theatre that culminated in the disappearance of the facade stage by 1660. The masques were allegories designed to honour a particular person or occasion by comparing them favourably with mythological characters or situations.

Introduction of court masques in England

Inigo Jones, the foremost English architect of his time, was responsible for the productions of the masques and other entertainments at the English court from 1605 to 1640. He had visited Italy between 1596 and 1604 and was the individual most responsible for acceptance of Italian stage design in England. From his sketches it is known that Jones went through several phases in his designs, starting in 1604 with the *décor simultané,* mentioned above, in which different localities were represented on different portions of the stage. In 1605 he introduced simple perspective settings—two painted representations of houses with a painted back shutter. The same year he experimented with *periaktoi,* creating a globe, with no visible axle, that revolved to reveal eight dancers sitting inside.

Devonshire Collection, Chatsworth. Reproduced by permission of the trustees of the Chatsworth Settlement



Stage designed by Inigo Jones in the Italian manner for a production of *Florimène,* 1635. In the Devonshire Collection, Chatsworth, England.

His masques all had painted proscenium arches, into which he set a falling curtain. By 1635 Jones designed a setting that utilized four angled wings, like Serlio's, and four shutters at the rear, three of which could be drawn to the sides in two parts. Jones's design for a masque in 1640 is considered the first design of the Baroque theatre. Four sets of side wings were placed on each side of the stage, each wing consisting of two flats. There were four shutters at the rear, with each dividing into two parts for easy removal. Each wing had either a header (*i.e.,* a horizontal unit that joins two upright wings, to form a flat arch) or a sky border (a horizontal piece of scenery, designed to look like the sky, which masks the space above the set). To give the illusion of distance, the side wings were made in exaggerated perspective, with each succeeding wing in significantly smaller scale than the one preceding. The disadvantage of using smaller sized flats toward the back of the stage was that if the actors stood too close to them the illusion would be destroyed. The actors were therefore restricted to the front of the stage. In successive decades, attempts were made to give the stage area greater depth by multiplying the number of flats on each side. In sum, with the exception of the chariot-and-pole system of scene changing, Jones introduced all of the major Italian developments into England.

Jones's first Baroque design

After Inigo Jones, English scenery practices and stage conventions were similar to those of Italy. Sets were changed by sliding them in grooves in the stage floor and overhead. Since the curtain was raised after the prologue

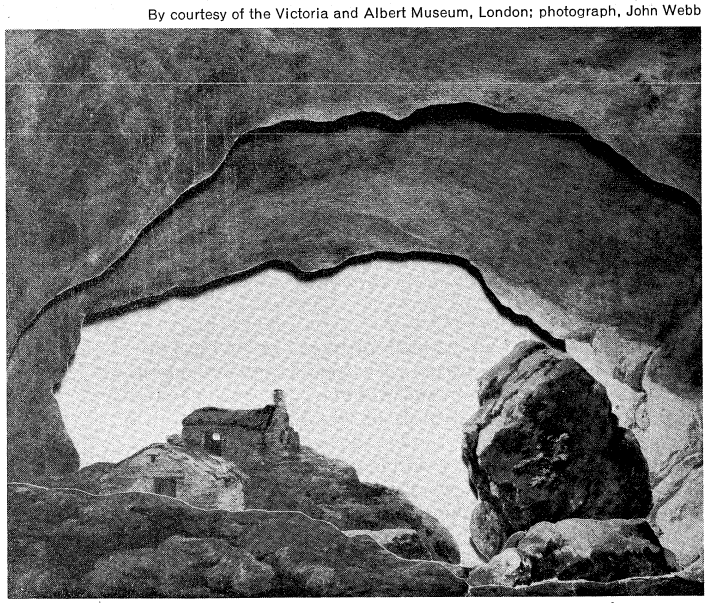
and remained up throughout the performance, all scene shifting was in view of the audience. It was not until 1750 that an "act drop" was used; previously, even intermezzi were performed in front of a full stage setting.

The development of spectacle was slow in England because members of the audience sat on the stage. Between 1700 and 1762 benches extended upstage from the proscenium and amphitheatrical seating was sometimes built across the back of the stage. Finally, in 1762, the distinguished actor David Garrick removed the spectators from the stage. As interest in spectacle increased, the scene painter became more important, and by the late 18th century each theatre had two or more permanent scene painters. In 1750 the actor-manager John Rich started to import from the Continent designers as eminent as the Italians Giovanni Servandoni and Giovanni Cipriani.

Loutherbourg's contributions

The best known designer around the end of the 18th century was Philip James de Loutherbourg, a painter; from 1771 he worked for Garrick as scenic designer at the Drury Lane Theatre in London and is credited with changing the orientation of design from the architectural to the landscape era, thus marking the end of the Baroque in England. He was one of the pioneers of the cut cloth, a double back cloth in which there was an opening in the one nearer the audience that revealed a vista painted on the back one. He also utilized transparent scenery; in one production he cut the moon out of the canvas back cloth, replaced it with gauze, and lit it from behind. The importance that Loutherbourg's landscape painting of the back cloth assumed is shown by the fact that the Drury Lane pantomime of 1779 was specially written for the scenery he had designed while on a trip to Derbyshire. His reproduction of actual places in England started a vogue for "local colour."

By courtesy of the Victoria and Albert Museum, London; photograph, John Webb



Philip James de Loutherbourg's model set for the pantomime "The Wonders of Derbyshire, or Harlequin on the Peak." In the Victoria and Albert Museum, London.

Loutherbourg was intrigued with miniaturization and invented a miniature theatre, the Eidophusikon, which consisted of nothing but scenic effects. He used miniature figures at the rear of the stage, which was six feet wide and eight feet deep, to represent soldiers in battle or marching. He was also a specialist in sound effects, such as waves, rain, hail, and gunfire. Loutherbourg's single most important contribution was that he achieved a unity of design because he directed both the scenery and the lighting and effects of a single production. At the beginning of the Baroque period in England, the three important theatres were the Drury Lane, the Lincoln's Inn Fields, and the Dorset Garden. Although each theatre was different from the others, they shared many significant details. The auditorium was divided into a raked pit, boxes, and galleries. There were two or three galleries, with the lowest partitioned into boxes and the upper one open. The theatres were small, each holding about 600 people. The stage included both a proscenium arch and an open platform in front of the proscenium. The stage was about 35 feet deep (11 metres), divided into equal parts by the arch. At either side of the apron were several doors, each with a balcony above it. These doors were the entrances for all the actors since the action took place on the apron and would continue to do so throughout the Baroque era. The floor of the stage was raked from the front of the apron to the back wall. There were grooves behind the proscenium for wings and shutters. Throughout the 18th century the theatre buildings grew larger, and the only other change was about 1700, when Christopher Rich, John's father and the manager of Drury Lane, removed one set of proscenium doors and shortened the apron.

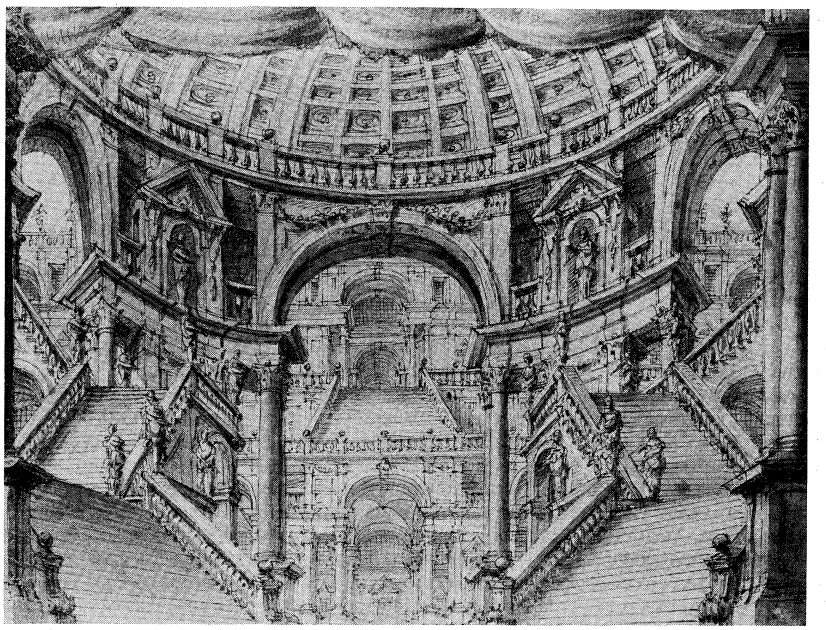
Angled perspective designs

Italy in the 18th century. Significant changes in scenic design were made by Italy's Bibiena family, of whom the best known members were Ferdinando, Francesco, Giuseppe, Antonio, and Carlo. Around 1703, at Bologna, Ferdinando introduced angle perspective. Previously, stage design was based on one-point perspective using a single vanishing point, in which all lines appear to recede with distance toward one point at the centre of the background (as parallel railroad tracks might appear to meet in the distance to someone standing amid them). Bibiena, however, used perspective with two vanishing points; if this technique were used to render a large, flat building as seen from one corner of it, for example, the base and roof lines of one side extended into the distance would appear to meet at some point off to the right and those of the other side at another point off to the left. Furthermore, by locating these points quite low, Bibiena gave the structures the effect of immense size. Buildings, walls, or courtyards were placed in the centre of the set and vistas at the sides, and the scale of settings, which had previously been designed to make the scenery an extension of the auditorium, was altered. The front wings were painted as though they were only the lower part of a building. The result of these innovations was that Bibiena's sets seemed so large that they created a feeling of fantasy. Despite its apparent size, though, an angled-perspective setting required less stage space than one with a central alley. Many of the settings divided the stage into a foreground, for the actors, and a background, for distant objects. A drop, designed to resemble a series of arches or columns, often marked the rear of the acting area.

Another designer who developed the angled scene independently of Bibiena was Filippo Juvara, a major Italian architect of the early 18th century, who began his work, as did many artists of the period, with the decoration of court entertainments. Some of his most intriguing scenic designs, in terms of architectural fantasy, were executed for a marionette theatre owned by his patron. These designs are particularly helpful because Juvara usually included plans of the stage settings showing exactly how the wings were placed. Juvara's sets were basically curvilinear, leading the audience's eye to the foreground instead of the sides. In several sketches, he designed a permanent set with a large archway opening up to a series of vistas, which varied from a landscape to a perspective corridor. His designs foreshadowed the interest in landscape that was not to find its complete development until almost a century later.

Late 18th-century developments

In the second half of the 18th century, several new directions in thinking led to stylistic changes in scenery. One was the increased interest in history, spurred by the rediscovery of Pompeii in 1748. Scenery of classical ruins with vines rambling between the cracks became popular. Plays dealing with folk literature were produced with Gothic architectural settings. The most important new direction was that scenic designers introduced "mood," they started to emphasize light and shadow to create an atmosphere. The best known artist of this period was the Italian engraver Giambattista Piranesi, who executed more than 1,000 engravings of Roman ruins and prisons. He did not particularly apply himself to the theatre, yet



Drawing for an interior setting by Giuseppe Bibiena. In the Victoria and Albert Museum, London.
By courtesy of the Victoria and Albert Museum, London

his designs were inspired by contemporary stage settings and were in turn themselves an inspiration for other designers. One admirer of Piranesi, for instance, brought to Poland a characteristically native modification of the Italian style.

Baroque theatre in France and Spain. Although Italian-style scenery was introduced to the French court before, it was not popular until after 1640. The first theatre in France with a permanent proscenium arch and a stage designed for flat wings was constructed in 1641 for Cardinal de Richelieu. In 1645 an Italian designer, Giacomo Torelli, popularly called "the great sorcerer," was imported by Richelieu's successor, Jules Cardinal Mazarin, to design for the new theatre, the Palais-Royal. Torelli's first assignment was an opera, for which he installed his chariot-and-pole system of scene shifting at the Petit-Bourbon. In 1646–47 Torelli remodelled the Palais-Royal to accommodate his invention. Pierre Corneille, the founder of French classical tragedy, was commissioned to write *Andromède* for the remodelled theatre. Although the play progressed through spoken episodes, each act provided an excuse for Torelli to introduce elaborate machinery, including the revolving stage. Until Torelli, the changing of scenery had marked a structural break in the dramatic presentation. Torelli put aside all previous methods, which were so distracting to the audience, and introduced set changes within, rather than at the ends of, the scenes. The audience was mesmerized by the scenery mysteriously changing while the action of an opera or ballet proceeded without interruption.

For the wedding of Louis XIV, in 1660, Gaspare Vigarani went to France from Italy to build the Salle des Machines, the largest theatre in Europe. It was 226 feet (69 metres) long, only 94 feet of which was occupied by the auditorium. Its stage, 132 feet (40 metres) deep, had a proscenium arch only 32 feet (ten metres) wide. One of Vigarani's machines, 60 feet deep itself, was used to fly the entire royal family and their attendants on stage for the finale. Vigarani was succeeded by his son, who was in turn followed by the Jean Berains, father and son. The Berains established a distinctly French style of design, emphasizing heavy lines, reverse curves, and encrusted ornamentation. Berain usually designed only one set for each act but used many machines and special effects. This combination of static scenery and dynamic machinery was to remain in vogue in French opera through the 18th

Designs of Jean Berain

By courtesy of the Metropolitan Museum of Art, New York, Elisha Whittelsey Fund, 1951



Set design by Giacoma Torelli for Act II of Corneille's *Andromède,* as performed in the Palais-Royal theatre, Paris, 1650; engraving by F. Chauveau. In the Metropolitan Museum of Art, New York.

century. After 1682 the influence of the court on public theatre declined. After 1690 the King became so puritanical that plays were no longer encouraged.

In contrast to court opulence was the Comédie-Française, for which a theatre was built in 1689 from a converted tennis court. The stage was equipped for flat wings and shutters, but since scene changes were few, the machinery was minimal. The typical background for tragedies was the *palais à volonté* (literally "palace to order"), a neutral setting without particularized details. For comedy the typical scene was *chambre à quatre portes* ("room with four doors"), an informal interior. By 1700 Paris had two types of theatres—the Opéra, with its Baroque scenery and machines, and the Comédie-Française, which did not rely on spectacle.

The first true spectacle in a public theatre was the 1755 production of Voltaire's *Orphan of China*, with its supposedly accurate Chinese scenery and costumes. The *comédie-italienne*, permanently performing in Paris, introduced local colour and increased the use of spectacle. Giovanni Nicolò Servandoni, originally a Florentine who lived in Paris from 1724 to 1746, translated Italian styles to French taste. His use of perspective gave the illusion of space without falling back on the obvious geometry that characterized the High Baroque. He took over as director of the Salle de Machines, where his work in its turn influenced the designers of Italy and Germany. Throughout the 1700s, spectacle was utilized at the Opéra. Its best known designers included the painter François Boucher. And although many famous painters designed settings, there is still no comparison with the works of the Italians of this period. The financial problems of the theatres did much to prevent a forward thrust in French design. Ironically, the most sumptuous spectacle of the latter part of the 18th century occurred in France in 1794, after the Revolution: the National Convention decreed "The Festival of the Supreme Being," designed by the painter Jacques-Louis David and conducted by the revolutionary government leader Robespierre—Neoclassicism in the service of the Republic.

Until 1763 the theatre buildings remained unchanged. After that date they were built in the Italian style, with ovoid auditoriums and enlarged stages. By the 1780s, when the standing pit was abolished, all of the spectators were finally seated, off the stage. The seated pit was not to become accepted in France until the 19th century.

In Spain, during this period, theatre started the decline from which it has not yet recovered. It was not until 1598 that theatre performances were presented regularly at court. Although Italian-style scenery was used occasionally, it was not common until brought from Florence in 1626, by Cosimo Lotti, who staged many outdoor productions on the palace grounds. For one, he built a floating stage on a lake, and the special effects included a shipwreck, a water chariot drawn by dolphins, and the destruction of Circé's palace. This production was lit by 3,000 lanterns, and the spectators watched from gondolas. In 1640 Lotti built a permanent theatre, the Coliseo, which probably had the first proscenium arch in Spain. The next decade saw a decline in both court and public theatres. By 1650 the Coliseo was reopened, but its popularity had diminished by 1700. By the late 1600s Spain had lost most of its economic and political power, and its theatre through most of the Baroque was a watered-down version of Italian concepts.

Developments in northern Europe. One country outside the Italian influence during this period exhibited an interesting theatre plan. In 1638 Jacob van Campen, an architect, designed a theatre in Amsterdam that had no counterpart elsewhere in Europe. The auditorium was elliptical, with two tiers of boxes on one side. The opposite side had a stage facade with open balconies over the sides. The facade itself consisted primarily of pillars with cornices above them, and flat painted panels were set between the pillars. The panels themselves represented different localities, rather than a unified setting. The whole concept of these panels was probably directly derived from the simultaneous stage of the Middle Ages.

The constantly shifting politics of northern Europe, and particularly of Germany, in the 16th and 17th centuries was the primary cause for the late development of professional theatre in this region. Theatre in Germany owes its impetus to the English troupes that started touring the country in the early 17th century. At first, the English performed in courts, and by 1650 they were travelling across the country regularly. To attract the non-English-speaking audience, the troupes gradually inserted German phrases, speeches, and then whole scenes. As early as 1626, German actors were joining the troupes, and by 1680 English actors had completely disappeared. Because Germany was politically divided and there were no large cities, the troupes had to travel constantly to find new audiences. To attract the disparate spectators, a company might have as many as 100 plays in its repertory. These conditions combined to make it impossible to rehearse and mount plays with any great thought.

First Viennese opera house

Between 1650 and 1775 the aristocracy ignored public theatre in favour of opera. The first Viennese opera house was built in 1652 by Giovanni Burnacini and is similar to the courtyards of Italian Renaissance palaces in the two levels of its auditorium, framed by a double arcade of rounded arches. Italian influences were bound to be strong, since Italians played an important part in every aspect of theatre at the imperial court until 1770. The first performance at the Viennese opera house was of an Italian opera.

In the public theatre, the work of the manager Johann Friedrich Schönemann led to the establishment of the Hamburg National Theatre, the first noncommercial public theatre. This marked the beginning of a movement that gained force during the next decade and can be found even today in East German theatre. The prevailing attitude was that theatre is a cultural institution that, like museums, should be available to all citizens.

This movement toward permanence greatly influenced the evolution of staging in the 18th century. As audiences grew larger, and as cities developed, a troupe was able to reduce the number of plays in a repertory, and this in turn led to better productions. And as the theatre building grew in size and substance, the chariot-and-pole system of scene shifting was introduced. Interest in historical accuracy of sets and costumes was stimulated by the chivalric plays popular in the 1770s. During the next two decades dramatists were writing plays that relied on complex set pieces, including bridges and walls. Doors and windows were set up between the wings, marking the first development toward the box set—a set representing three walls of a room, the fourth being the proscenium.

Early Russian staging. Russian theatre can be traced back to the pagan rituals of the ancient Slavs, the later Christian festivals, and, in the Middle Ages, the mixed rituals of these two influences. As early as the 10th century there were mummers, called *skomorokhi*, probably itinerant comedians who performed in small towns and villages. Suffering the same fate as the actors of western Europe during the Dark and early Middle ages, the mummers were the victims of measures taken by the princes as well as the church. Their specialties were trained-animal acts, usually with bears, and puppet shows. The Greek Orthodox Church introduced some morality plays in Russia, but they were limited in scope and number. Since the mass was celebrated in the vernacular, the church did not need the aid of morality plays for the dissemination of its message. At the beginning of the 17th century, the only theatre using a literary text and a trained company was the school drama brought from Poland to the Ukraine.

The Tsar's theatre

In 1672 the first public performance of a play was offered in Moscow. A special theatre was built in the Tsar's residence; there seems to have been no scenery outside two fir trees on either side of the stage, although there was a sliding curtain. The Tsar (Alexis), who was so excited that he sat in the theatre for ten hours on the day of the first performance, paid for the founding of a theatrical school. The problem with the theatre during the next 80 years was precisely that it had been founded on the Tsar's order: it was attached to the court and subsidized by the aristocracy (who chose German direc-

tors); it assumed political overtones, and aimed at pleasing only the monarchy.

The beginning of change came in 1730, when an Italian group brought the commedia dell'arte, elaborate sets, and machinery to Moscow. During the next ten years, Italian opera and ballet were also introduced to Russia. Despite the success of foreign companies, the significant event was the birth of the national theatre in 1749. This enterprise was open to the public, and, although it was popular, lack of funds made it a weak rival to court theatre. Its impact lay in the fact that it introduced the concept of theatre to a large audience.

By the end of the 18th century, companies were touring the major cities, and privately owned theatres were being opened. In general, only the aristocracy patronized the theatre, and nobles established their own serf troupes. One prince, who owned 21,000 serfs, established his own ballet, opera, and dramatic companies. Owners occasionally rented their troupes for public performances, until the abolition of serfdom in 1861. Little is known of the staging of their productions beyond the fact that it was imitative of French and Italian theatre.

THE 19TH CENTURY

Theatre in France after the Revolution. Under Napoleon, French theatre was little different from that of the 1780s, specializing in Neoclassical drama. Popular drama, as performed by what were known as "boulevard theatres," introduced melodrama, a form that was to dominate theatre in the 19th century. Melodrama, in turn, by popularizing departures from Neoclassicism and capturing the interest of large audiences, paved the way for Romantic drama.

The dramatic debut of Romanticism is dated at 1830, when public pressure forced the Comédie-Française to produce Victor Hugo's *Hernani*. After a spirited opening at which Hugo's Bohemian claque overwhelmed the staid regular theatregoers, Romanticism was victorious and ruled the Parisian stage for 50 years. The grandiose bombast of Romanticism did not overturn the Baroque, it merely diluted it; the formal artificial structure was broken up into sentimental, melodramatic episodes depicting the distraught hero buffeted by an unfeeling world and the awesome elements. The melodramas introduced natural disasters that were significant to the plot, so that emphasis could be placed on special effects and spectacle. Dramatists also deliberately included exotic locales or examples of local colour, so that a variety of historical periods and fantastic sets would hold the attention of the audience. Throughout the 19th century, architectural perspective was replaced by neo-Gothic sentimentalization of nature. Painted Romantic scenery, in the style of Loutherbourg, was the rage in France.

Staging of romantic drama

The two important designers of this period were Louis-Jacques-Mande Daguerre, who was also the inventor of the daguerrotype, an early photographic technique, and Pierre-Luc-Charles Ciceri, the most important designer of this period. The panorama, a major scenic innovation, was invented in 1787 and first used on the London stage in 1792. The panorama was set up in a circular building in which the audience, sitting on a central platform, was totally surrounded by a continuous painting. Daguerre started his career as one of the first panorama painters. He went on to invent the diorama, in which the audience sat on a platform that revolved to show paintings on proscenium-like stages. Although the scenery remained stationary, Daguerre created the illusion of constant change by controlling the light on the semitransparent sets. The panorama became more popular than the diorama because it did not depend on the ability to alter stage lighting. Its shape, though, was altered to resemble the diorama, as it could not surround the audience in a standard theatre. The next development in spectacle was the moving panorama, in which a continuous scene was painted on a long cloth, hung from an overhead track, and attached at both ends to spools. When the spools were turned, the cloth moved across the stage so that the actors with their carriages, and other props could move from one location to another without changing wings and drops. The sky borders were dispensed with, and flats of architectural units or natural objects forming an arch were placed at the front of the stage, through which was seen a distant view painted on a curved panorama stretching across the back and sides of the stage.

Ciceri's importance arose from his abilities to depict local colour, ruins, and historical backgrounds. He founded the first scenic studio in Paris independent of a theatre, with specialists in various types of design. After the opening of the new Opéra in 1822, interest in spectacle was so great that promptbooks were published, describing scenery and special effects and how they could be adapted for theatres with less equipment.

Ciceri's contributions to staging

By the end of the 19th century, the process of scenic design and construction had become standardized. The director gave the requirements to the scenic designer, who made cardboard models. The scenery was constructed by the theatre's carpenters and then sent to a scenic studio for painting.

A number of changes were made in theatre architecture during this period. Theatre interiors, still decorated with Baroque ornamentation, were transformed from intimate playhouses for the aristocracy into huge auditoriums. Most of the private seats, such as boxes, were done away with in favour of benches. This change was in keeping with the egalitarian spirit of France after the Revolution. The larger theatre also made practicable the use of the chariot-and-pole method of scene shifting.

A significant factor was that whereas the Baroque scenic artist wanted his scenes to shift in front of the audience, the 19th-century designer used the curtain between acts. The setting was now seen as an image of reality to be revealed by the raising of a curtain, not a series of dazzling transformations. This spirit of treating the set as a picture became even stronger with the advent of the Realist school later in the century.

Scenic practices in Italy in the 19th century followed the same paths as in France. Although Italy continued to be a leader in opera, its influence in other countries declined, as did the demand for Italian scenic artists.

German Romanticism and Naturalism. The 19th century in Germany was a study in contrasts. The beginning decades saw the rise of Romanticism, which, 50 years later, was still strong, primarily in the figure of the composer Richard Wagner. The century's middle decades, of political and economic disillusionment before the unification of Germany, were conducive to the emerging Naturalist school, the philosophy of which was first personified by the Meiningen troupe, organized by the Duke of Saxe-Meiningen.

By the middle 1820s, after the defeat of Napoleon, the political turbulence in Germany led to municipal control over the theatre and strict censorship. Repertories were composed of "safe" classics and insipid new plays, resulting in competent but uninspired theatre. This competence was reflected in the staging. One of the few important designers of this period was Karl Friedrich Schinkel, who had been trained in both Italy and Germany. He introduced the diorama in Berlin in 1827, and it encouraged the extended use of moving and stationary panoramas and dioramas.

One true innovator during the first half of the 19th century was Ludwig Tieck, who advocated realistic acting on a platform stage. With the help of an architect, he tried to reconstruct an Elizabethan public stage. He also championed the open stage in the belief that pictorial realism destroys the true illusion of the theatre. Invited by William IV of Prussia to stage *Antigone* at the court theater in Potsdam in 1841, Tieck extended the apron in a semicircle over the orchestra pit and built a skene as the only background for the drama. In 1843 he adopted Elizabethan conventions to the proscenium theatre for a production of *A Midsummer Night's Dream*. Using the front part of the stage as a large open space, he built a unit in the rear consisting of two stairs leading to an acting area eight feet above stage level. The stairs framed an inner stage below the platform. He then hung tapestries at right angles to the proscenium, thus masking the sides of the stage. Although Tieck was universally respected, it was

Innovations of Ludwig Tieck

not until the 1870s that widespread support was given to his innovations in staging. The works of Wagner and Saxe-Meiningen were responsible for this change in the public's attitude.

Wagner's works were the dramatic culmination of Romanticism and the contribution it was destined to make to modern theatre practices. German Romanticism was in great part a protest movement against the dominance of French Neoclassicism. Instead of structuring dramatic action according to fixed patterns of logical progression, the Romantics wanted dramatic structures born of human experience. This stress on what was to be called "organic form" was a protest against the received tradition of dramatic theory and staging practices. German Romanticism, also known as *Sturm und Drang* ("storm and stress"), which is generally attributed to the influence of the young Goethe at the end of the 18th century, turned to a revival of the Gothic style of the Middle Ages to escape Neoclassicism. The new middle class audiences identified with the lonely soul against the world—a sentimental world of windswept mountain crags and gloomy, moldering castles. The Baroque backdrop that enclosed the retreating wings of the formal stage was expanded to a large painting by a landscape artist.

Romanticism later broke into two camps. The first of these, called "historical Romanticism," held that history is continuous, and once its import has been grasped, the present can be recognized to be as "historical" as anything that occurred in the past. The second Romantic group, with which Wagner was associated, was concerned only with the use of history to reach absolute truths. Wagner believed that the study of history ultimately leads to prehistory and thus to transhistorical mythology, the realm of absolute truths. What is particularly interesting is that the first type, historical Romanticism, eventually found its home in the theatre in the Realist school. Wagner's myth-Romanticism faded by the late 1800s, although their personal contributions to staging exerted their influence well into the 20th century.

Wagner wanted to use myth to reunite modern man with the passion from which rationalism, the industrial process, and capitalism separated him. For him, the theatrical manifestation of the myth was the music drama, and he hoped to combine music, acting, stage space, design, and lighting to establish the primeval mood of myth. To house his music drama, Wagner designed a new opera house in Bayreuth, which opened in 1876. Although he did not invent it, Wagner established the practice of darkening the house lights so that the attention of the audience would be rivetted to the stage. He had the orchestra lowered into a pit, so that it became the hidden source of an enveloping sound. He did away with the box seats, from which the wealthier theatregoers had watched each other instead of the stage for hundreds of years.

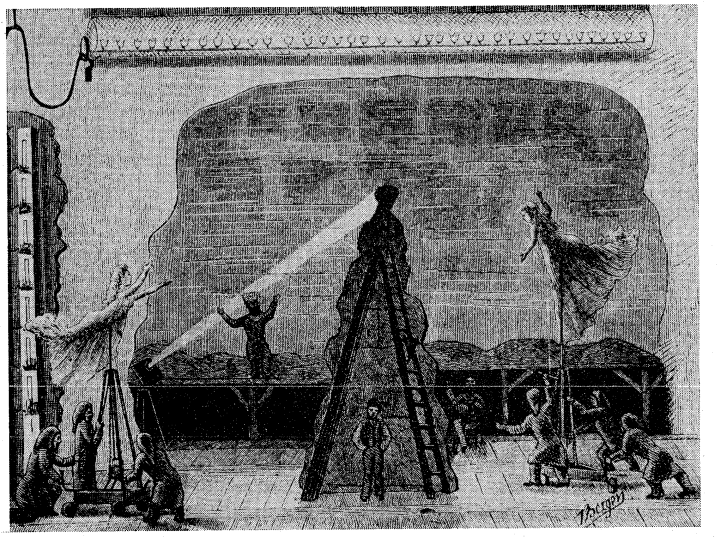
The stage itself was raked upward toward the rear, and the scenery was shifted by using the chariot-and-pole system. He introduced a system of steam vents to make a steam curtain to hide scene changes. In sum, he saw the theatre as a single instrument of which every part would function to move the audience from the commonplace to the mystical-mythical.

The Meiningen troupe

At the time Wagner was introducing his music drama, George II, duke of Saxe-Meiningen, began to take an interest in the theatre of his court. The theatre itself, built by his father in classical-revival style in 1831, had a facade decorated with pillars crowned by a Greek pediment. The building contained not only the auditorium and stage house but also an assembly hall for balls, banquets, and other nontheatrical festivities. Meiningen tried to create the illusion of reality with accurate spectacle and lifelike acting. He studied the distinctions between nations within the same historical period; the result was unprecedented historical accuracy. Meiningen also insisted on using authentic period furniture, and the success of his troupe led to the opening of theatrical supply houses. The Duke designed all scenery, properties, and costumes for his troupe. He saw a scene design as similar to an architectural design in that it directs whatever activity it shelters. He insisted on the continuous and direct relation between the design of a set and the actors' movements within it. Among his innovations was the abandonment of the practice of using only pastel colours for scenery; instead, he used richly coloured banners, foliage, and other devices instead of sky borders as overhead masking. Meiningen was one of the first designers to add to the plastic possibilities of a stage by breaking the surface of the stage floor into different levels. In one play, in which Roman generals go to negotiate with a German chieftain in his forest, the inaccessibility of the German was emphasized by fallen logs on the floor.

From 1866 to 1874 the troupe performed only in Meiningen; for the following 16 years it toured 38 cities in nine countries. It had enormous impact on the history of the theatre, as it achieved complete illusion in every facet of its productions. Saxe-Meiningen is considered the single most important influence on the directors who inaugurated 20th-century stagecraft; he introduced the pictorial mass grouping of crowds. He made every director conscious that scenery must be designed to fit the movements of actors; that lighting, properties, and costumes contribute to the creation of the visual mood; and that every detail of interpretation and stage business merits research, planning, and rehearsal. The Meiningen movement toward unified productions influenced such other major figures as André Antoine in France and Konstantin Stanislavsky in Russia, great apostles of realism and the

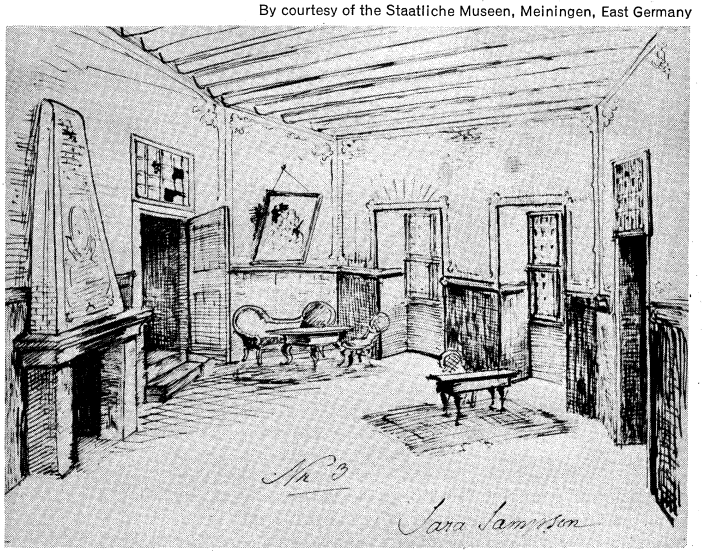




Wagner's staging techniques.
(Top) *Das Rheingold,* painting by J. Hoffmann of the 1876 production in Bayreuth. In the Theatermuseum, Munich. (Bottom) Print by Bergen showing a backstage view of the same production. In the Bibliothèque de l'Arsenal, Paris.
By courtesy of (top) Theatermuseum, Munich, (bottom) Bibliotheque de l'Arsenal, Paris, Collection Rondel; photograph, (bottom) Marc Garanger

founders of modern theatre. Perhaps the reason that the Meiningen productions were more successful than Wagner's is that they sought not to create a passage to the world of absolutes on the stage but rather to present a world of illusion to which the theatre audience could relate directly.

By courtesy of the Staatliche Museen, Meiningen, East Germany



Scene from *Miss Sara Sampson* by Gotthold Ephraim Lessing, designed about 1884 by George II, duke of Saxe-Meiningen. In the Staatliche Museen, Meiningen, East Germany.

Russian imperial theatre. Russian drama in the 19th century also got off to a slow start because of strict government censorship, particularly after 1825. This atmosphere was conducive, as in Germany, to the flowering of Romanticism, especially as manifest in patriotic spectacles. Melodrama, Shakespeare, and musical plays were the backbone of repertory until the 1830s. The best known plays of the new realistic school were those of Aleksandr Ostrovsky, Nikolay Gogol, and Ivan Turgenev.

Until 1882 the imperial theatres, under strict government controls, had a monopoly on productions in Russia's two major cities, Moscow and St. Petersburg. It was not until the monopolies were rescinded that public theatre was able to expand, although the state troupes, such as the Bolshoi in Moscow, continued to offer the most professional productions.

Through the 1850s each theatre had its own few stock settings. The box set had been introduced in the 1830s, but it was to take several decades to become popular. Realism began to dominate scenic design by the 1850s, particularly at the Maly Theatre. Archaeologically accurate settings started to appear in the 1860s, when one theatre hired an archaeologist to help the designer of Aleksey Tolstoy's *Death of Ivan the Terrible.* Previously, settings had followed the neutral style of the German designer Alfred Roller, whose pupils were the major designers for the state theatres. The unified production techniques exemplified by Meiningen were not to be seen in Russia until the Moscow Art Theatre flourished at the beginning of the 20th century.

Victorian theatre. In 19th-century England, the audiences shaped both the theatres and the dramas played within them. The upper class favoured opera, while the working class, whose population in London alone tripled between 1810 and 1850, wanted broadly acted theatre with scenic wonders and machinery. And as the audience grew in number, the Georgian theatre building, which was small and intimate, began to disappear.

Designs of William Capon

In the early part of the 19th century, an important designer was William Capon. Influenced by Loutherbourg, he utilized raking pieces set at various angles and elaborate back cloths as an alternative to flats and wings. His sets were also large enough not to be overpowered by the larger theatres. One of Capon's sets, of a 14th-century cathedral, was 56 feet (17 metres) wide and 52 feet (16 metres) deep. His sets were historically accurate even though the practice of several designers independently decorating the same play was still in effect.

The productions of John Philip Kemble, the manager of first the Drury Lane and then the Covent Garden, marked the shift from Neoclassicism to Romanticism in English stage design. He put theatricality ahead of historical accuracy, as the audience demanded an increased use of spectacle with each passing year. As melodrama became more popular, the effects multiplied until, in an 1820 production of *King Lear,* the storm noises were so realistic that Lear could not be heard. In 1823 Kemble began to reverse this trend and started to use accurate sets and costumes for Shakespeare, and within 15 years archaeological accuracy was dominant.

Another important contributor to the history of staging was Lucia Vestris, an actress and manager at the Olympic Theatre. She controlled all the elements of a production and combined them into a single, integrated unit. She was also enamoured of spectacle and is credited with introducing the box set on the London stage. There is some evidence pointing to the use of a box set as early as 1794, but the first clear evidence of its use was by Mme Vestris in 1832. The sidewalls of rooms were built solidly from front to back so that the actors, instead of entering as usual between side wings set parallel to the footlights, came in through doors set on hinges. She treated the box set realistically, attaching knobs to doors and adding such touches as real books in bookcases.

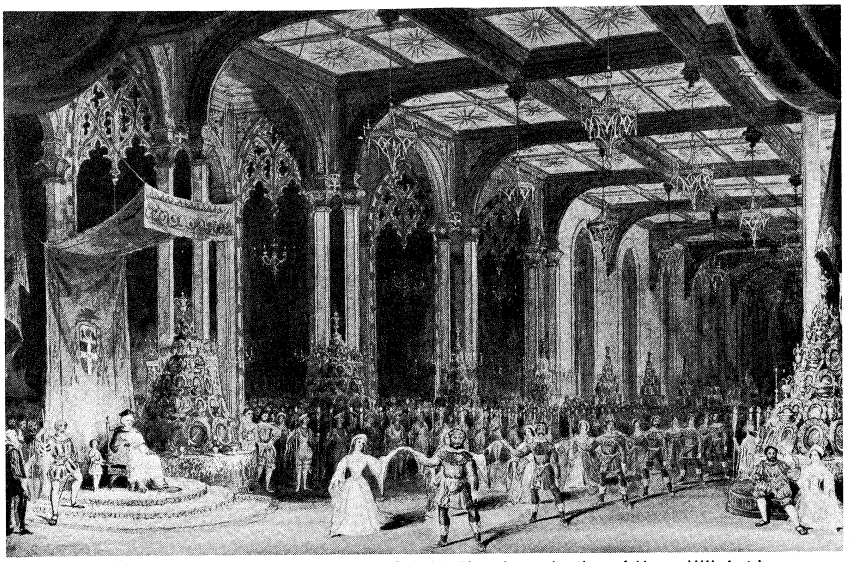
By the 1840s, because of political conditions, many theatres were bankrupt. The next 20 years saw a gradual recovery, with few dramatic innovations in design. One important manager of this era was Charles Kean, a pictorial realist, whose first major attempt to insure accuracy in every production detail was made in 1852 with *King John.* In the following year, Kean gave the audience a printed list of the authorities he consulted to maintain the authenticity of each production. With the assistance of the designers Grieve, Telbin, and Lloyds, he mounted Shakespeare as lavishly as possible while at the same time emphasizing historical accuracy. In this he contrasted strongly with Samuel Phelps, whose Shakespeare productions at Sadler's Wells were simple in staging. Kean, with historical costumes, settings, and pageantry, almost buried his actors.

In the second half of the 19th century, burlesque, extravaganza, and musical drama held the largest audience appeal. The music hall also came into prominence, as incidental entertainment was separated from regular drama. In the period from 1860 to 1880, the theatre continued to expand, and the number of buildings alone increased 50 percent in the first ten years. The first manager of significance was Charles Fechter, who revived interest in the box set. He also discontinued entrances from the wings, heretofore a standard practice of actors even when the wings represented solid walls. Fechter also used a sinking stage, later perfected by the Germans, which allowed scene shifting to be handled in the basement.

The most important management team was that of Sir Squire Bancroft and his wife, Marie Wilton, at the Prince of Wales Theatre. Producing plays by Thomas W. Robertson, they succeeded in melding character and stage business. Spectacle was no longer embellishment but an emphasizing of realistic visual details. The Bancrofts' productions also finally won general acceptance for the box set; they were as accurate in modern plays as in Shakespeare; and they firmly anchored acting behind the proscenium arch. Theatre boxes had been replaced by open balconies, which no longer extended to the proscenium wall. This made it possible for the Bancrofts and others to lower the proscenium arch. In 1880 they even extended the arch across the floor of the stage to emphasize the picture frame. Modern, realistic interior settings were constantly used, and the acting was keyed to the settings. Although Fechter had employed realistic stage business as early as 1860, the Bancrofts were the first to standardize it and make it a tradition.

Trend toward realism

The greatest of the actor-managers was Sir Henry Irving, the manager of the Lyceum Theatre from 1878 to



"The Masque," painting by F. Lloyds, 1859, of Charles Kean's production of *Henry VIII*, Act I, scene 4. In the Victoria and Albert Museum, London.
By courtesy of the Victoria and Albert Museum, London

1901. The best known designers of the period, Craven and Harker, worked for Irving, who also hired archaeologists to aid in the accuracy of productions. He also imported Pre-Raphaelite easel painters: Sir Lawrence Alma-Tadema for *Cymbeline* and *Coriolanus* and Sir Edward Burne-Jones for *King Arthur*. After Irving saw the Meiningen performances in London in 1881, he modelled his Shakespearean productions on what he had seen. His most important staging innovation was to eliminate grooves for shifting scenery. In 1881 Irving introduced the "free plantation," a system by which scenery could be placed anywhere on the stage. Irving was also the first manager to use the front curtain to hide major scene changes, and he completely darkened the auditorium.

Technical innovations in staging methods

The widespread changes in staging methods during the 19th century were possible because of changes in architecture and the development of machinery. The complex scenery of the Victorian theatre required increased flying space about the stage, with a complicated counterweight system, and this in turn fostered one of the most popular features of the staging of this period. Flying, of actors as well as scenery, became an essential element in productions both of melodrama and of Shakespeare.

During this period the arrangement of the stage floor also changed to fit the requirements of spectacle. After 1850 the supporting joists of the stage were usually arranged so that the spaces between them could be used for lowering or raising scenery. Sometimes a vertical panorama would run from overhead through a cut. Floorboards laid over the joists could be moved to form "bridges" through which large set pieces could be raised or lowered.

The changes in stage floors made possible new scenic effects to meet the audience demand. The traps of the Elizabethan and Georgian eras, for instance, were greatly elaborated. The most famous trap was a "ghost glide," which made actors appear to rise from the earth and glide through space. The most lavish of the spectacles were Imré Kiralfy's *Nero* in 1890 and *Venice* the following year.

In the 1870s such three-dimensional scenic elements as platforms and steps became common. The trend was started by Edward Godwin, the father of Gordon Craig, in his sets for *The Merchant of Venice* in 1875. Probably from a desire to increase the number of three-dimensional pieces, Irving sought an alternative to the groove system of scene shifting. In its place, 135 people were hired to arrange and move the scenery. This development itself probably led directly to the removal of the raked stage, as the angled floor made it difficult to set up freestanding pieces. Another factor promoting three-dimensional pieces was that, after the introduction of gaslights, two-dimensional scenery tended to look tawdry under the new brightness. Irving, who was instrumental in accepting new developments in lighting, also introduced black masking pieces at the front of the stage to prevent light spill. Irving's importance to theatre is comparable to that of Saxe-Meiningen. His emphasis on pictorial realism was the high point of 19th-century English theatre, and this era was to continue until World War I.

Development of American theatre. The first record of a performance of a play written by an American was in 1690 at Harvard College. And the first permanent theatre, which was made of brick, was built in Philadelphia in 1766; it was imitative of English buildings in arrangement and general architecture. In 1752 Lewis Hallam, a member of a distinguished theatrical family, arrived with a troupe from England, thus marking the beginning of professional theatre in America. The theatre in America for the next 40 years was similar to English provincial theatre, with simple sets for easy travelling; there were as yet few cities that could afford a theatre. By the 1790s troupes were based in Boston, New York, Philadelphia, and Charleston, and many permanent theatres were being erected.

Early American theatre

The first decades of the 19th century brought not only multiplication of playhouses in the larger Eastern cities but also the extension of theatre to interior regions. The frontier spirit was embodied by Samuel Drake, who took the first company west (to Kentucky) in 1815. Drake designed an adjustable cut drop as a proscenium that could be set up in any large room. The front curtain was a roll drop, and three sets of wings (one each for exteriors, ornate interiors, and simple interiors) and six roll drops, including a garden, a street, and a wood, completed his scenic repertoire.

After 1825 New York City had higher standards of theatrical production and more theatre buildings than any other city in America. Although the audiences thought the theatres luxurious, they were in fact filthy and rat infested. They had little or no fire protection, and between 1820 and 1845, no less than 25 theatre buildings burned down. Although most productions used stock sets and props, occasionally an elaborate or specific set was added. Interest in historical accuracy was not a major production concern until 1830, when Charles Kean travelled with his Shakespeare troupe from England.

In 1849, within months after the Gold Rush, San Francisco started building the first of its many theatres. During the 1860s and 1870s it was a theatrical metropolis, in which vast sums were spent for European opera companies and legitimate theatre from the East Coast. Nothing noteworthy happened in stage design, however, as most troupes were geared for touring and needed only portable

scenery. Real changes came from New York, the only important centre for new ideas after the growth of the railways allowed touring companies to travel with full productions. Now it was finally possible to devote money and space to scenery and properties.

During the latter third of the 19th century, the general scenic trend was toward greater naturalism, with particular emphasis on local colour. The major design innovations came from the managers of troupes permanently based at a theatre. One such manager was Edwin Booth, whose new theatre, opened in 1869, introduced several new concepts. The most important innovation was that the stage floor was flat and had no grooves; elevators raised set pieces from the 50 feet (15 metres) of working space below the stage, and flying machines moved other pieces into the 76 feet (23 metres) of overhead space. Thirteen years before its introduction in England, the "free plantation" method of shifting scenery was used in an American theatre. In Booth's new theatre he abandoned the apron and used box sets almost exclusively.

Another important manager was Augustin Daly, who furthered the trend of realism. While the members of his troupe changed considerably during its three-decade history, it was the best example of the permanent stock company in the history of the American theatre. Daly stressed the unity of every production and controlled each element himself. His first success, *Saratoga*, by Bronson Howard, in 1870, was the first play to give a realistic picture of American life of the day.

American innovators: MacKaye and Belasco

Steele MacKaye, who also was active during this period, holds a unique place in theatre, as an actor, manager, playwright, inventor, and designer. In an age of mechanical inventions, producers were seeking a means of effecting scene changes that would not require an intermission. In 1879, MacKaye filed a patent for a "double stage," a feature he subsequently introduced in the Madison Square Theatre. He put the orchestra on top of and behind the proscenium arch and built an elevator platform on which the next scene might be set while a scene was being played below. The new scene was then merely lowered, with its own stage floor, to the appropriate level.

Among MacKaye's other mechanical innovations were a folding theatre chair with coatrack attached; a sliding stage; a theatre ventilation system; the first installation of an electric lighting system in a theatre (1885); devices to produce cloud, wave, and rainbow effects; the substitution of overhead lighting for footlights; and a process for fireproofing scenery. Some of his more grandiose schemes were never executed for financial reasons. For the Chicago Exposition of 1893, for example, he designed a "Spectatorium" for musico-spectacle-dramas; it called for a sky dome that encircled the stage, a curtain made of light, a sliding stage for scene changes, and an auditorium seating 10,000 people.

In 1896 six men formed the Theatrical Syndicate, which acquired almost complete control over American theatre. They were interested only in commercially profitable works, such as productions involving stars with large numbers of fans. The major opposition to the Syndicate came from David Belasco, a producer and playwright who had been educated in a monastery but ran away to join a circus. His aim as a producer was to bring complete realism to the stage, and it is difficult today to appreciate how spectacularly far he carried this attitude. In 1879, in his production of *The Passion Play*, the story of Jesus Christ from birth to Resurrection, he used a live flock of sheep to follow Joseph and Mary down the mountainside to Bethlehem. When the actor James O'Neill (the father of the playwright Eugene O'Neill), who played Christ, was dragged before Pontius Pilate and crowned with thorns, members of the audience fainted. And after the performance, when O'Neill walked around the city, people sank to their knees and prayed to him. The play aroused such religious frenzy that Jews were assaulted on the street outside the theatre, and a court injunction forbade further performances.

The power of the Syndicate was so great that in 1904 Belasco had to rent Convention Hall, in New York City, which had a leaky roof, for his productions. During the first performance there was a violent rainstorm, and the audience, including Admiral Dewey, had to sit through the last act holding umbrellas. Belasco's productions became so popular that the Syndicate was finally forced to compromise with him, thus breaking their stranglehold on the American theatre.



David Belasco's realistic setting for the New York production of "Tiger Rose," 1917.

By courtesy of the Theatre Collection, New York Public Library at Lincoln Center, Astor, Lenox and Tilden Foundations

THE BEGINNINGS OF MODERN THEATRE

The new Naturalism. Ideologically, the drive to realize the full three-dimensionality of the stage was in accord with late-19th-century positivism—a mixture of science, Darwinism, sociology, and Socialism that postulated that the meaning of human character was to be found in its interaction with the physical, social, and economic environment. If such interaction is to be represented on stage, the production must create the atmosphere, as well as the appearance, of an environment. The plays of Henrik Ibsen and Émile Zola best captured the new Naturalism, and the man who did most to establish this movement in staging was André Antoine, who made his living as a clerk for the Parisian gas company.

Antoine's innovations

Antoine set up the Théâtre-Libre ("free theatre") in Paris in 1887 to produce the new Naturalist plays that were being censored by the French government, including the works of Ibsen, August Strindberg, Gerhart Hauptmann, and Henry Becque. The Théâtre-Libre became a testing ground for authenic production techniques, especially after Antoine saw the Meiningen troupe and Sir Henry Irving's company. He set out to reproduce environment in every detail—the "slice of life" school. In *The Butchers*, for instance, he hung animal carcasses on the stage. When he designed sets he arranged rooms as in real rooms and later decided which was the fourth wall to be removed.

Although Antoine insisted that furniture and props should be not only real but visibly used, he did not exclude the element of distinctive design. His schoolroom for *La Clairiere* was Naturalistic but included a formal patterning that in the objective view of Naturalism shouldn't be there. Though he did capture the fervour of the movement, Antoine's Naturalism has been overstressed.

Antoine's innovations did much to establish the principle that each play requires its own distinct setting. In 1906, as director of the state-subsidized Odéon Theatre, he produced classical plays in which he recreated theatrical conventions of the 1600s. What is significant was that he attempted to establish a realism based on past theatrical conventions instead of on architecture and dress.

The pattern of theatre reform set by France was repeated in Germany during the same period. Similar to the

Théâtre-Libre, the *Freie Bühne* ("free theatre") was founded in Berlin in 1889. Equally significant, a people's theatre, the *Freie Volksbühne*, was organized by a Socialist group in 1890. By 1900 the new realistic drama, including the works of Hauptmann and Arthur Schnitzler, had been accepted throughout Germany.

In England the rising realism was embodied in the plays of Henry Arthur Jones and Sir Arthur Wing Pinero. The first English independent theatre was organized by J.T. Grein, a native of Holland living in London. The founding producers included such eminent writers as George Moore, George Meredith, Thomas Hardy, Pinero, Jones, and William Archer. In the beginning Grein presented controversial continental plays, such as Ibsen's, but was unable to locate any daring English drama until he finally found his man in George Bernard Shaw. The Independent Theatre produced Shaw's first play, *Widowers' Houses*, in 1892. Shaw was a member of the Naturalist school, perhaps the only one with a sense of humour as well as a concern for ideas and social issues.

Russia also followed the pattern of the independent theatre movement as seen in France, Germany, and England. Although Russian playwrights had founded a realistic school of drama, theatrical productions were still similar to those of the 18th century. The Moscow Art Theatre, established by Stanislavsky and Vladimir Nemirovich-Danchenko in 1898, was the first independent theatre to emphasize theatrical production rather than censored or neglected plays. Its first play, Aleksey Tolstoy's *Fyodor Ivanovich*, accurately recreated 17th-century Russia and created a sensation. Stanislavsky sought to create the environment of a play only after visiting the actual place where the play's action occurred and after extensive research.

Reactions to realism. Between 1850 and 1900, the most famous group to oppose Naturalism were the Symbolists. The Belgian Maurice Maeterlinck was the chief theatrical spokesman for the group, which included artists from every medium. In theatre, the group was initially influenced by Ibsen's late plays and Wagner's theories about music drama. In their manifesto of 1886 they proposed that subjectivity, spirituality, and mysterious internal forces represent a higher form of truth than the objective observation of appearances. In 1890 the French poet Paul Fort set up the Théâtre d'Art to produce nonrealistic plays. Subsequently, the Théâtre de l'Oeuvre was organized in 1893 by Aurélien-Marie Lugne-Poe, who, guided by the belief that the spoken word creates the scenery, attempted to create a unity of style instead of an illusion of place and employed as designers such illustrious French painters as Henri de Toulouse-Lautrec, Maurice Denis, Odilon Redon, Edouard Vuillard, and Pierre Bonnard. Lugne-Poe's technique was typified in his production of Maeterlinck's drama *Pelléas et Mélisande*—no furniture or props were used, the stage was lit from overhead and most of the time it was in semidarkness, a gauze curtain was used so that the stage looked clouded in mist, and backdrops were painted in gray tones to convey a general air of mystery.

Appia and the music drama

The two important designers of the nonillusionistic movement were the Swiss Adolphe Appia and Gordon Craig. Appia assumed that the fundamental goal of a theatrical production was artistic unity. He concluded that there are three conflicting elements in a production —the moving three-dimensional actor, the stationary perpendicular scenery, and the horizontal floor. He recommended the use of three-dimensional units, such as platforms and steps, to enhance the actor's movement and create a bond between the horizontal floor and the vertical scenery. To unify the visual elements in a production, Appia advocated the use of light. He saw lights as changing in response to shifting emotions, moods, and actions, and he wanted to manipulate it as carefully as a musical score.

Since Appia designed few settings, and those only for Wagnerian operas, his influence spread primarily through his three books on staging and lighting design published from 1899 onward. He was brutually rebuffed by Wagner's widow, who considered Appia's projects for Wagnerian opera to be the work of a madman. In the Swiss educator Émile Jaques-Dalcroze, however, Appia found a collaborator whose rhythmic gymnastics realized the performer side of Appia's staging theories. In 1911, at Hellerau-Rähnitz near Dresden, they built a revolutionary theatre in which the stage and auditorium formed a single unified space with translucent walls that glowed with variable colouring. The acting areas, which were simple levels at the stage end of the rectangular enclosure, were a concrete expression of Appia's insistence on the primacy of the human body in movement.

Craig and the designer-director

Gordon Craig, the son of the English designer Edward Godwin and the actress Ellen Terry, started his career as an actor and became a designer at the turn of the century, just before the publication of the first of his many books on the theatre. Although he agreed with the major formulations of Appia, he disagreed with Appia's belief that the actor and the text are of prime importance. Craig held the director-designer to be supreme, and the essence of his designs was a revolt against the falseness of the realistic method. He created symbolic form and drew inspiration from Greek theatre. He substituted masses of light and shade for naturalistic detail and used imaginative statuesque shapes instead of illusionistic interiors, and he designed stages in various planes to aid movement. He dreamed of a theatre that would appeal to the emotions through movement alone, giving the audience a kinetic experience. He used curtains occasionally, but the primary and distinctive element in his work was the three-dimensional form.

By courtesy of the Bibliotheque Nationale, Paris; Gordon Craig Collection



Set design by Gordon Craig for the ghost scene in *Hamlet;* frontispiece from *Towards a New Theatre,* 1913, by Craig.

One of Craig's most interesting innovations was scenery made of screens with two-way hinges, which could be assembled in any shape and made in any size. The Abbey Theatre in Dublin was the first to use these screens; its director, the poet William Butler Yeats, was a great admirer of Craig. His work was similar to Appia's in that they both rejected detail and painted ornamentation. For both designers, light was used to create mood, not merely to illuminate the actors and certainly not to simulate nature. Craig's value to modern theatre has been debated. With his flamboyant rhetoric he hastened the end of archaeological painting and static stage pictures, and his campaign for the organic unity of production inspired new generations of theatrical craftsmen. On the debit side, his dismissal of the human being, the actor, has been called messianic arrogance; and his arid abstract shapes seemed interchangeable from drama to drama—the Russian painter-designer Léon Bakst called it northern pur-

itanism. His bombast provided too little guidance to concrete changes in the practicing theatre. In the long run, Appia, in his quieter, thorough fashion, was a more positive influence on contemporary staging.

Confrontations between actor and audience were the prime concern of Georg Fuchs, who founded the Künstler Theatre in Munich in 1907. He held that to be relevant, the theatre must reject the picture-frame stage and the Italianate auditorium. He proposed an indoor amphitheatre in which, on a projecting stage, the action could be thrown forward plastically into the audience. According to Fuchs, the stage designer should not try to produce an illusion of depth since depth is part of the theatre architecture and cannot be added by scenery. Fuchs's view was the culmination of the search for three-dimensionality that had passed through five essential stages since the 18th century. At first, an illusion of depth was achieved by painting perspective scenery on canvas; then the ground plan of the set was rearranged to envelop the actor with the set. The third phase was the introduction of objects for the actor to touch. With Appia and Craig there came the realization that an actor's movement shows up in contrast to inanimate objects, such as platforms and other masses. Fuchs introduced the final phase, in which the actor joins the place for playing to the area in which the audience also belongs. In Fuchs's theatre, designed by Max Littman, the acting area could be extended forward by covering over the orchestra pit, and the size of the stage opening could be changed by adjusting the inner proscenium, which had a door at stage level and a balcony above. The floor of the stage was divided into sections, each of which was mounted on an elevator so that it could easily become a platform. Four cycloramas, surrounding the stage, could be changed electrically.

Search for the plastic stage

The next major director was Max Reinhardt, who started in 1893 as an actor at the Deutsches Theater, in Berlin, and later took it over. He initiated a series of productions that rapidly made Berlin one of the outstanding theatrical centres of Europe. While at the Deutsches, he experimented with staging at a cabaret and was strongly influenced by the advantages of an intimate atmosphere. In 1910 he staged *Oedipus Rex* in the Zirkus Schumann, an amphitheatre, in an attempt to recapture the union of actors and audience that had existed in classical Greek theatre. From 1915 to 1918 Reinhardt directed the Volksbühne Theatre, and 1919 he opened his own theatre, the Grosses Schauspielhaus, which had an open stage and every mechanical device known. This theatre was obviously derived from study of the Dionysian theatre at Athens, and he hoped that it would embody modern life as the arena had embodied the Greek community. Reinhardt was not a traditionalist, however (he showed a completely different approach when he converted a ballroom in Vienna into a formally designed intimate theatre); rather, he was a true eclectic who tried the whole gamut of theatrical invention, directing over 500 plays that represented every style. He believed that theatre, which had become shackled to literature, must be offered instead for its own sake. He re-examined the physical layout of the theatre building and the spatial relationship between the actors and the audience as part of his concept of theatrical style. Believing that the director must control every facet of a production, Reinhardt worked closely with his designers, Ernest Stern, Alfred Roller, Oscar Strnad, Emil Orlik, and the Norwegian Expressionist painter Edvard Munch. His productions usually centred on a particular motif or the staging conventions of a historical period. After beginning with a three-dimensional, drab naturalism, he adapted the indeterminate solids that Appia had inspired and then the decorative surface styles derived from contemporary art movements such as Art Nouveau, the Austrian Sezession, and Munch's Expressionism. He used unit settings with plugs (detachable parts of a setting that alter its appearance when in place), revolving stages that presented different facets of the same construction; he adapted the conventions of Oriental theatre; and he even totally renovated an auditorium interior for his production of *The Miracle* in 1911. He exerted a strong influence on the designers of the German expressionistic cinema as well as on stage artists. His eclecticism helped to reconcile the differences between conflicting movements by romanticizing the realistic and fleshing out the idealistic with solid structure.

Reinhardt's eclecticism

While most English productions during this period were in the realistic tradition, several steps were being taken toward nonillusionistic staging. One director, Sir Frank Benson, began by mounting plays in the realistic style of Sir Henry Irving but by 1900 had started to simplify his staging. He produced Shakespeare's plays with only a few stock sets, placing primary attention on the actors. William Poel, also producing Shakespeare, attempted to recreate an Elizabethan theatre. His attempt to reproduce the actual theatrical conditions of the original performance instead of historical accuracy was also being tried at the Royal Court Theatre in Munich, at the Odéon in Paris by Antoine, and elsewhere by Reinhardt and others.

English nonillusionistic staging

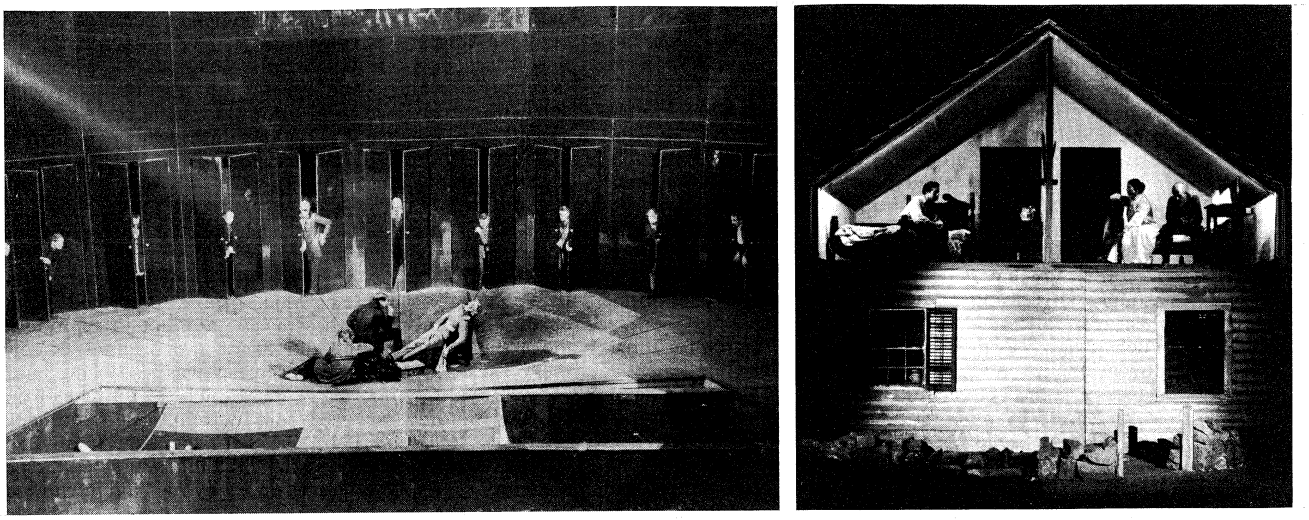
Between 1912 and 1914, the actor-manager Granville-Barker staged Shakespeare so that the action could be continuous, an approach influenced by his having worked with Poel. He remodelled the Savoy Theatre by adding an apron and doors in front of the proscenium. He divided the stage into three parts, the apron, a main acting area, and a raised inner stage with curtains. This permitted a continuous flow of action and eliminated the rearrangement of scripts that had previously been necessary for nonillusionistic staging. Norman Wilkinson and Albert Rutherston, artists with reputations outside the theatre, were his principal designers, and their usual settings were brightly painted, draped curtains.

Russian influences. There was, during this period of ferment, another movement that ran counter to many of the others. For the most part, the reaction against realism had taken the direction of simplifying or rendering symbolic the ordinary elements of the set. The other direction was that of the Russian impresario Sergey Diaghilev, the founder of the Ballets Russes, whose designers deliberately painted in two dimensions but emphasized the artificiality through false perspective and exaggeration. This was particularly evident in the work of Léon Bakst, Aleksandr Benois, and other members of the St. Petersburg school founded by Wronbel. In opposition was a Moscow group that included Mikhail Larionov and others who were in touch with such Parisian movements as Cubism. Diaghilev enlisted their services and later those of French painters of the stature of Pablo Picasso, Henri Matisse, Joan Miró, and André Derain.

In 1905 Stanislavsky established a studio to experiment with nonrealistic production styles, with Vsevolod Meyerhold in charge. They did not get along, for Meyerhold rejected the introspective drama of Anton Chekhov and the Moscow Art Theatre. He conceived a stage composed of machinery and abstract arrangements of scaffolding, with the actor as one who had to undergo rigorous training to climb around the scaffolding. This training, which became known as Meyerhold's system of "biomechanics," assumed that the actor is a large and wonderful engine composed of smaller engines, such as muscles and tendons. For him, the new problem of the theatre was how to get the engine in full motion—a machine-age equivalent of the commedia dell'arte principle of the actor as gymnast.

Meyerhold was succeeded at the theatre of the actress Vera Komissarzhevskaya, in Moscow, by Nikolay Yevreinov, a director who emphasized theatricality. His most important innovation was the "monodrama," in which he attempted to make the audience the alter ego of the protagonist. In his productions, Yevreinov treated the staging as though it were seen only through the mind of the protagonist. The scenery, lighting, and sound effects reflected the major character's moods and feelings.

To Alexandr Tairov, the actors under Meyerhold were puppets in his control, while the actors at the Moscow Art Theatre under Stanislavsky were dominated by the playwright. He attempted to train a company of actors who would be capable of improvising on an idea in the tradition of the commedia dell'arte and developing it in front of an audience. He believed that the future of the theatre



Early 20th century.
(Left) Anti-illusionist stage of Vsevolod Meyerhold designed for his production of *Revizor* (*The Government Inspector*) by Gogol, Moscow, 1926. (Right) A simultaneous setting by Robert Edmond Jones for O'Neill's *Desire under the Elms*, Provincetown, 1924.
By courtesy of (left) Society for Cultural Relations with the U.S.S.R., (right) Theatre Collection, New York Public Library at Lincoln Center, Astor, Lenox and Tilden foundations

lay both in the synthesis of all the arts and in the development of complex machinery that would act as an extension to the performer. Influenced by Appia and Craig and assisted by lighting and the choreographed movements of actors, Tairov used space to create dynamic relationships. He saw the stage as the "keyboard" of actors and varied it with ramps, levels, stairways, and abstract shapes.

The appearance of the Ballets Russes in Paris in 1909 started a new wave of nonrealistic experiments in France. Among those who were particularly impressed was Jacques Rouché, the director of the Théâtre des Arts. His aim was not the extreme stylization of the Russians but a simplicity in which colour and line would subtly characterize a mood. He in turn inspired Jacques Copeau, who in 1913 published a manifesto for a new theatre. Copeau's philosophy was the reverse of Meyerhold's and Tairov's; he argued that the director's job is faithfully to translate a play into a "poetry of the theatre." Believing the actor to be the only important element in a production, he advocated return to the bare platform stage; he was bolstered in his stand by Appia's precepts. In the Théâtre du Vieux-Colombier, he removed the proscenium arch to create a raised open platform. At the rear of the stage he placed an alcove with a balcony, making the acting space similar to that of an Elizabethan theatre.

Copeau's formal stage

Late-19th-century technical developments. The theatre is a social institution as well as an art, and it responds more slowly than other arts to changes in form. The profound innovations of the 19th century in playwriting, directing, and acting were not immediately assimilated into the staging of plays. It was not until the late 1800s that significant technical innovations were made. Some of these innovations were motivated by the need to shift the heavy three-dimensional scenery that had replaced wings and drops. One such invention was the revolving stage, which was discovered in Japanese Kabuki theatre and adopted in Munich in 1896. Able to accommodate several settings, which were changed simply by revolving the stage, it became very popular after 1900. The rolling platform stage, introduced in Berlin in 1900, allowed sets to be mounted offstage on the platform, which then could be wheeled in on tracks. In Dresden a few years later, a complex variation was designed to use both rolling platforms and elevator stages.

Many German theatres also installed plaster domes, *Kuppel-horizants*, which curved over and around the stage, thus eliminating the need for overhead and side masking pieces (and ultimately rendering difficult any scene changing except those using a complete elevator stage). Under the influence of Wagner's theatre at Bayreuth, auditoriums were also modified during this period. Boxes were removed, the number of balconies was decreased, and centre aisles were eliminated. And as a result of the tendency to remove the separation between the actors and the audience, the apron stage returned and the proscenium arch disappeared. For the first time since the 1600s, the Italian-style theatre building had competition.

American stagecraft. The currents of innovative stagecraft eventually reached the United States. The first migration was represented by the Viennese Joseph Urban, who went to the Boston Opera before World War I, taking with him an entire atelier of draftsmen and scenic artists. They decorated the opera with the patterns of the Austrian Sezessionist movement combined with a lush, fairybook world. Urban moved into the musical comedy field and displaced the acreage of painted vistas and box sets that were manufactured by the stock scenic studios. The next change grew out of marginal experimental groups, such as the Provincetown Playhouse and the Neighborhood Playhouse, which fostered designers such as Robert Edmond Jones, Cleon Throckmorton, and Aline Bernstein. By the middle 1920s, their simple, tasteful romanticism had invaded Broadway as the groups had become commercial and as the more artistic theatre managers extended commissions to the free-lance designers. The industrial designer Norman Bel Geddes entered the growing ranks of theatre artists and devised grandiose projects and engineering marvels. With the impetus provided by ecstatic reports from Europe on the work of Reinhardt, Copeau, Meyerhold, and Tairov, American directors such as Arthur Hopkins and Phillip Moeller stressed the synthesis of all the elements of production into a persuasive whole. The imaginative poetry of Robert Edmond Jones was balanced by the sensible craftsmanship of Lee Simonson for the Theatre Guild. Simonson as the exponent of "selective realism" was more attuned to the practicalities of the earthbound psychological problems that provided the staple fare of Broadway's "serious" drama. By the 1930s, scenery consisted of solid carpentry and tasteful furnishings that were tailored to the mood, atmosphere, and mechanical requirements of the individual play. The Urban style in musical comedy design was replaced by that of Albert Johnson—a style of loose colour plus calligraphic line that was more appropriate to the sharp revues that prevailed until World War II. In staging musicals, a peculiar division persisted between the direction of the plot and comedy segments and that of the "production numbers" —the sumptuous song-and-dance displays under the overall supervision of a "stager" who was noted for his taste. Director-producer George Abbott surmounted this artificial departmentalization in an important step forward in the development of the rhythmic, lively musical show that became America's contribution to world theatre.

Staging musicals

During the 1940s many American dramatists began to break up the standard three-act format into linear strands of brief episodes. The designers devised scenery that provided a shorthand indication of locale, which could be wheeled onstage and off rapidly. The remnants of three-dimensional architecture were soon supplanted with more mobile and graceful skeletonized fragments. The convention of the toothpick framework combined with gauzes for smooth transitions was particularly helpful for the simultaneous scenes spread across the stage: indoors-outdoors, upstairs-downstairs, separated places. The skeleton-plus-gauze pattern was the characteristic scenery in the theatre of the 1940s, 1950s, and 1960s. Musicals also borrowed from the graphic arts such techniques as collage and transparencies.

In the 1950s the bitterly realistic plays of Britain's so-called Angry Young Men brought back naturalistic detail and a consciousness of environment to flesh out the abstract framework, though little attention was paid to scenery in these productions. In decided contrast were the lush revivals of Britain's Royal Shakespeare and Old Vic companies. The Absurdist dramatists, such as Eugène Ionesco in France, when they alluded to staging at all, tended to prefer "normal" surroundings as a counterpoint to the wild behaviour and nihilistic philosophy of their characters.

Recently, there has emerged a dissatisfaction with the skeleton tracery; particularly for certain classics, fragments of primitive masonry—Stonehenge fashion—were adopted. This type of scenery reflected not only an interest in myths and archetypes current in all the arts but also a romantic reaction to the chaos of the present and a search for artifacts of a time when life appeared to have had more meaning and grandeur.

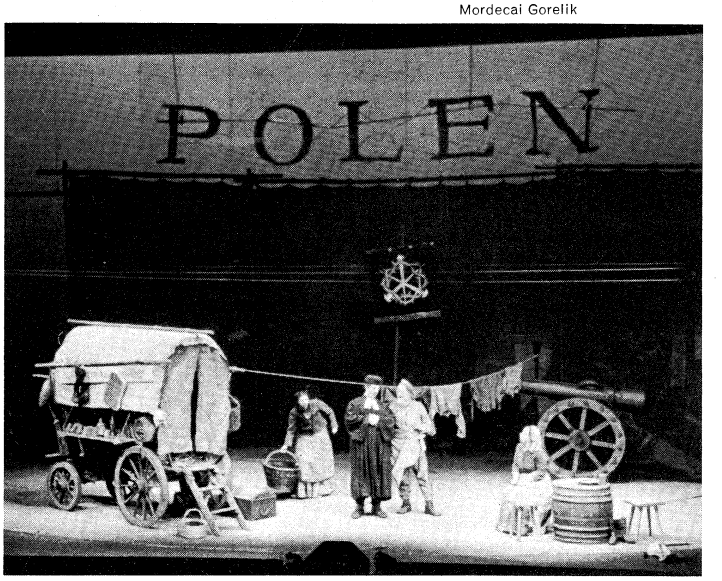
The internationalism of theatre today has erased national and local styles in decor and staging. Two prime tendencies in modern production—one stemming from the work of the Germans Bertolt Brecht and Erwin Piscator and the other from the Pole Jerzy Grotowski—are global in influence.

MODERN STAGING AND DESIGN

Brecht and Piscator: epic theatre. Just as the modern theatre can be said to have come into being with Antoine and the Naturalists, the contemporary theatre of the "theatrical" was ushered in by the collaboration of two Germans—director Erwin Piscator and dramatist Bertolt Brecht. After World War I, during the Weimar Republic, Germany sheltered a wide variety of drama including Max Reinhardt's productions, the People's Theatre under Socialist Party sponsorship, the cabaret theatre of the artists of the Dada movement, and theatricals staged by the influential Bauhaus school of design, as well as standard boulevard theatre. Piscator's approach to production was personified in his production of *The Good Soldier Schweik* in 1928, which returned to the concept of the stage as a platform and added all the devices of modern technology. The result was strikingly different from traditional dramaturgy; Piscator utilized his assorted machinery to shake up his audiences, to force them to think, and, hopefully, to change the world as a result of the lessons learned in his theatre. Treadmills, revolving stages, slides, films, cartoons, vaudeville, music, and sound were used to drive home the content. This total bombardment of the audience, this use of the stage as tribunal, was called "epic theatre." The theoretician and codifier of epic theatre was Brecht, who maintained that in conventional theatre the uncritical synthesis of words, music, acting, and production were passively absorbed into the emotions of the audience in a manner that tended to perpetuate the ideology of the status quo. In place of that, he called for the conscious separation of the elements of theatre, a breakdown into logically analyzable quantities. This requires a certain objective distance from the material on the part of both the performer and the spectator, which Brecht called the "alienation effect." Conventions were borrowed from Meyerhold and the Oriental theatre to formalize the cool approach to staging. With a score by Kurt Weill, Brecht's *Three Penny Opera* (1928) established him as a dominant figure in Western theatre. That position was consolidated after World War II with the founding of Brecht's own company, the Berliner Ensemble, by the German Democratic Republic. The ensemble under Brecht presented exciting, concrete works that corrected many previous misinterpretations of Brecht that had been staged in London, Paris, New York, and elsewhere. The ensemble had a profound influence on Peter Brook and Peter Hall in England, Roger Planchon in France, and Giorgio Strehler in Italy. In the United States, Mordecai Gorelik's book *New Theatres for Old* (1940) was decades ahead of its time as a social history of theatre and as a polemic for epic theatre.

Staging of the "epic theatre"

Mordecai Gorelik



Setting for a scene in *Mutter Courage und ihre Kinder*, staged by Bertolt Brecht for the 1949 production by the Berliner Ensemble.

Brecht affected production concepts in a fundamental sense. Although there had been a revival of presentational, as opposed to representational, staging by Meyerhold and in the various attempts to reconstruct the Elizabethan stage, Brecht went far beyond these formal exercises. He used the bare platform as an arena, unmasked the full stage, and exposed the lighting equipment as tangible preconditions of an approach that negated all Romantic-Symbolic-Illusionist-Idealistic theatre. The present was treated as a historical period, and physical objects used onstage were chosen as artifacts to be handled by the performers in a manner that lifted the drama out of its customary context of an unchanging human nature set in front of different backgrounds. Brecht's designers Neher, Otto, and von Appen, evolved stage structures that were in themselves pertinent comments rather than stylistic entities. In keeping with Brecht's insistence on the radical separation of the production components, to avoid a synthesis of the kind that engendered passivity in the audience, the scenic element of his productions tended to be confined to side and back blackboards, on which the designer sketched his opinions. The danger inherent in this departmentalization is that the actor could fail to develop any organic interaction with the stage environment. Such an interaction is essential if the spectator is to put together the man and the world presented onstage.

In the 1930s Brecht's political ideas reached the United States principally through the agitational-propaganda troupes of the progressive labour movement. A full-scale expression of his ideas was a section of the U.S.-government-sponsored Federal Theatre called the Living Newspaper. Based on factual data, the Living Newspaper took social topics and developed them by any device—films, projections, loudspeaker commentary, burlesque bits, patter songs, ventriloquists, masks, puppets—to render graphic the dry statistics. The exuberant theatricality of the Living News project has never been rivalled.

Grotowski's Lab Theatre. The other major tendency in today's theatre arises from an investigation of the sources of the theatre's uniqueness and strength. The prophet of this search was Antonin Artaud, a French stage and film actor, who had a vision of a total, visceral theatre with the potency and terror of primitive myths. Like Craig and Appia, his actual performance was minimal but his writings inspired many directors, such as Peter Brook in England and Gerald Savory in France and Jerzy Grotowski, whose Polish Lab Theatre, a fiercely dedicated acting ensemble, sought to cut through the bonds of the polite literary tradition to rediscover basic human drives and conflicts. The actors undergo exhaustive exercises, designed to break down the layers of superficial technique and repressions. Each production is conceived as a unique entity, requiring its own playing space and actor–audience relationship; the rectangular hall-theatre is rearranged for each production to create a configuration consisting of onlookers, company, properties, dividers, and boundaries appropriate to the single dramatic piece. Grotowski relies on his own dramaturgical skills in freely molding the texts and on his architect-designer for the carving of the space in which the ritual takes place.

The influence of Artaud

By discarding pictorial backgrounds, disciples of both Brecht and Grotowski put more emphasis on objects, which appear, impinge on the performers in a meaningful fashion, and then disappear. Thus, the theatrical event in process supplants the story enfolded within a static environment. For the director and his designer, staging and design fuse into an indivisible organism.

II. Scenic design

DEVELOPMENT OF SCENIC DESIGN

The execution of scenic effects passed from the guilds in the late Middle Ages to artists, with their apprentices and assistants, during the Rennaissance and the Baroque and to commercial scene painting studios from the mid-19th century until today. Each tiny Baroque state maintained a court painter and staff to embellish the court revels, masques, and operatic presentations. Large permanent companies such as the Paris Opéra, La Scala in Milan, Covent Garden in London, and the Vienna State Opera employed their own artists and shops. The independent actor-managers, stock companies, and travelling troupes and acts of the late 1800s ordered stock settings from firms that were prepared to turn out acres of mountain crags, sumptuous temples and palaces, picturesque hovels, and town squares. Historical research was crammed into every meticulously painted border, side wing, and backdrop, but no thought was given to the particular requirements of the individual scripts. The major firms printed illustrated catalogs so that single hangings or complete backgrounds could be ordered by number through the mails. Leading managers, such as Sir Henry Irving in England and Max Reinhardt on the Continent, and the French art theatres commissioned painters who either did the painting themselves or supervised its execution from their sketches. In the United States, even on Broadway, the standard procedure well into the 1920s was for the actor-producer-director to dispatch a ground plan to the scenic shop, which made up a coloured model based on research and, after approval of the design, made the finished setting. The scenic artists who made such settings were highly skilled artisans, with thorough art schooling; though they frequented art galleries and museums, they seldom attended the legitimate theatre. As a result, free-lance designers with a better understanding of theatre began to play a greater role on the New York stage, and studios arose that catered to their new, personal styles. Bergman Studios, for example, executed works by Robert Edmond Jones, Lee Simonson, Norman Bel Geddes, Donald Oenslager, and Jo Mielziner, and Triangle Studios, an outgrowth of the Urban migration from Vienna and Munich, concentrated on musical comedies. Generally, carpentry, painting, property, and drapery shops were separate businesses. Not until the 1950s would all branches of scenery making in New York be found under one roof. With the diminution of theatrical activity, the all-in-one studio is an economic necessity. In London, the shops are still separate. Most resident theatres, such as the Metropolitan Opera and the Los Angeles Civic Light Opera, maintain their own shops.

Early scenery shops

PROCEDURES

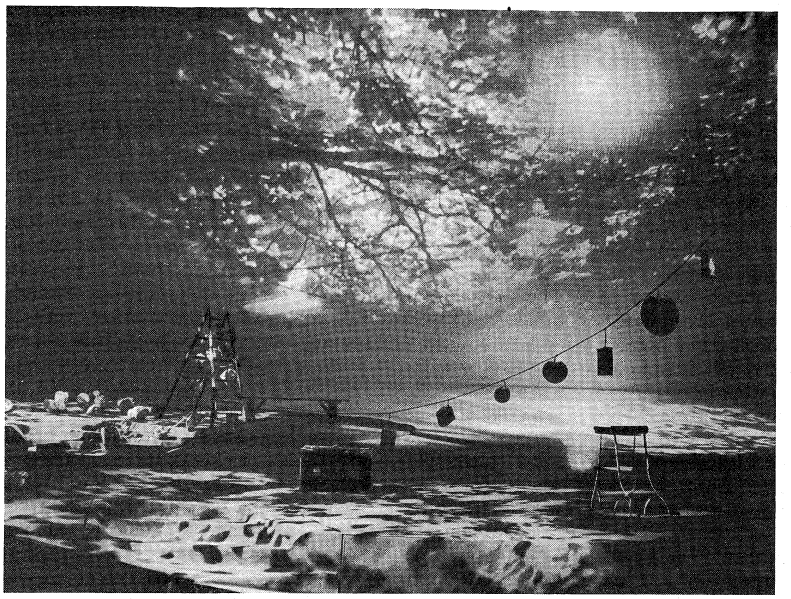
Design procedures are fairly similar throughout the world. First, the script is studied closely and notes are made of mandatory action and mechanical requirements. Next, there is a consultation with the director on a production scheme. Thumbnail sketches are expanded into colour renderings or a model, if the structure is complex or the setting is primarily plastic, usually on a scale of one-half inch to one foot. Extensive research is undertaken and incorporated into the designs. Upon approval, the sketches are translated into working drawings. The execution shops are sent a ground plan, a section to determine masking and light placement, elevations of all units, and all details that require larger scale. The designer does not break down the elements into construction pieces or into actual framing or building details; this is done by the layout man in the scenic firm. Usually there is a great deal of questioning and discussion between the artist and the technicians over the mechanics and the use of materials before the final drawings are consigned to the shops.

In spite of the marginal financial status of the legitimate theatre, the physical layout of a modern scenic shop is extensive; vast amounts of space and equipment are required for the wide variety of designs to be translated into solid items, especially in the rush of preparations for the theatre season. Templates for flats, which once occupied the major amount of floor space, have been supplanted by three-dimensional fabrication, with scene-shifting wagons and overall stage decks. More recently, a blacksmith shop has been a requirement, since metal has replaced lumber in the scaffolding and in delicate open-tracery sets. The power tools, both for lumber and for metal, and all the workbenches are mounted on castors for mobility; essentials such as traveller tracks, winch motors, and the material stock for immediate use are stored around the edges. The paint shop also has required a larger space since looser, wetter techniques were adopted and bulk objects replaced flats and drops, which could be painted on upright frames. The drapery department produces soft maskings, drapes, upholstery, and any other special scenery covering. Vacuum form machines turn out ornaments wholesale and also make sizable, lightweight plastic objects that could not be carved in Styrofoam or wood. The drafting layout room is where mechanical problems are solved, the work distributed, and models translated into working drawings. Working models have come back into use to represent the sculptured units on the open stages that have replaced traditional flat scenery of the peephole tradition. If there is room, an acting company may practice on an elaborate structure for a few days before it is moved to the theatre. Ideally, the electrical shop is incorporated into the scenic complex to service the units that onstage will come within the province of the production electrician.

Property making

Properties, once produced in quaint workshops, are now a part of the scenic firms' mundane output. Celastic casting has taken over from wood turning and whittling. The procurement of ready-made props, furniture, and dressing is handled by the property master, but the artistic choice resides with the designer. On today's open stages the primacy of properties demands the absolute rightness of every object in the visual ensemble. The theatre has sheltered the last remaining medieval craftsmen, the scenic painters and the blacksmiths; the production property master is the last all-around handyman, able to find or make any intricate contraption called for by the dramatist.

Despite the ingenuity and invention of the accredited shops, they remain deficient in two key sectors: engineering, beyond a simplistic mechanical level; and multimedia, as an organized production technique. Backstage-nurtured electricians and carpenters have not been able to



Stage design by Josef Svoboda for the 1958 production of *August Sunday* by the National Theatre, Prague, showing his use of patterned lighting to achieve special designs.
© Jaromir Svoboda

cope with complex electronic gear, and advanced film-slide devices have been left to gypsy groups. Only Josef Svoboda's laboratory at Prague's Czechoslovakian National Theatre succeeds in truly integrating science with the theatre.

Working relations in the theatre still reveal influences of their long historic development from their origins within the medieval church, through the guilds, the artists' ateliers under noble patronage, the permanent companies and their staffs, the itinerant scenic painters (the term "journeyman scenic artist" continues until this day), to the modern practice of using shops plus free-lance designers. In Britain there are now artists' and technicians' organizations primarily concerned with information exchange. In the United States there is a rigid union and guild structure geared to economic concerns. Nearly all the crafts in film, theatre, and television are under the jurisdiction of the International Alliance of Theatrical Stage Employees of the United States and Canada or of the United Scenic Artists. The production heads of departments—the master electrician and the property master—are personally selected for each attraction; under the designer's direction, they oversee first the execution of the designs by the shops, then the setup in the theatre, and finally the running of the production in the theatre. The commercial playhouses of America have no built-in equipment beyond the overhead gridiron system (*i.e.*, a series of girders from which scenery can be suspended) and the electrical power, so that incoming shows must travel with all the equipment they need, including portable switchboards and cable, as well as all lighting units.

The American system of specialization of crafts has resulted in a loss of creativity on the one hand, balanced by the maintenance of minimum craft standards, scales, and working conditions on the other. This closed system has made designing a profession and prevented the dilettantism found in other Western nations, where fine artists often accept theatre commissions and professional theatre designers are rare. This system accounts in part for both the conservatism and the technical brilliance of Broadway musical design.

The economy of New York and London theatre, which is based on one production at a time, has suffered under the competition from mass media. The relative inefficiency of the handicraft mode of operation versus automated industry has aggravated the economic problem.

Scenic design in the future

Designers have gravitated toward subsidized resident theatres in which they can work with good physical equipment and permanent creative staffs.

The future of design and staging is difficult to chart, since Western theatre is in a period of transition. The competitive nature of commercial entertainment has encouraged a pervasive exhibitionism, which in turn has prompted a rebellion against the pictorial and the bombastic and a search for the roots of what is unique in theatre—the actor-human and his immediate object world. The uncritical rejection of all technique, however, has made the concrete dramatic product tend to be arid. Ultimately, this new theatre may be fleshed out with a fresh type of colour and magic that will attract and excite more of the population than its initial primitive and puritanical experiments have done. The designers, with their command of visual tools, will then play a leading part in such a theatre. (Ho.B.)

III. Stage lighting

EARLY FORMS OF STAGE LIGHTING

The classic Greek *theatron* was built in the open air, usually on a hillside, and placed so that the afternoon sunlight came from behind the audience and flooded the performing area with light. The larger Roman theatres were also outdoors, but the added luxury of coloured awning stretched over the spectators softened the glare of the sun. Later, in the Middle Ages, the miracle and mystery plays were performed outdoors on the front steps of the church and the adjoining square. In England, the pageant wagon, complete with actors and properties, was drawn through the main street of a town. Until the 16th century, the theatre continued to be basically an outdoor institution.

Under the patronage of the aristocracy in Italy, private performances, pageants, and tableaux began to be given indoors. Serlio, the Italian architect, gave considerable attention to theatre design and in a treatise written in 1545 discussed theatre construction and the creation of lighting effects. He recommended placing candles and torches behind flasks filled with amber- and blue-coloured water. Palladio's indoor theatre, also in Italy, used the common everyday light sources: torches, pine knots, open wick lamps, and tallow candles. In England at the end of the 16th century, the Globe Theatre was used for summer performances of Shakespeare's plays, but in winter performances were given in the completely enclosed Blackfriars Theatre. Artificial light, mainly candles, was used in several indoor theatres to light the stage and the auditorium.

Beginnings of artificial lighting

In the early 17th century, Inigo Jones introduced several innovations in lighting and stagecraft, using reflectors to intensify the light sources and making use of colour on stage. The earliest known definite description of stage

lighting may be found in *Sciena d'Comoedia* (1628) by Joseph Furtenbach. He describes the use of oil lamps and candles set in a row along the front edge of the stage but out of sight of the audience, as well as vertical rows of lamps behind each wing at the sides of the stage. The common method of lighting the stage and auditorium was by means of tallow candles. As seen in old prints, these candles were mounted in crude hoops or chandeliers, which were hoisted aloft on pulleys to hang in dripping splendour. Gold decorations applied to the interior of the auditorium caught the many reflections. The inconvenience of the lighting system was that candles were expensive and hard to control. The twisted wicks had to be constantly trimmed during the performance, and this was the duty of the snuff boy. A transformation from light to darkness was effected by the agile skill of the candle snuffers.

When David Garrick used footlights at the Drury Lane Theatre in 1765, he masked the candles with metal screens. By 1784, when Richard Brinsley Sheridan managed the Drury Lane, all lights used to illuminate the stage were out of sight, hidden by the now familiar wings and borders.

A floating oil wick lamp was replaced after 1783 by the Argand oil lamp, in which the cylindrical wick was enclosed in a glass chimney to steady the flame and provide a brighter, whiter, and cleaner light source. And so the chimneyed oil lamp replaced the candle, but it was still hung in clusters above and bracketed to the walls. At the Haymarket Theatre in London, the oil lamps had chimneys of white and green glass that were controlled by levers, so that raising or lowering the chimneys could effect light changes. Actually, the result merely made the actors and scenery visible. Stage design and stagecraft had now advanced as far as was technically possible under the limitations of low-intensity stage lighting.

Introduction of gas lighting

The first major advance in several centuries was the introduction of gas lighting. A Scottish engineer, William Murdoch, developed a practical method to distill gas from coal for illumination. The first successful adaptation of gas lighting for the stage was demonstrated in the Lyceum Theatre, London, in 1803 by a German, Frederick Winsor. The Chestnut Street Opera House in Philadelphia installed a gas lighting system in 1816 and supplied its own gas by installing a gas generator on the premises. (Gas stations and city mains did not come into use before 1850.) The advantages of gas lighting were immediately realized and exploited, despite the initial cost. No new methods of lighting, however, were devised for stage lighting. The conventions remained the same: footlights (*i.e.*, a row of lights across the front of the stage floor); border lights (a long horizontal row of lights used for the general lighting of the stage from above); and striplights (a row of lights, usually mounted in a trough reflector and placed in the wings to illuminate specific portions of the stage or setting).

Even without a chimney, an open gas jet flame was brighter than oil lamps or candles. The additional advantage was control: by varying the control valves from a central point, a smooth increase or decrease of light could be effected, and at variable speeds. For the first time, to add to the realism of the play, the auditorium lights could be darkened. Elaborate central control systems were devised, with a main regulator, branch mains, more and more regulators, branch mains, and valves. This growing array of valves and pipes was organized into circuits and displayed on the "gas table"—the forerunner of the modern switchboard.

But there were also disadvantages to gas; namely, heat, offensive vapours, and the serious fire hazard of the open flame. A protective code was established to provide guards, screens, and glass chimneys and was made compulsory. In 1890, after the introduction of electric lighting, the incandescent gas mantle was developed. Although the mantle greatly improved the quality of light —brighter and whiter—the hazards of fire still remained.

Introduction of the limelight

Although Thomas Drummond, a British engineer, invented the limelight in 1816, it did not come into general use until some 30 years later. Limelight is produced by directing a sharp point of oxyhydrogen flame against a cylindrical block of lime. The tiny area of lime becomes incandescent and emits a brilliant white light that is soft and mellow. As the lime is slowly consumed by burning, the block of lime has to be slowly and constantly turned by an operator to supply the flame with a fresh surface. Since the brilliant area was very small, the addition of a mirrored reflector was necessary to give accurate control.

The intensity of the limelight permitted it to be directed onto the stage from the auditorium. Since it offered control as well as intensity, the limelight was quickly adapted to follow individual performers about the stage. The sharpness produced by the small point source made possible the creation of realistic effects, such as sunlight and moonlight, and moving effects, such as clouds, water, and fire.

ELECTRIFICATION

An advance of great importance was the introduction of the electric carbon arc, which was exhibited in experimental form in 1808 by Sir Humphry Davy. Davy used a large battery to provide current for his demonstration, since no means of mechanically generating electricity had as yet been developed. In principle, an "arc" of current across the gap between two rods of carbon heated the tip of the positive carbon to brilliant incandescence. As the carbons gradually vaporized, however, they had to be continually adjusted to maintain a constant gap. The use of the carbon arc was not practical until some 50 years later, when adequate generating sources were developed. The Paris Opéra developed the earliest electric arc effect —a beam of sunlight—as early as 1846. By 1860 the Paris Opéra had also developed a lightning machine, a rainbow projector, and a luminous fountain. Most important, the company made the earliest spotlight, a carbon arc and reflector housed in a hood, which included a lens and a shutter.

Development of incandescent light

The next great advance in lighting was the development of the incandescent electric lamp, in which light is produced by a filament electrically heated to incandescence. The invention of a practical electric lamp by Thomas Edison in 1879 marked the beginning of the modern era of stage lighting. Gas was quickly discarded; within one year the progressive Paris Opéra introduced the new system. Two years later, at the Electrotechnical Exposition in Munich, a small theatre was erected that used electric lighting exclusively for both stage and auditorium. The success of the experiment received worldwide acclaim. In London, the Savoy Theatre was the first to install the new lights; while in Boston, the Bijou Theatre followed the new trend in 1882. The following year, 1883, the Landestheatre in Stuttgart, the Residenztheatre in Munich, and the Staatsoper in Vienna were the first completely electrified theatres.

At the turn of the century, incandescent lamps were in almost universal use for stage lighting, but no new methods or techniques of lighting appeared. The conventional footlights, border lights, and striplights were merely electrified, and the arc light was used for concentrated light sources. Gradually, new improvements provided brighter lamps that were both more durable mechanically and available in larger wattages. Metallic filaments replaced carbon, and finally drawn tungsten filament lamps appeared in 1911. The use of inert gas in place of a vacuum produced lamps of even higher efficiency and larger sizes. The introduction of concentrated coil filaments made practical the development of the incandescent spotlight. And the refinement of the incandescent spotlight added an exciting new tool for the advance of stage lighting and the further development of stagecraft. Gradually the arc spotlight was replaced by the new incandescent spotlight, which provided all the advantages of flexibility and control.

In his music dramas, Wagner suggested new possibilities for the use of light and design in a unified production—a lyrical synthesis. Adolphe Appia and Gordon Craig gave tremendous impetus to the new plastic stagecraft. Their conception of the stage was a cubic volume of space bathed in a continuous play of functioning light. All the

vast optical effects of Baroque design previously obtained with paint were now possible by means of light.

About 1902, in Germany, Mariano Fortuny developed an elaborate system of soft reflected light using arc lights bounced off of coloured silk fabrics. The simulation of natural lighting was remarkable, but the entire mechanism was too bulky and intricate and required the construction of a special theatre. In the course of his experiments, Fortuny evolved a dome-shaped cyclorama, a rear wall surfaced in plaster. Flooded with light, it gave the illusion of infinite space and was the perfect means of simulating spectacular sky and background effects. Because it was dome-shaped, however, it occupied a large amount of stage space and tended to distort optical projections. In modified form, as a curved, hanging cyclorama, it became an indispensable tool of the new stagecraft. Earlier, Sir Henry Irving had used transparent coloured lacquers to coat lamps to produce colour effects, using separate circuits for each colour. Irving was also the first producer to introduce organized light rehearsals in his productions. David Belasco, with his electrician Louis Hartman, developed a standard of realism in stage lighting that anticipated the motion picture and that remains in effect in today's theatre. In their lighting laboratory, Belasco and Hartman developed and refined many new lighting instruments. Individual sources were developed and used to light the acting areas from above the proscenium as well as from the auditorium.

The first dimmer boards

The earliest types of electrical dimmer board—a centralized control of the intensity of all of the lights—followed the conventions of the gaslight era. Groups of resistance-type dimmers were permanently wired to particular circuits in the footlights or borderlights. The only open or flexible dimmer circuits were the floor pockets (*i.e.*, electrical outlet boxes set in the floor of the stage) used to accommodate a wide variety of floodlights and special effects. Like the pipes and rubber tubing of the gas mains, all electrical circuits were wired solid. The location of the large banks of resistance dimmers in permanent installations of this kind was backstage in the wings and in close communication with the stage manager, who could see what was happening on stage. Later, however, even in large Broadway musicals, the array of portable switchboards may be (and is sometimes) relegated to the cellar or some other place with a scant view of the action on stage.

With the inception of individual spotlights for controlled-area illumination, in contrast to the earlier banks of footlights and borderlights, there arose a need for a flexible patch panel, in which temporary connections of such lights could be easily made and controlled. A group of spotlights could now be placed for the most effective direction, colour, and overall intensity. It became possible, for example, to place several spotlights together to achieve the realistic effect of an actor lighting an oil lamp, upstage right, without flooding the entire set with light.

The first electricity that was commercially available was direct current. For arc lights and resistance dimmers, today or 50 years ago, direct current presents no problems, and many theatres still draw their main supply of electricity from direct current. Most up-to-date theatres have the benefit of alternating current, which allows the use of more advanced dimmers.

Use of colour in stage lighting

The use of colour in stage lighting adds immeasurably to the realistic and atmospheric mood of a setting. Through the use of contrasting colours, the power of light to reinforce plastic qualities may be enhanced, and accents of appropriate intensity may be obtained. Coloured glass is the oldest medium of coloured light still in use in the theatre. As glass does not fade and comes in a variety of pure colours as well as some tints, it may be used in permanent installations for footlights or borderlights. Glass is available in cast roundels or in sheets; in sheet form, it is cut into narrow strips and mounted in holders to protect it as it expands and contracts with changes in heat. On the negative side, glass is heavy, fragile, and expensive, and its colour range is limited.

Gelatin is an expendable colour medium made of purified animal matter that is dyed and dried in sheets. It comes in a wide range of colours, it is cheap, and its light-transmission properties are excellent. Gelatin, however, is affected by moisture, and it deteriorates rapidly under heat. Plastic colour mediums are now available in the same range of colours as gelatin and have the advantages of durability, resistance to water, and strength. Although plastic costs more than gelatin, it lasts about five times longer. Later plastic colour mediums have been made of Mylar, which is virtually indestructible and more resistant to heat than sheet glass, which is much more expensive.

LIGHT SOURCES

Illumination sources. The earliest incandescent spotlight used a rather heavy plano-convex lens. The socket assembly with attached reflector moved forward and backward in the housing in relation to the lens and widened or narrowed the spread of the beam of light. As higher wattage lamps appeared, larger-diameter lenses were required, and soon they grew too bulky. Larger and more practical spotlights were made possible by the adaptation of the Fresnel, or stepped, lens, a type of lens invented in the 1820s, consisting of thin concentric sections that allow much more efficient use of a light source and, therefore, much smaller lamps. A slightly pebbled effect on the rear of the lens helped give an even, soft light with soft edges. Fresnel spotlights became available in wattages as low as 50 or, with a 24-inch (61-centimetre) lens, as high as 10,000.

The ellipsoidal reflector spotlight

The refinement of the coiled filament made possible the development of the tubular lamp, which in turn made possible a second type of spotlight, the ellipsoidal reflector spotlight. This type of spotlight uses an ellipsoidal reflector to gather the light at the focal point of the reflector. In front of the reflector, the beam of light is shaped by means of adaptable shutters, an iris, or a predetermined pattern. One or two lenses are used as an objective to project the image. The field of the beam is of even intensity, and a hard or soft edge can be obtained by focus adjustment. The focal lengths of the lenses can be changed if necessary to compensate for the length of throw or to widen or narrow the field. Special wide-angled lenses are useful in projecting patterns, such as leaves, at close distance. Available in a range from 250 watts to 3,000 watts, the ellipsoidal spotlight is a highly efficient and versatile light source. The larger ellipsoidal spotlight can also be fitted with coloured quartz patterns of any design, from realistic clouds to abstract rotating patterns. Used in combinations at different intensities, a cyclorama can be flooded with light in either bold or subtle patterns.

Borderlights and striplights are available with wiring for control in three circuits. Older types had reflectors that produced a soft, even blend; later versions use a range of highly efficient sealed-beam lamps, such as the quartz halogen sealed-beam lamps, which are available in a variety of wattages and reflectors—from flood to narrow spot. Mounted in striplights, the sealed-beam lamps have become a versatile and effective source, either for lighting large surfaces or as a strong backlight wash. Television and motion-picture lighting has rapidly developed such sources, which were initiated in the theatre.

Use of floodlights

A floodlight is a refined version of the old bunchlight—a half dozen old filament lamps grouped into a square housing, usually painted white on the inside to serve as a reflector. Later, a single large-wattage lamp replaced such multiple lamps. Floodlights today are usually ellipsoidal reflectors in a wide or narrow beam. The reflectors are mat-finished, accommodate a variety of lamp sizes, and do not require a lens. The quality of light is soft and diffused; it is ideal for a wash of light over a drop or a cyclorama. Usually, more than one circuit is used to achieve a variety of colour changes. Large banks of fluorescent lamps with colour filters are effective in lighting a large and very high cyclorama, but the difficulty in dimming fluorescent sources has limited their use.

The parabolic "beam" projector was developed for outdoor lighting and slightly modified for stage lighting. The

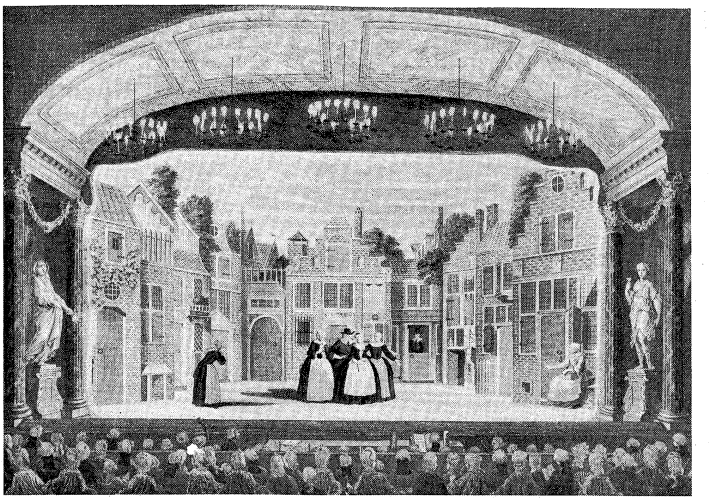
familiar carbon arc searchlight, which throws an intense narrow beam of light several miles into the night sky, is a simple parabolic unit without a lens. Such a searchlight consists of a strong pinpoint source (the arc) and a large parabolic reflector of silvered glass. The same simple optical principles were adapted to incandescent lamps for floodlighting the exteriors of buildings. The output is highly efficient, but to reinforce it and control it a lens can be added in front of the source. Used in a phalanx as far offstage as possible, the beam projector can serve as a strong source to give the effect of long rays of sunlight streaking across the stage. A most significant advance was made by hanging beam projectors so that they provided backlighting; this resulted, for the first time, in the actors' being strongly separated from the background, creating an illusion of depth and a haze.

Projections and special effects. A significant amount of lighting equipment has been developed for "special effects." Standard effects include moving clouds, water ripple, fire effects, rain or snow, painted rainbows, and fireworks. For practicality, most special effects are built around a standard spotlight housing. The effect head, containing a painted or photographic transparent disk and the mechanism for revolving it, is placed in front of the spotlight housing. An additional objective lens is used to magnify and focus the image.

The oldest effect projector, dating from the World War I era, is the Linnebach lantern, often called a "scene" projector. It is simple both in principle and in construction. A concentrated light source is placed in a deep black box, and a painted slide is placed on the side of the box left open; since light travels in straight lines, the design painted on the glass is thus projected against a drop onstage, greatly enlarged, at a relatively short distance. Since no lens is used in a Linnebach lantern, the light source must be powerful and concentrated. The design must be simple and bold, for any line narrower than the point source itself will be lost. The overall effect is stylized and borders on the abstract. Rear projection with at least two projectors is required for any ambitious production. Large incandescent lights replaced the original carbon arcs in the Linnebach lantern.

About mid-20th century there was renewed interest in the use of projections. Fortunately, the development of new projection equipment provided a powerful instrument to produce effects not possible before. In the post-World War II years at the music festivals at Bayreuth, Germany, Richard Wagner's grandson Wieland reduced three-dimensional scenic elements to the barest essentials and then flooded the stage with multiple, overlapping projected patterns. In later years, more scenic elements were added to give variety of texture and depth to the flow of light and pattern. Still later at the Festspielhaus in Salzburg, Austria, the productions of Wagner's music dramas designed by Gunther Schneider-Siemssen elaborated this concept to achieve even more dramatic and sumptuous productions; he filled the vast, extra-wide stage with patterns of light in depth, softened with scrims (loosely woven meshes that diffuse the light) and translucent drops (backdrops with sections dyed to transmit some light). The Czechoslovakian designer Josef Svoboda did more than any other contemporary designer with "visions in space." For some productions, Svoboda used a direct, journalistic approach, massing three-dimensional screens to create a montage effect with slides and film. Polyvision, a production conceived and executed by Svoboda for the Czech pavilion at the 1967 international exhibition at Montreal, was a brilliant multimedia experience. In his other productions, which were equally stylized but more indirect and abstract, he used alternating surfaces of scrim and scenic elements to catch the patterned light, cast complex shadows, and float in depth before a seemingly infinite background. The concept is reinforced by a constant play of changing light upon the performers. No matter how flexible the technique of projection may be, however, it does not take the place of conventional scenery, such as a door or a flight of steps. In Svoboda's productions these diverse elements are blended together with light.

Svoboda's productions



"Street Setting, Schouwburg," hand-coloured etching, *c.* 1765, of the scenic design, showing an 18th-century theatre lit by candles.
By courtesy of the Toneel-Museum, Amsterdam

Innovative contributions to lighting and the use of projections were also made in dance. The American Alwin Nikolais made very original use of dancers, costumes, light, and projections to form moving geometric and abstract designs. At times, the moving bodies of the dancers become the screen for the projections. The Robert Joffrey Ballet, which is also based in New York City, in its production *Astarte*, created a unique combination of film and slides on a moving, pulsating screen. The screen is a "living" element that repeats the images of the dancers as it changes contour with the development of the dance. The screen moves, strains, relaxes, contracts, and, in the end, slowly lifts to reveal the empty stage behind.

All such approaches to the theatre as a total concept combine the old and the new. The designer reaches out to try new materials, to create new textures and dimensions. The most versatile technique of blending the designer's concept into a "total environment" is the art of stage lighting.

LIGHTING CONTROLS

The dream of a practical system of remote control of lighting has always tantalized designers. Relays for remote switching have been commonplace for years. Motor-driven dimmers have proved practical for controlling auditorium lighting but for stage lighting are too cumbersome, expensive, and limited in speed control and are unpredictable in responding to changes. Banks of resistance and transformer switchboards were too bulky to be installed in a light booth at the rear of the auditorium with a clear view of the stage.

The first primitive but successful remote-controlled dimmer was used in 1890 but not perfected until later; it used a reactor coil operating on a relatively small amount of direct current to control a huge amount of alternating current.

Remote control became a reality largely as the result of the development by George Izenour of Yale University in 1948 of a dimmer using the thyratron, a type of electron tube. His concept brought continuous and instantaneous control to remote systems for the first time. In addition to its simplicity of operation, the development gave the lighting designer the facility for an "infinite" system of preset cues as an integral part of the system. It also became practical to install the control console at a vantage point in the auditorium so that the lighting designer (or operator) could see every nuance of lighting on stage.

Izenour's system of remote control

Meanwhile, transformer manufacturers developed a major breakthrough in design using new magnetic alloys. The result was a reactor with great efficiency and speed. This new reactor, combined with the silicon rectifier, another post-World War II development, resulted in a device called a magnetic amplifier—with no moving

parts or tubes, and capable of controlling large amounts of alternating current with no time lag.

Tube manufacturers, meanwhile, were furthering the development of the transistor, namely, the silicon-controlled rectifiers, or SCR, transistor-like devices which handle heavy currents and perform in the same manner as the thyratron tube, but with the advantage that they are compact in size, light in weight, low in cost, and provide instantaneous response, high efficiency, and long life. Control consoles using all three types of dimmers are in current use; future control boards will probably use dimmer units of the silicon-controlled rectifier type.

The design of lighting control boards is no longer a matter only of electrical engineering, in which the only concerns are engineering requirements or code specifications. Previously, the lighting designer had to make the most of whatever the engineers had given him. Now the remote lighting control console gives the lighting designer a flexible, artistic tool. Freed from the limitations of the mechanical controls and of the personnel who would operate them, the lighting artist can concentrate on the execution of his concept. No longer is there any reason for the mechanical limitations to interfere with the lighting.

One tool still remained to be developed—a memory bank to store all the cues of any number of productions. A console developed by George Izenour can hold ten mechanically preset cues and is practical and functional but nowhere near big enough. Various systems of punch cards, readout cards, tapes, and memory cores have been proposed to meet this problem, but the designs have been overcomplicated. Moreover, the cues were inaccessible, so that they could not be checked until they either were executed or failed to be executed. With the older ten-preset system, the operator could always turn to the particular dimmer of the preset in question and check the reading. He could see the actual value of the dimmer reading as well as every dimmer value in the entire preset. The great drawback of all subsequent systems was the lack of total visual readout of all circuits. A total visual display of the dimmer values in each cue (whether the cue is in actual operation or in a preview mode) is a practical necessity.

A television studio installation suggested one possibility for future stage lighting systems. Incorporated into an existing electronic switchboard, a small general-purpose computer functioned with remarkable success as the basic memory system. The console includes a large television monitor, and the cues are annotated with a "light pencil." Four fader handles (*i.e.*, devices controlling fade-ins and fade-outs) control any sequence of overlapping cues. These handles can be interfaced with any electronic dimming equipment, so that the need for an expansive array of control levers is eliminated. Any number of display monitors can be installed at strategic locations to monitor cues. Its unique value is in its constant display of cues. During a cross fade between two cues, every value change is indicated. Though cues normally follow in sequence, additional cues can be inserted or cues can be cancelled. The operator can jump forward or backward at his discretion. An agile operator can run through 70 cues in 15 seconds. A common application might be a sequence of cues timed to the beat of music.

A generous number of cues is available for storage and recall in this installation. Assuming that its lighting console has 100 dimmers, the memory bank of the computer is able to store from 400 to 2,400 cues on disk, or 5,000 cues on tape. For library storage purposes, an electric typewriter will provide printed copy of the cue sheets directly from the computer. Cartridge tape provides for storage on a season-to-season basis.

Since the heart of the system is a small, general-purpose computer about the size of a large suitcase, it can easily be moved with the rest of the production. The flexibility and adaptability of the computer system are not confined to lighting. It can be programmed for other theatre operations, including automated systems for moving suspended scenery, stage wagons or turntables, sound systems, and special effects. Since small, general-purpose computers are commercially available and need not be built to order, their price range is not forbidding.

Portable dimming equipment

There has been a long-standing need to develop portable dimming equipment for travelling. The practice of using banks of resistance dimmers (housed in a wooden crate that resembles an upright piano) came into use in the 1920s and is now considered an obsolete burden, bulky, extremely heavy, and mechanically primitive. Television has capitalized on small, lightweight silicon-controlled rectifier dimmers and produced a unitized system of dimmer racks with a five-scene preset console. A rack with six 12,000-watt dimmers is small enough to fit through a narrow door, and the five-scene preset console comprises controls for 24 dimmer units. Several consoles can be joined together.

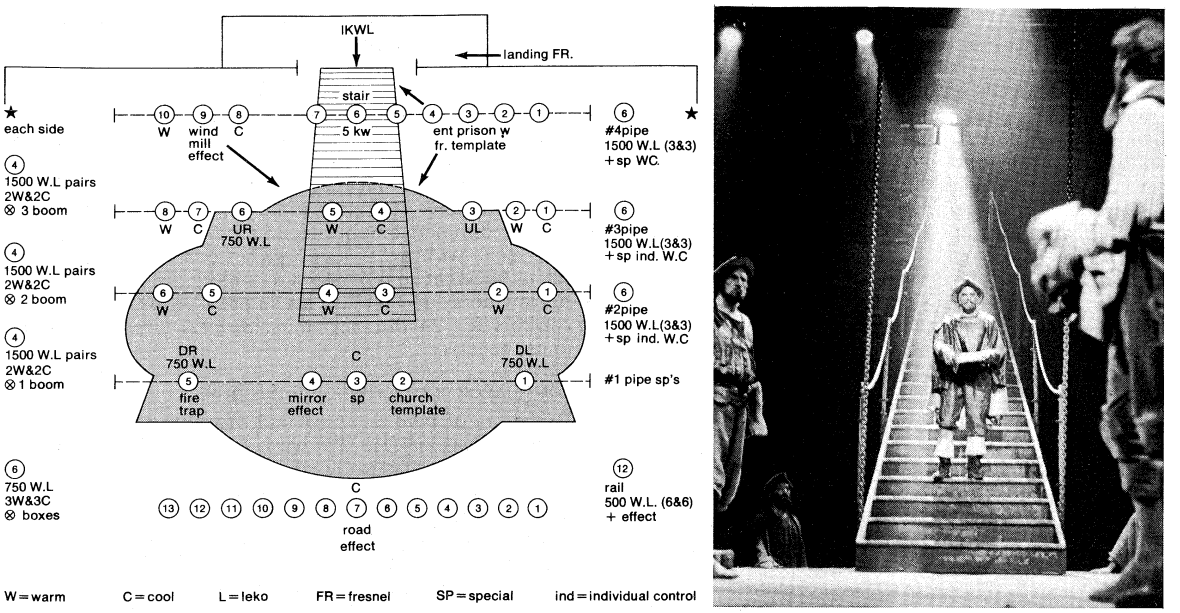
CONTEMPORARY LIGHTING TECHNIQUES

In the era of painted wings and drops, the standard illumination was an even, flat wash from footlights and borderlights. Architecture and furniture, including tables and chairs, were painted into the background with heavy shadows and exaggerated perspective. The actors had to resort to grotesquely heavy makeup to convey any facial expression; while their very presence (and every movement) was visually in conflict with the forced perspective of the painted scenery. It is, thus, understandable that the impressive effects of modern lighting made it seem almost limitless in its possibilities. Borrowing effects from Rembrandt and other acknowledged masters, the theatre designer began to "paint with light"—putting light where he wanted it and taking it away from where he did not want it. Thus, romantic effects were born, such as that of the lone figure in a shaft of searing light, isolated by darkness. The first requisite of stage lighting, however, is visibility. Good illumination provides plasticity and modelling of the actors and of the inanimate objects. Superior lighting is orchestrated to underline the development of the drama.

Importance of visibility

Many plays permit no approach other than realism, which must be achieved by suggestive means. Nature provides the model: the time of day is suggested by the predominant directionality of sunlight or moonlight, the depth and length of shadow cast, and the intensity of light. On a cloudy day, the overcast sky diffuses the direct sunlight and produces a soft, shadowless light of low intensity, and colder in colour. Intense sunlight and the attendant light of the bright sky together produce reflected light that diffuses or fills in daytime shadows, while the ambient light of the moon reflected from sky, trees, and buildings is too weak to wash out the shadows. So by means of the direction, diffusion, and intensity of light, as well as its colour, it is possible to suggest time, place, and season.

The means of suggesting natural lighting indoors are more arbitrary. The simulated sky or sunlight seen through a door or window is essential to indicate the time of day or night; in which case, the dramatist and designer have provided the necessary scenic element. To render the feeling of bright daylight flooding a room, the strong motivating light (*i.e.*, light from a source apparent to the audience, in this instance the sunlight) must be supplemented with additional light from other directions for adequate illumination. If only a sliver of sunlight creeps through a parted curtain in a dark room as the scene begins, the mood may be retained as the scene progresses as the illumination of important areas is slowly increased with specials. Artificial light indoors is easier to suggest because it more closely approaches the normal quality of stage lighting. Actual light fixtures are used on stage to suggest the sources of the light, and opaque shades can be used on some of these fixtures so that they cast actual patches of light against walls and furniture. The exteriors seen through doors and windows are darkened and different in colour from their appearance in a brightly sunlit scene. The walls fall off in shadow even though the general illumination is more diffused and even than in daylight. Light serves as a unifying medium for the stage composition. It is a mobile and changing accent that reinforces the action, sustains the mood, and focus-



Modern lighting.
(Left) Light plan for *Man of La Mancha.* (Right) Scene in the New York production showing the use of backlighting.
(Right) Alfred DeSio

ses the attention of the audience. Light and shade, brightness and contrast, define the size and shape of objects; but it is colour that creates emotional response, mood, and atmosphere.

Use of backlighting

The dynamic potential of lighting is best realized when the changes of light underscore and propel the development of the drama. Current techniques of stage lighting use arbitrary means to create realistic effects. The first real attempt at creating space and a real haze on stage (in a depth of 30 feet) came into being with the use of backlighting—light emanating from the rear of the stage rather than the front—which made the actors appear to be pulled away from the background. At first, parabolic beam projectors were used to achieve this effect, then ellipsoidal spotlights, and later, in some instances, sealed-beam parabolic reflector lamps. The ellipsoidal reflector spotlights afford the best control and the least spill. The beam projector may be deployed in a battery offstage to create the illusion of paths of sunlight or moonlight or a strong sidelight or crosslight. For strong crosslight, clusters of low-wattage units were supplanted by powerful ellipsoidal spotlights, hung low. With the addition of lamps attached to the auditorium's boxes, or box booms, besides the lamps already on the balcony rail or ceiling coves, modern lighting began to take shape. The open stage of the musical and dance furthered the development of effective lighting.

The square pattern of modern lighting

In bold outline, modern lighting can be described in terms of a square consisting of a lucid pattern of strong backlight, paths of strong crosslight, and the necessary fill light from the front from the boxes and first pipe (*i.e.*, lights hanging from a horizontal pipe just inside the proscenium) on stage, plus special lights as needed. Richer colours are used in the backlighting and crosslighting. Lights creating warm and cool effects are wired to circuits that allow them to be controlled right, centre, and left stage. The crosslighting coming from offstage and from the box booms is usually less saturated in colour. The lights on the balcony rail give a pale wash—passive illumination to produce visibility without destroying the sharp modelling of the backlighting and sidelighting. Special lights, in colour or clear, cut through particular areas from the box booms and the pipes on stage.

When box sets began to go out of style, and suggestive sets, skeletal fragments without ceiling and walls, were introduced, the lighting technique was already at hand in the layouts for musicals. There was need for more special lights, different circuiting, softer colours, more pattern projections, and more intricate, overlapping changes of light. The pattern of back, sides, and front, however, still holds; it is a clear-cut, three-dimensional approach, with distinct functions meshing into a controlled pattern.

For the most part, ellipsoidal reflector spotlights are used to cover acting areas from all locations. Fresnel spotights are used for area washes or to accent plastic scenery elements. The appearance of scenery as it is illuminated by the lighting intended for the acting areas is totally unpredictable; it is necesary to provide separate blending or accent lighting for most plastic scenery. Patterned projections, such as leaves or abstract patterns, may be used in wide-angled, ellipsoidal spotlights to project on scenery that is solid or even made of gauze.

The three-dimensional approach to lighting, developed within the confines of the proscenium stage; it is also valid for the thrust stage or arena theatre. For theatre in the round, the predominant angles for lighting will tend to be sharper and also to be steeper, to avoid spill into the audience area that would disturb the spectators. The ellipsoidal reflector spotlight is used for all acting areas, and an added high hat (a hollow tube that extends out 6 to 12 inches) will confine unwanted spill as well as hide the lens and colour frame.

In simple language, lighting is a matter of hanging the right instrument in the right place and focussing on the right target. Stage lighting cannot be mastered except through the observation and experience that are necessary to learn the practical and aesthetic functions of light. It is essential to know the physical characteristics of different light sources and how and when to use them to best advantage. It is also mandatory to learn the complexities of control and how to build lighting changes. Stage lighting must be more than good illumination. Light must be used as a subtle flux, exerting a subconscious influence that unifies all the media of the theatre.

The creative concept formed for lighting a production requires that the essential qualities of the play be understood and absorbed. The theme or main line of the script may suggest an overriding motif: enervating heat, damp clamminess, dappled sunlight, cold penetrating north light, a feeling of being in limbo or underwater. The final choice must satisfy the particular qualities of the production and the concepts of the playwright, director, and actors, as well as the designer. Attending run-throughs in the final period of rehearsals will reinforce previous impressions and suggest refinements in the rhythmic changes of light required. The actors' performances will suggest lighting changes that can enhance the emotional range of the total performance. (Ra.Ho.)

IV. Stage machinery

In the modern theatre, a variety of mechanisms are used to erect, position, and manipulate scenery. These devices fall into three categories: hoists, lifts, and horizontal drives.

Use of hoists

Hoists are used for flying (*i.e.*, raising or lowering) scenery up to or down from the stage penthouse; lifts are used to bring up scenic elements from beneath the stage floor; and horizontal drives are used to slide platforms, flats, and large properties from the wings onto the visible part of the stage or from one area of the stage to another.

Hoists are the most common type of mechanism in the theatre, and there are three types: counterbalance, electrical, and fluid drive. The simplest method of hoisting is by counterbalancing, and there are two types, the "hemp system" and the "iron-counterweight system." In the hemp system, sandbags and iron ingots are used as a counterbalance. Four to six lines are attached to a batten; each line passes over its associated "loft block" and then goes to the master head-block sheave. There are sandbags of varying weight used to counterbalance the scenery. The iron-counterweight system is similar except that the lines are made of wire rope and the operating rope is not directly attached to the hoisting line. Instead, it is a fixed system, held between the master head-block sheave and a fixed-tension sheave. For counterweighting, there is an arbor or carriage, in which the iron weights are placed. Finally, there is a lock-rail combination to insure the safe use of the counterweights.

In an electrically powered stage hoist system, electricity may provide only the pulling power, while still requiring counterbalancing (electrical counterweight-assisted hoists), or it may provide the lifting ability and the force needed to counterbalance (pure-power hoists). There are two types of electrical counterweight-assisted hoists, traction drive and linkage drive. In the traction-drive system, the hoisting line-counterweight system is not directly coupled to the electric motor drive, and slippage develops during acceleration and deceleration of the payload. It varies according to the weight of the payload and the velocity developed. Consequently, traction-drive hoists are utilized only when a relatively constant weight is lifted at a constant velocity, such as the asbestos curtains, light bridges, and orchestra shells.

The linkage drive is similar to the traction-drive hoist, except that the hoisting lines are attached directly to the motor.

The second category of electric hoist is the pure-power type, in contrast to the counterweight-assisted type. It is composed of a motor (usually placed on a gridiron), a brake, a gear reducer, and a drum around which several hoisting lines wind. At the Metropolitan Opera House in New York, there are 110 motors linked into a single console that can be run by a single operator.

The third type of hoist is the fluid-drive system, in which an electric power unit is used to run a fluid piston, which in turn moves the hoisting lines. There are several advantages to this system: the electric generator does not have to be physically near the fluid drive, so the system is virtually noiseless; and the operator may divide the power between any number of pistons, a feat not possible with an electric motor alone.

Use of lifts

Lifts are used in the theatre to move platforms, actors, scenery, and other production elements above or below the stage floor. In contrast to the hoist, which is supported by the "over stage" structure, the lift is supported and guided by the stage floor or the cellar floor below the stage. The two general types of lifts are the architecturally integrated and the temporary production apparatus.

Architecturally integrated lifts must be designed in accordance with local building-safety codes: in New York, for instance, only two types can be built into a theatre. One of these, the direct-plunger lift, is hydraulically driven, using a piston attached to a portion of the stage floor. The piston operates under hydraulic pressure (like the pure-power hoist) and is expanded and collapsed to elevate the associated platform. The other type, the screw-actuated lift, is either electrically or hydraulically driven and coupled to a vertical screw through a nut in which the upper end of the screw is connected to a portion of the stage floor.

The layout and installation of permanent architecturally integrated lifts must be carefully determined; the frequency of use and the type of events to be accommodated must be envisioned, along with the attendant inflexibility of the permanent mechanization of the stage floor. In general, architecturally integrated lifts are successfully employed where they are designed to meet the specific production needs of a permanent resident company.

Use of horizontal drives

Horizontal drives are used to rotate, propel, and project scenery, actors, properties, etc., from the offstage areas on to the acting area. In an effort to meet the audience demand for rapidly changing scenery, designers often utilize a system of horizontally moving platforms. Although the articulation of horizontal motion on the stage is unlimited, there are several established configurations that are easily identifiable. These include the wagon, in which scenery is built on a low platform mounted on casters so that it can be quickly rolled onstage and offstage; the jacknife stage, which is similar to the wagon except that it is anchored at one corner from which it pivots onstage and offstage; and the revolve or turntable, in which several settings are built on a huge circular platform, or turntable, which is revolved so that only the appropriate setting may be seen through the proscenium. In each of these, the scenery may be changed when the unit is offstage and then rolled back on. Treadmills and revolving segments are also used. (W.Cr.)

V. Theatrical costume design

Theatrical costumes are neither historical nor contemporary; they elaborate a style and convention of their own. Their success in any production depends in large measure on the designer's ability to understand and interpret the director's production concept. The spirit of the production and the director's aims may call for special emphasis on certain characteristics in clothing. Designs may be copies of period costumes or may attempt only to evoke a period. Whether real or illusory or a subtle balance between the two, research is an important start to the designer's work. Museums and other repositories of the decorative and literary arts are rich sources of inspiration. Since the subject of theatrical costume design is so closely related to the art and history of dress, the reader's attention is called to the article DRESS DESIGN AND BODY DECORATION, in which these topics are treated comprehensively.

HISTORICAL DEVELOPMENT

Classical theatrical costume. Theatrical costumes were an innovation of Thespis, in Greece in the 6th century BC, and theatrical costumes are still called "the robes of Thespis." Athenians spent lavishly on the production and costumes at the annual drama contests since each poet was given a wealthy citizen, the *chorēgos*, who, encouraged by the honour of a separate state impresario's prize, tended to make the event a demonstration of his spending power.

The earliest tragedies were played in long, rich robes similar to those worn by the priests of Dionysus. To increase the height and importance of the principal actors, Aeschylus introduced the buskin, Asiatic elevated boots, called *kothornoi* in Greek, which became one of the chief characteristics of the Greek tragic actor. The soles increased in thickness according to the status of the wearer. To balance his height, padding often was used to add bulk to the actor's stature. The chorus did not wear these boots, so that they could dance with ease. The performers were clad in stage tunics, *chitōn*s, which were longsleeved, high-girdled, and elaborately embellished, as were their long and short cloaks (*himatia* and *chlamydes*). Aeschylus was renowned for the brilliant mounting and costuming of his tragedies, and by the time of his death, in 456 BC, a traditional tragic costume had evolved. Each costume was in a symbolic color.

Greek masks

The most important feature of the Greek costume was the mask, which indicated an actor's age, sex, station, and customary mood. The masks were made of linen, cork, or

wood and were skillfully carved and painted. Their funnel-shaped mouths acted as megaphones to amplify the voice. In his *Onomasticon*, Julius Pollux, a Greek writer of the 2nd century AD, gives a detailed account of the special features given to each character. He enumerates 30 masks used in tragedy and 17 female masks and lists the characteristics of the comedy series, which are particularly exaggerated and grotesque. Crowning the mask and adding more inches and importance to its wearer sat the *onkos*, a high ornate headdress.

By courtesy of the Hermitage, Leningrad



Scene from a burlesque showing Heracles tempting Apollo. Actors wear short tunics and tubular pants of Greek Old Comedy. Detail from a phlyax vase in the Hermitage, Leningrad.

In the Doric mimes and Old Comedies, the upper class characters wore stage *chitōn*s and cloaks and the lower classes and slaves wore short tunics, revealing pendant phalli. These character tunics were often worn under light-fitting vests and over grotesque padding of torso and buttocks.

Mimic horses, satyrs, bird figures, and other animal imitations were much in evidence. Aristophanes, in *The Wasps*, *The Birds*, and *The Frogs*, calls for all manner of such figures and clothing. Actors performed in skins and wore horses' heads, birdlike visors, and mock wings.

In the later comedy of Menander, the phallus and mythological elements were abandoned, for his intention was to represent urban life, and the costumes worn reflected this intention. Masks became more stereotyped; they were used over and over again for character parts in different plays. The symbolic colour for characters still held great importance.

Roman drama had its indigenous roots in the Etruscan mimetic dances. Mime without masks, together with acrobatic dancing, became a favourite part of the day's entertainment at the games or circuses. Plays, too, were among the diversions; to satisfy the crowd's taste for realism and sensation, legions of spectacularly dressed soldiers were introduced to the tragedies. Costumes for tragedy were modelled on Greek styles; by Roman times the name *cothurnus* had come to designate the tragic genre itself. Kings and queens in tragedies wore appropriate padding, tall wigs, and sleeved *syrma* (the robe corresponding to the *chitōn*). Bands of bright hues decorated the costumes of happy characters, and gray, green, or blue those of fugitives. Gods and goddesses were distinguished by their insignia; seers were clad in woolen garments over a shorter, less full *syrma;* and huntsmen rolled a purple shawl around their left arms. Slaves wore leather jackets and light breeches or braids indicating their barbarian origin. In the *fabula palliata*, Roman comedies on Greek subjects and based on Greek models, actors wore *chitōn*s and the *pallium*, a cloak resembling the *himation*. In tales of Rome based on Greek models, and *fabula togata*, actors were costumed in the mantle and toga. The heroes of plays dealing with Roman history, called *fabulae praetextatae*, wore togas with the *praetexta* decoration indicating magistrates.

The *fabula Atellana*, rustic plays originating in southern Italy, were given both public and private performances. In the homes of patricians, young noblemen often took the parts of the standard characters—the bragging, greedy Maccus, the stupid Bucco, the foolish miser Pappus—wearing masks to disguise their identities. Their masks, known as *personae*, were decorated with hair; they were not worn by professional actors until introduced by the actor Quintus Roscius in the 1st century BC. Before their introduction, wigs (*galeri*) were worn. Also, Italian comedians wore a flat slipper called the *succus*.

Roman mime costume

The garment peculiar to mime productions was a gay coloured-patchwork jacket called *centunculus*, and this name became applied to mime actors in general. At Corneto, a tomb painting depicts Stupidus (the fool of the mimes) wearing another coat covered with coloured patches, a tall pointed cap decorated with a tassel, and a square, short cloak known as a *ricinium*. Because of this garment, the mime players were also known as *riciniati*. The producers of the mimes also provided their performers with gigantic phalli. "The mimes," it was said, "pulled down the gods and heroes from on high and provided them with phalli."

As drama in Rome declined, the mimes introduced plays on the theme of the oldish cuckold and the frail and plotting lover, which became most bawdy and obscene during the 6th century AD, as did the pantomimes featuring lavish licentious entertainments. Such entertainment continued to be presented until the fall of the Roman Empire later in that century. Through the Dark Ages, jugglers, fools, and small mime troupes roamed through Europe, but a record of organized drama has yet to be found.

Medieval costume. Originally, mystery plays were performed before the altar, with the actors, priests, and clerics wearing church vestments. The miracle plays, which retold incidents in the lives of saints, were also originally played by clerics and actors. Inventories were kept of garments made and bought, and these lists indicate that Adam and Eve were clad in close-fitting white leather, God in bishop's robes, and Jesus in a simple white robe. Cain, Abel, Joseph, Lazarus, and other biblical characters would appear in contemporary clothes, hoods, pourpoints or doublets, or loose gowns. A female character was indicated by the simple addition of a kerchief on the head. During the 12th century, when the performances moved outside the church, priestly vestments were still the main costumes.

When European craft guilds became responsible for the mounting and dressing of these productions, their scenic plans became ornate and ambitious and the early simplicity was lost. The exotic robes, angels with gilded limbs, halos, and ornate wings may be seen in the paintings of the 15th-century Flemish masters Lucas van Leyden, Memling, and Van Eyck. Satan and his devils enjoyed great popularity with the large audiences. Their grotesque masks, lashing tails, fangs, and snouts in lurid blacks, reds, and blues are well recorded by artists of the period. There is a large collection of devil masks preserved at the Tiroler Landesmuseum Ferdinandeum, Innsbruck, Austria.

In the morality plays, concerned with allegorical subjects, that became popular toward the end of the 14th century, costumes personified the virtues and vices, life and death, and similar abstractions. The Bible stories and morality dramas were also taken through the streets on two-story pageant carts; these processions of gorgeously dressed groups and tableaux can be seen in 15th-century paintings such as Piero di Cosimo's "Triumph of Theseus and Ariadne" and Botticelli's "Triumph of Chastity."

Renaissance costume. From the early 14th century, in addition to religious performances, secular theatre also developed in the form of popular celebrations, spectacles, royal and ambassadorial entrances, weddings, and other lavish entertainments that surround the royal courts. A monarch entering a town or dominion would do so with a large entourage in full pomp and heraldic dress. These entrances included a series of stops at stages placed at various intervals en route. *Tableaux vivant* and mimes

Costumes of royal celebrations

were performed in costumes similar to those worn in the mystery and morality plays. With the gradual decline of church power and the revival of classic ideas, Renaissance designers found their inspiration in the myths and legends of Greece and Rome.

Countries and designers competed in the sumptuousness of their displays. The progresses in England, *entrees* in France, *trionfi* in Italy were all based on the triumphal processions of the ancient world. The monarch or emperor was glorified as the hero, and his entourage and vassals appeared in semiclassical grandeur. There were floats of allegorical figures and legions of splendour such as Caesar had never dreamed.

The Italian designers, who included artists as eminent as Leonardo da Vinci, led the field. Costume plates of specific festivals indicate the awe-inspiring splendour of these occasions.

At court, the tragic, comic, and satiric dramas of classical theatre were revived, with the addition of entertaining entr'actes. From these descended the intermezzi—pantomimes that included song and dance. Again, the designers wished to appeal to the eye of the Renaissance public. The designer in charge of theatrical productions for the court of Mantua in the mid-16th century explained in his treatise on stagecraft that the Italian object was to captivate and charm the audience:

> I tell you especially that I make efforts to dress the actors always in as noble a fashion as is possible for me, but in such a manner that there is a sense of proportion among them, in view of the fact that the rich costume . . . particularly in these times when pomp is at its highest peak, adds much reputation and beauty to comedies, and even more to tragedies. I would not hesitate to dress a servant in velvet or colored satin, as long as his master's costume were embroidered or decorated with gold, so rich that there would be maintained the proper proportion between them. . . . I do my utmost to dress the actors very differently from one another, and this is of great help, both in adding beauty and in facilitating the understanding of the plot.

In 1589 the sophisticated Florentine court produced an intermezzi called "Harmony of the Spheres," a spectacular type of masque, with music, that was the immediate forerunner of opera. Etchings of the Grand Ducal fetes in Florence of 1606, 1608, 1615, and 1616 show groups of dancers in towering plumed classical helmets, Roman costume, and cuirass (body armor) worn over doublets, with Roman labels hung from the cuirass, over breeches and hose, to cover the thighs. The Renaissance developed these labels into an exaggerated ornate skirt and named it the *tonneler*. Each item of costume was decorated with a profusion of curved ornaments, flowers, vines, and animal and human forms, suggestive of Baroque design.

Strongly influenced by these Florentine specialists during a visit to Italy, Inigo Jones transformed English court masques and entertainments in the early part of the 17th century. The Duke of Devonshire's archives at Chatsworth preserve his designs. Through him, English designs followed the Italian pattern: breastplates molded to the body, plumes, helmets, and various Roman kilts mixed with modified elements of contemporary dress. Ladies' costumes followed the dress of the period more closely but strove for a looseness and for the transparency of fabric typical of classical dress. Decolletage and low-cut bodices were much in evidence, with bosoms often veiled with gauze. The cost of these costumes was borne by the wearers, the lords and ladies of the court. There seemed at times no limit to their embroidered, bejewelled luxury and fine headpieces.

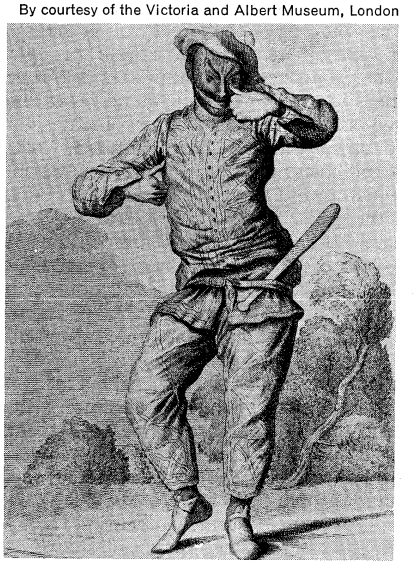
Besides court entertainments, literary noblemen maintained companies of professional actors to perform the new plays and classical revivals.

Early English costumes

Little pictorial evidence of the first public theatres in England survives. It is known, however, that the best part of the actors' wardrobes were gifts from wealthy patrons. One drawing of Shakespeare's *Titus Andronicus* dated 1595 shows contemporary costume mixed with *costume a la romaine* similar in design to the courtly work of Inigo Jones. Mixtures of styles and periods were accepted by the audiences. Certain famous characters such as Shakespeare's Sir John Falstaff became identified with a particular type of dress; a sketch that Inigo Jones made of Sir William Davenant in *The Temple of Love*, produced in 1635, was inscribed: "a roabe of russet Girt low w^{t} a great belley . . . the sleeues shorte . . . buskines to shew a great swolen legge . . . a great head and balde. Like a Sr Jon fall staff."

In Italy, the commedia dell'arte presented a multiplicity of character and types. Their fine costumes evolved from garments representing different Italian provinces and rustic personalities and gradually became stylized, subject to the dictates of changing taste. To make the costumes decorative or funny there were exaggerations in form and colour. The characters of the lovers appeared in contemporary street clothes, and comic personages in curious or grotesque garb with leather half-masks. They were ideal theatre clothes for they identified the characters, allowed complete freedom for movement and acrobatic dancing, and charmed the spectator with their amusing devices, patterns, and colour motifs. A typical character, Arlecchino, originally dressed in shreds and patches, emerged by the 17th century in a suit of red, blue, and green triangles arranged symmetrically and joined together with yellow braid. One hundred years later, the triangles became diamonds, and his soft cap was exchanged for a pointed one. Now as Harlequin he still appears as a central character in English pantomime, carrying his original wooden sword and wearing his black mask. A fine pictorial record of the commedia characters may be seen in the works of the 17th-century French artists Jacques Callot and Antoine Watteau.

By courtesy of the Victoria and Albert Museum, London



Typical costume worn by players in commedia dell'arte. Anonymous German engraving of Arlecchino, early 18th century. In the Victoria and Albert Museum, London.

The commedia dell'arte toured and performed throughout Europe and was particularly popular at the French court. Molière's company also toured France in improvised farces based on the style of the commedia dell'arte. Actors in his comedies, such as *Le Tartuffe*, played in contemporary dress, but his company also performed with the King and court in brilliantly dressed spectaculars at Versailles. Texts were by Molière, music by Jean-Baptiste Lully, and costumes and decor by Jean Berain.

The introduction of actresses

The Restoration of the English monarchy in 1660 led to the opening of theatres again, and the great innovation was the introduction of actresses. Talented, confident, and flamboyant ladies replaced the Elizabethan boy actors. Records indicate that "splendidly clothed" persons lent costumes, and wardrobe keepers now aided players in selecting from available stock. The ladies chose the gown most becoming and elaborate. No attempt was made at character dressing or historical accuracy.

Costume in Baroque opera and ballet. A galaxy of specialists joined the courts of Louis XIII and Louis XIV.



(Left) Jean Berain's design for the costume of a female dancer with castanets, 1685. In the Victoria and Albert Museum, London. (Right) Costume design for a French peasant girl by Jean-Baptista Martin, early 18th century. In the Bibliothèque de l'Arsenal, Paris.
By courtesy of (left) Victoria and Albert Museum, London, (right) Bibliotheque de l'Arsenal, Paris

with Giacomo Torelli, the chief originator of Baroque theatre, arriving from Venice. This monumental scenic style included magnificent colonnades, temples, palaces, and scenes so lavish that only the great ruling houses could afford them. Designers such as Torelli brought great prestige to their patrons. An outburst of Baroque opulence bore witness to the power and splendour of the Sun King. In France in the early 17th century, the designer Daniel Rabel worked inventively, producing many witty and droll effects and costumes of grotesque conception. Burlesque costume had found its way to amuse the court.

Henri Gissey and the great Jean Berain were attached to the Royal Cabinet of Louis XIV. Gissey is most famous for his celebrated Carrousel (1662), a horse spectacular never since surpassed in its magnificence—500 noblemen in plumed regalia escorted by a greater number of elaborately dressed attendants. Costumes represented different nations, each having a particular colour scheme: the Romans were in red and gold led by the King; the Turks in black and yellow; the Indians in white and gold; the Americans in green and gold. The marvellous scene played in front of the Louvre was indeed the grandest fantasy in costuming, with both horses and riders decorated in full Baroque trappings.

Chroniclers record gorgeous costumes appearing in ballet, opera, and dramatic offerings: gold brocades covered with lace, diamonds, emeralds, ribbons, and immense trains graced the stages of France. The French surpassed the Italians in their opera costumes, which were richer than any elsewhere. In the designs of Jean Berain there is a homogeneity of style that is the mark of a great master. In 1680 women first appeared on the operatic stage. Berain's costumes for them were serious and noble in style, based on contemporary court dress. Their tight bodices and flared basque shirts had decorative overskirts with trains; they followed an almost uniform silhouette. The headdress, also carefully designed, was a formal arrangement of feathers or lace. Male garments were in the Roman style, the tunic fitting tightly to the chest, worn with *tonneler* (short skirt), high boots, and cape. To avoid monotony, Berain devised an amazing variety of ingenious trimmings—embroideries, fringes, slashings, puffs and paddings, inlaid ornamental motives, pastes, and semiprecious stones to punctuate every small decorative panel and part. Scaly bodies, bejewelled ones, serrated leaves —every elegant or fantastic theatrical device was used to create a sophisticated ensemble. Berain coordinated all aspects of scenic decor in an equally industrious way and as a decorative artist dominated French and European taste and costume design.

Berain's costumes

After Berain died in 1711, the sumptuous Baroquerie of the early years of Louis XIV declined. Berain's son Jean succeeded his father as court designer and carefully documented his father's drawings, which are now kept at Musée de L'Opéra and at the Victoria and Albert Museum in London.

By 1721 Claude Gillot's designs for the ballet *Les Elements* showed a great change in taste. Heavy fabrics and embroideries used by Berain were replaced by lighter, more delicate weights and appliques. Ladies' costume, following the caprices of the contemporary modes, included a panier. Peasant and rustic characters began to appear and were most popular with the court, in be-ribboned garments of pastel satins and silks. Extensive records of these costumes exist in the form of engravings.

Costume of the 18th and 19th centuries. Jean-Baptiste Martin, who was appointed designer for the Paris Opera in 1748, devised decorative and amusing Rococo variations for the male dancer's traditional costume. Incas, Africans, Chinese, and Mexicans graced his ballets and under his direction the *tonneler* took on an elliptical shape.

All the elegance and sophistication seen in Rococo court circles of Louis XV were brought to the stage by Martin's successor, Louis-René Boquet. His designs were theatricalized versions of the new fashionable silhouette. Boquet clothes were delicate, artificial, and pale in tone, trimmed with garlands and Rococo finery. All Europe imitated the French ideas, the English and German facsimiles without Boquet's innate good taste.

Denis Diderot wrote in 1758 of the current fashions in the French theatre:

> Ostentation spoils everything. . . . Wealth has too many caprices: it can dazzle the eye, but not touch the heart. Beneath the garment that is overloaded with gilding, I never see more than a rich man, and it is a *man* I look for. . . .
>
> Comedy ought to be played in informal dress. On the stage it is not necessary to be either more or less dressed up than one is at home. . . .
>
> The more serious the play, the more austerity there must be in costume. . . .
>
> What expense were not our actors put to for the production of *l'Orphelin de la Chine*? How much has it not cost them to rob this work of part of its effect?

Historical authenticity of 18th-century French costumes

In 1755 the great French actors Henri-Louis Lekain and Mlle Clairon, who were also bent on more simplicity and historical accuracy, dressed Voltaire's play *L'Orphelin de la Chine* without the hoops or overskirts of currently fashionable dress, while the men appeared dressed as Tartars and Chinamen. In 1789 the French tragedian François-Joseph Talma provoked a scandal by appearing in Voltaire's *Brutus* in a severely simple toga appropriate to the ancient Roman setting of the play.

In England David Garrick abandoned heroic Roman garments in favour of realistic, contemporary dress, which he designed himself to insure the simplicity he desired. The Irish actor Charles Macklin was first to play Shakespeare in authentic "ancient dress" when he appeared on the stage in 1773 as Macbeth in a highland military habit.

By courtesy of the trustees of the British Museum; photograph, J.R. Freeman & Co. Ltd.



David Garrick costumed for the role of Richard III; engraving by William Hogarth, 1746. In the British Museum.

Public knowledge of historical costume was increasing, though some critics disapproved of it, believing it weakened the plot. The thirst for historical accuracy won the day, in part because of the efforts of its major champion, the 19th-century playwright and antiquary James Robinson Planché. A playbill of a Planché production of 1824 read:

> Shakespeare's Tragedy of King John with an attention to Costume never equalled on the English Stage. Every Character will appear in the precise HABIT OF THE PERIOD, the whole of the Dresses and Decorations being executed from indisputable Authorities. . . . King John's Effigy in Worcester Cathedral . . . Queen Elinor's Effigy in the Abbey of Fontevraud. . . . Illuminated Mss in the British Museum, Bodleian and Bennet College Libraries . . . &c.

Planché's success amply reimbursed the management for the enormous expenditures required to make even the meanest extra's garments authentic. Correct costumes became the order of the day. The actor-manager Charles Kean splendidly mounted a series of Shakespearean productions in London in the 1850s. In Germany, August Wilhelm Iffland's productions closely followed the same reforms, and costume designers were urged to emulate the past.

Dance costume

Dancers, too, wished to be liberated from the pannier and *tonneler*. In *Pygmalion*, a ballet of her own composition, Marie Sallé danced in London in 1734 dressed only in a muslin robe like that of a Greek statue; she wore no panniers, petticoat, or bodice, and her hair was loose and without ornament. The French ballet master Jean-Georges Noverre had advocated drastic reform in dance dress as early as 1760. He proposed to put his dancers in simple draperies of contrasting colours, worn in such a manner to reveal the figure in movement. Both male and female dancers became increasingly impatient with the hampering Rococo conventions; and as dancing technique grew even more intricate, designers prepared innovations that anticipated the general development in fashion. By the end of the 18th century, every woman had abandoned her panniers and was "dressed only in a single muslin robe."

Civil codes at the start of the 19th century greatly aided those who would reform theatrical costume. The fashion for classic robes and tunics modelled on Greek and Roman patterns took to the streets; the high-heeled shoe was replaced by a flat slipper.

Leading French ballet designers of the period were Auguste Garneray and Hippolyte Lecomte. The former's work shows ingenuity in adapting contemporary dress to suggest different lands and periods. The latter was originally a painter of historical episodes; accuracy rather than imagination is the distinguishing quality of his designs. In 1832 the influence of the Romantic period was first seen in ballet with a production of *La Sylphide*. Eugène Lami designed for Marie Taglioni, the greatest dancer of her day, a muslin dress, an ethereal costume that became the new uniform of the classical dancer and is still worn today.

Opera and dramatic productions followed Planché's regard for period and regional detail. Unfortunately, the elaborate stage settings and the costumes were seldom coordinated successfully. This led the Duke of Saxe-Meiningen to design his own productions, as he wished to achieve an integrated style. He was the first of the directors who controlled all aspects of production. His example has been followed by the most prominent directors of the 20th century.

Amid spectacular scenery, large ensembles danced and mimed in fanciful versions of contemporary and historical dress in the ballet-extravaganzas of the 1880s (forerunners of the Folies-Bergères revues of Paris), that played at La Scala in Milan and the London Alhambra. The ingenious designer Wilhelm translated insects, flowers, birds, and reptiles into dance costumes. The main interest, however, lay in framing the female figure, and most of the costumes were designed to reveal as much as the law permitted.

20th-century costume. At the turn of the century the theories of Appia and Craig called for symbolism and voiced a strong reaction against the naturalism of Meiningen and the Moscow Art Theatre. Appia advocated that stage costume should evoke and suggest, but never copy historical lines.

A painters' theatre arose after Diaghilev's presentation of the Ballet Russe in 1909 in Paris. The brilliant palettes and well-coordinated decors (of Bakst, Benois, and Roerich) were praised. Natalya Goncharova's design for *Le Coq d'or* in 1914 was unprecedented in its use of vivid colours, chiefly shades of red, yellow, and orange, with other colours for discordant emphasis. The forms of the costumes and their decorations were based on Russian folk dress (of fairly recent times, though transformed and made uniquely of their epoch by the jaggered influence of Cubism). Avant-garde artists of many of the flourishing movements of modern art—cubists, constructionists, surrealists—brought about an acceleration of innovation in design concepts, which in previous centuries had evolved gradually throughout Europe.

Expressionist theatre costumes

Appia, Craig, and Diaghilev led others to experiment. Meyerhold's 1907 production of Leonid Andreyev's play *The Life of Man*, with expressionistic costumes designed by Theodore Komisarjevsky, was purely mechanical in its design. The German advocates reasoned that since the actor man is enclosed in the space of the stage, either the stage must be arranged according to the illusion of reality so as to fit the natural man or the natural man must be transformed to match the cubist and abstract space of the stage.

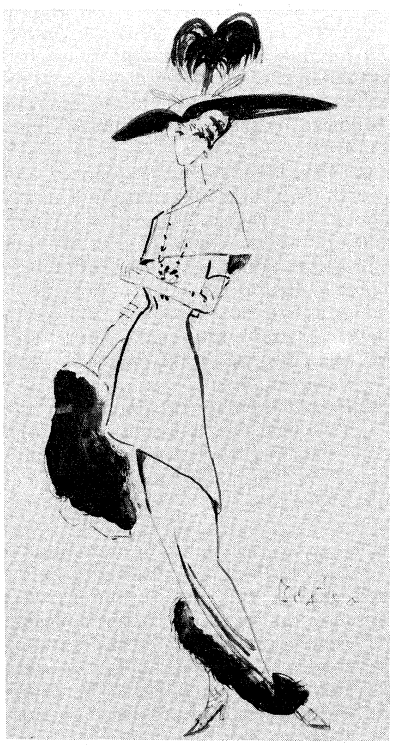
The most daring stage work of the 1920s was created in Dessau, Germany, by Oskar Schlemmer, of the faculty of the Bauhaus school of design He devised a costume of coils of wire projecting from black tights; his masklike designs reduced the human head to a sphere, and by means of vertical white tapes he sought to transform the human body into a spinning top. Many of his ballet designs anticipated the space suits of the astronauts.

As modern dance evolved, its rapid rhythms and pace offered the costume designer new challenges and scope for original work. The creative presentations of Merce Cunningham and Alwin Nikolais in New York presented unique shapes that attempted to express the exploration of time and space. Nikolais made his costumes part of a total stage design, a theatrical abstraction of the way he saw man—as part of a socio-economic mechanism, an agreeable but not a central part. Accused of dehumanizing the dancers, he maintained they are rather an expression of a greater state of being for man, of the experience of living in a world of motion, sound, colour, and action, which strongly affects them and is affected by them.

The rise of the American musical owes much to the close collaboration of the composers, writers, and choreographers with the designers. In a musical, the costume

designer's task is manifold. The designs must capture the spirit of the music and lyrics, interpret the period, heighten the characterization of the actors, and help the dancers in their varied, often athletic routines. The costumes for *West Side Story* (1957) and *My Fair Lady* (1956) were especially successful in these respects.

Sir Cecil Beaton



Sir Cecil Beaton's design for Liza Doolittle in the "Ascot scene" of *My Fair Lady*.

New concepts in 20th-century costume

The radical creative talents of Bertolt Brecht created new production concepts and styles: the clothes worn conveyed to the audience in a satirical fashion not only a characterization of the wearer but also his social status. Another compelling force in contemporary experimental theatre, Jerzy Grotowski, conceived his production *Akropolis* at the Polish Laboratory Theatre as a poetic paraphrase of an extermination camp. There is no hero and no individuality among the characters. The costumes are bags full of holes covering naked bodies, and the holes are lined with material suggesting torn flesh; wooden shoes cover the feet, and anonymous berets the heads.

A form of 20th-century pageantry can be seen in the designs of Tanya Moiseyevich for the simple thrust stage of the Tyrone Guthrie Theatre, Minneapolis, Minnesota, and the Festival Theatre, Ontario, Canada. For the brilliant productions of these theatres, with their deft handling of large crowd scenes, highly skilled costumiers are engaged each season. The costumes and props that they make enhance the productions' limited scenic elements, recalling the style of the Greek classic theatre and that of Japan. The growth of such regional theatres in Europe and the United States has afforded many opportunities for costume designers. Most of the opera and repertory companies that operate in larger cities have their own well-established costume workshops.

CONTEMPORARY COSTUME-DESIGN PROCEDURES

Coordination of costume sketches with the scenic and lighting designers' schemes is essential from the start. The line, shape, and colour of the entire company's costumes must be viewed in relation to the set and the production's lighting plot. Rough preliminary sketches are made and discussed, revised, and redrawn. A thorough and creative meeting of minds at this stage is invaluable. It may be necessary to call many production meetings before the final rough sketches are approved by the director.

The costume designer

The costume designer can now search for the appropriate fabrics. Fabric buying is a time-consuming and arduous task, conditioned by the size of the production budget. Costumes can be made from an enormous variety of materials; often what is needed can be located only with ingenuity and imagination in some unusual place. The correct weight and weave must be selected if the cutter and draper are to translate the sketches into fact. The fabrics must be suitable for shaping and molding the silhouettes and styles as indicated by the designer, and they must hold up for as long as the production runs. Heavy-textured wools, fine cottons, upholstery fabrics, industrial felts and rubbers, natural fibres, and man-made stiffeners all help to create costumes for varied stage characters. When they are finally selected small samples of the fabrics are stapled to the rough sketches.

Once the cast list is made known, final costume renderings can be completed. Many designers sketch details of all millinery, wig work, and accessories. Hand props, such as snuff boxes, are also often considered the costume designer's responsibility. Spectacles, parasols, and walking sticks also fall in this category.

The costume maker

The management submits approved final designs to costume makers and invites bids. The costumers estimate prices according to the number of man-hours required for manufacture and the price of the fabrics to be used. The designer often confers with the producer regarding the figures submitted by the costume makers. The designer may recommend a particular firm because it has the craftsmanship and skills required for a particular style of work. Contracts are finally drawn up and construction begins.

A costume house is staffed with experienced and creative crafts people, who translate all items of the costume sketch into shapes fashioned in the chosen fabrics. Millinery, jewellery, armour making, and dyeing can often be supervised by the designer under one roof. The construction of costumes starts with the appropriate underwear and necessary padding. Corsets, hooped skirts, panniers, bustles, and petticoats form the base on which period silhouettes are fashioned; they are fitted on the actors before the basic muslin pattern of the overgarments. On the skills of the cutter, draper, and fitting hands rests the ultimate success of the finished garment. All fittings are crucial, and the designer normally is present to supervise.

At the second fitting, shoes, wigs, and headwear must be tried on; and at the third and final fittings, the complete outfit is presented for approval of the designer and cutter. The dress parade follows, usually held onstage before the appropriate setting. This parade, ideally, is held at least ten days before the dress rehearsal to allow for revisions by the director. Despite the long continuing collaboration, the designer still may find himself spraying, re-dyeing, and artificially aging garments to make the stage picture perfect in every detail. Last-minute adjustments can be most satisfying. When the costumes are completed, the production's own wardrobe staff takes over responsibility from the costume house. Supplied with inventories and wardrobe plots, the staff takes care of all items during the run of the production. (M.H.)

VI. Theatrical makeup

Theatrical makeup is the practice of painting, enhancing, or altering the face and body of the actor with cosmetics, plastic materials, and hair; it is also the collective term for the materials used in making-up. Actors have used makeup in the theatre for a long time, not only to look their best and to transform their appearance but also to ensure that they will be seen and recognized by the entire audience. How well an actor is seen depends upon the distance between him and the farthest spectator and the amount of available light. Distances blur the features and make recognition of the actors extremely difficult for the spectators.

EARLY DEVELOPMENT

It is impossible to say when the first actor daubed paint on his face in an attempt to make his performance more effective. Some maintain that Thespis, the first actor to

step out of the chorus in Greek theatre in the 6th century BC, smeared his face with white lead and red cinnabar. He may have done so, but the very large size of some of the Greek theatres (containing up to 15,000 spectators) made the use of the masks a more practical solution. Not only did masks enable the actor to play more than one role but their larger-than-life size and exaggerated features enabled the audience to identify the actor's role from a distance.

Use of masks

The Roman theatre borrowed and continued the use of the mask. Evidently some actors thought that their own faces would be more effective, however, since there are contemporary descriptions of certain mimic actors with painted faces. Like the Roman mimes, the travelling actor-comedians of the commedia dell'arte in the 16th century developed a set of stock characters using masks. There may have been some continuity of tradition, for there are resemblances betweeen the comedians' leather masks and those of the Roman farces.

As actors in the 17th century began to perform in permanent theatres, in which they were separated from the audience by the proscenium arch, they had to deal with the problem of making the face look distinctive in illumination from candles and oil lamps. It is probable that the actors adapted the cosmetics of fashionable ladies to their own use.

The use of cosmetics increased greatly during the 18th century. Women, and men to a lesser extent, painted their faces with lead oxide for a pale complexion and cinnabar (mercuric sulfide) for rouge. Both were poisonous compounds. With the widespread use of cosmetics it was noticed that the paler faces were easier to see in dim light. By the 1770s the use of white paint was general upon the stage even though it tended to conceal the expressive motion of the muscles.

The impact of artificial light

The introduction of limelight in 1816, of gaslight in 1817, and arc lights in 1846 created the necessity for a more subtle use of cosmetics on the stage. The actor's palette consisted at this time of white chalk, carpenters' blue chalk, papers impregnated with red colouring, and India ink. Some actors were known to make up their faces with whitewash rubbed off the walls, dust scraped from red bricks, and black from a burned match, paper, or cork. Credit for the invention of greasepaint belongs to Carl Baudin of the Leipziger Stadt Theatre. Wishing to conceal the join between the front edge of his wig and the forehead, he mixed a flesh-coloured paste of zinc white, yellow ochre, vermillion, and lard. By 1890 theatrical greasepaints were available commercially in many colours, and other items of makeup also began to be manufactured for use in the theatre.

CONTEMPORARY PRACTICES

The practice of makeup in the theatre has changed greatly with the wave of vitality that swept over the theatre after World War II. As the stage was thrust into the auditorium and the audience was arranged experimentally on all sides of the acting area, the actor was brought closer to the spectator, making close scrutiny possible. Poorly executed and heavy makeup looked obvious, and several time-honoured practices had to be abandoned. Red dots in the corners of the eyes and the use of bright eyeshadow (which had first been used to counteract the effects of footlights) appeared ridiculous. Actors began to reject greasepaint in favour of a makeup that had been developed for the motion picture industry, a water-soluble cake makeup capable of being spread in a thin wash of colour, making a completely natural look possible. The desire for naturalism in appearance became very important.

Many improvements in makeup techniques were developed primarily for films and television and rapidly adopted by the theatre. There has been a great change in theatrical wigs, for example, which once had a band of silk that was blended with colour into the forehead This type of wig was not successful for film work since the join was obvious to the camera. A new type of wig was developed, and now most wigs for films and theatre are made with an edge of plastic lace into which hairs are knotted. When the lace is glued to the forehead the hairs seem to grow from the skin, resulting in a natural-looking hairline.

Film and television makeup artists have perfected the use of foam rubber and plastic facial appliances to the point that a young actor can be transformed into a very old man and yet withstand the scrutiny of a camera close-up. This process frequently takes a makeup artist and an assistant four hours to complete. Audiences accustomed to these miracles of transformation are beginning to expect the same verisimilitude in the theatre. But this kind of makeup requires more time than is generally practical in the theatre, and the cost could be prohibitive. Theatres cannot hope to compete with the motion picture industry: one film budgeted $1,500,000 to transform a large cast of actors into realistic-looking apes.

There is a growing trend in the theatre to include makeup in the total production design. What was once considered the prerogative and responsibility of the individual actor is now being influenced by designers, directors, and, in some instances, by makeup artists. In most German opera houses, for example, singers are not allowed to do their own makeup. A staff of makeup artists assures that each makeup is executed as designed and that the style is consistent. This expert supervision has hitherto been considered possible only in the motion picture industry or in commercial television because of costs, but government subsidy of theatres and television in England and other European countries has enabled the hiring of such specialists. Rubber latex makeup can be used on the stage to give the skin the textured look of extreme old age, to make bald caps, and to make realistic-looking false noses, jowls, and other forms of sagging flesh.

The emphasis upon design and production techniques in recent years has, in the opinion of some, tended to minimize the importance of the actor and, therefore, of makeup. Jerzy Grotowski, director of the Polish Laboratory Theatre, said that his experiments indicated that theatre could exist without makeup, though he did not wish to eliminate it entirely. Rather, he advised the actor to try to form himself without its help; then, if it is used after all the changes possible without makeup have been studied, it will be much more expressive and will surpass all technical tricks.

Others have sought a return to primitive origins, to a ritual theatre of magic and fantasy, unencumbered by the constraints of objective reality. Experiments in ritual theatre tend to result in the use of masks or a decorative style of makeup, as is found in Chinese opera and Kabuki theatre. (S.D.)

BIBLIOGRAPHY

Staging and stage design: A.M. NAGLER, *A Source Book in Theatrical History* (1959), the primary reference for basic documents; ALLARDYCE NICOLL, *World Drama: From Aeschylus to Anouilh* (1949), the definitive historical survey; *The World of Harlequin* (1963), the authoritative work on the commedia dell'arte; and *A History of English Drama, 1660–1900,* 6 vol. (1952–59), from the Restoration through the 19th century; PHYLLIS HARTNOLL (ed.), *The Oxford Companion to the Theatre,* 2nd ed. (1957), the only one-volume encyclopaedia in the field; MARGARET BIEBER, *The History of the Greek and Roman Theatre,* 2nd ed. rev. and enl. (1961); E.K. CHAMBERS, *The Medieval Stage,* 2 vol. (1903, reprinted 1954); FAUBION BOWERS, *Theatre in the East: A Survey of Asian Dance and Drama* (1956); A.C. SCOTT, *An Introduction to the Chinese Theatre* (1959); BALWANT GARGI, *Theatre in India* (1962); BARNARD W. HEWITT (ed.), *The Renaissance Stage* (1958), documents on stage scenery; *Theatre U.S.A. 1668 to 1957* (1959), the best historical survey of American theatre; ASHLEY H. THORNDIKE, *Shakespeare's Theatre* (1958), the stages and physical productions from Elizabeth through Charles I; LEE SIMONSON, *The Stage is Set,* rev. ed. (1964), the background of modern stagecraft; MORDECAI GORELIK, *New Theatres for Old* (1940), the historical roots of modern production and scenic design; INTERNATIONAL THEATRE INSTITUTE, *Stage Design Throughout the World,* 2 vol. (1956–64), a pictorial record of contemporary scenic art.

Lighting: WILLARD F. BELLMAN, *Lighting the Stage: Art and Practice* (1967), the best textbook on American equipment and procedures; THEODORE FUCHS, *Stage Lighting* (1929), classic work including treatment of electricity, optics,

and control systems; RICHARD PHILBROW, *Stage Lighting* (1971), contemporary English practice.

Theatrical costume design: ALLARDYCE NICHOLL, *The Development of the Theatre*, 5th ed. rev. (1967), a study of theatrical art from the beginnings to the present day; PHYLLIS HARTNOLL, *A Concise History of the Theatre* (1968), and BAMBER GASCOIGNE, *World Theatre* (1968), two books containing the practical physical details in the theatre's history with many illustrations; THEODORE KOMISARJEVSKY, *The Costume of the Theatre* (1968), and JAMES LAVER, *Drama: Its Costume and Decor* (1951), comprehensive accounts of costume history with well-chosen illustrations; MILLIA DAVENPORT, *The Book of Costume* (1948), a comprehensive detailed account of costume through the ages covering dress, jewellery, ornament, coiffure, and all other elements including social manners; HILAIRE and MEYER HILER (comps.), *Bibliography of Costume* (1939, reprinted 1967); ISABELLE MUNROE and DOROTHY E. COOK, *Costume Index* (1937); for Oriental references: IWAO KONGOW, *Nō-isho: Japanese Nō-Play Costume*, 2 vol. (1934), RUTH M. SHAVER, *Kabuki Costume* (1966), and MAX TILKE, *Oriental Costumes* (1923).

Stage makeup: RICHARD CORSON, *Stage Makeup*, 4th ed. (1967), a comprehensive study of all aspects of stage makeup and its application, and *Fashions in Hair* (1970), a survey of hair styles through the ages.

(Ho.B./Ra.Ho./M.H./S.D.)

Stahl, Georg Ernst

Georg Ernst Stahl, German physician and chemist, developed the phlogiston theory, the first comprehensive explanation of the phenomenon of combustion and of such related biological processes as respiration, fermentation, and decay. While Stahl's theory was eventually replaced by that of Antoine Laurent Lavoisier (1743–94), it dominated chemical thought for almost a century and provided a rational basis for much useful experimental work. It led Stahl himself to define the relationships between acids, bases, and salts; to determine the relative strength of acids; and to distinguish between the bases of sodium and potassium.

By courtesy of the Bildersammlung des Universitatsarchivs, Martin-Luther-Universitat Halle, Wittenberg, East Germany



Stahl, copper engraving attributed to a member of the Bernigeroth family, 18th century.

Early training

Stahl was born on October 21, 1660, in Ansbach, Franconia, Germany. He grew up under the influence of Pietism, a vigorous movement for moral reform in the Lutheran Church that had begun to spread across Germany. German Pietism stressed the inward, practical, devotional, and tolerant aspects of Christianity in a manner not unlike that of the Friends (Quakers) in England. Stahl studied medicine at Jena, Thüringen. His teacher, Georg Wolfgang Wedel (1645–1721), was a proponent of the chemiatric school, which sought to explain life processes in terms of chemical composition. Even when Stahl later broke new ground in chemistry as well as in medicine, his early training remained apparent in his firm belief in the essential unity of these two sciences. At Jena, Stahl formed a close friendship with his fellow student Friedrich Hoffmann (1660–1742). Stahl graduated in 1684, having begun to teach the year before. In 1687 he was appointed court physician by Johann Ernst, duke of Sachsen-Weimar.

In 1694, Frederick III, elector of Brandenburg, founded the University of Halle. When Hoffmann was asked to organize the medical school of the new university, he asked Stahl to join him. For the next twenty years the two worked side by side; Stahl taught theoretical medicine, physiology, pathology, dietetics, pharmacology, and botany, while Hoffmann lectured on practical medicine, anatomy, physics, and chemistry. But gradually deep temperamental differences and increasingly divergent views on theoretical questions divided them. Hoffmann became an ardent convert to the school of iatromechanics, which saw living organisms as machines to be explained by the laws of classical mechanics. Stahl could not accept such an oversimplification. He remained faithful to his iatrochemical training in seeing physiological processes as chemical changes, but he insisted that neither mechanical nor chemical laws alone were sufficient to account for the phenomenon of life. To live, an organism required direction by a force for which Stahl used the Latin word *anima*. That word, the full meaning of which is not well caught in the usual translations of "soul" or "spirit," gave rise to the name animism for this theory. Vitalism and antimaterialism are later words for very similar views.

Their opposition in scientific matters eventually destroyed the friendship between Stahl and Hoffmann. The friends became rivals and, at last, enemies. Stahl must have welcomed his call in 1716 to Berlin, where he served as personal physician to Frederick William I, king of Prussia. He died there on May 14, 1734.

Assessment

Stahl's contributions to science are to an extraordinary degree reflections of his personality. By character he was contemplative, melancholy, and deeply inclined toward the inner-directed religious Pietism of the late 17th century. His intellect, on the other hand, was analytical and sweeping and compelled him toward large abstractions in the manner of the 18th century. Stahl bridged two ages, as did his phlogiston theory. It was the first systematic account of chemical transformation, but without the concepts of atom or element. The roots of his ideas lie not in the prior use of the word phlogiston by Jan Baptist van Helmont (1577–1644) or in the vaguer theories of Johann Joachim Becher (1635–82) but in Aristotelian thought: phlogiston is the property of being combustible that is lost in the process of combustion but that may be regained by contact with materials especially rich in phlogiston, such as coal. On an ancient philosophical foundation, Stahl built a practical theory that clarified fundamental chemical concepts.

BIBLIOGRAPHY. There is no published biography of Stahl; biographical information may be found in the article by C.W. BECK, "Georg Ernst Stahl," *J. Chem. Educ.*, 37:506–510 (1960). See also H. METZGER, *Newton, Stahl, Boerhaave et la doctrine chimique* (1930), a specialized monograph on chemical theory; and B.J. GOTTLIEB, *Bedeutung und Auswirkungen des Professors G.E. Stahl auf den Vitalismus* (1943), a specialized study of vitalism.

(C.W.B.)

Stained Glass

All coloured glass is, strictly speaking, "stained," or coloured by the addition of various metallic oxides while it is in a molten state; nevertheless, the term stained glass has come to refer primarily to the glass employed in making ornamental or pictorial windows. However, the singular colour harmonies of the stained-glass window are due less to any special glass-colouring technique per se than to the exploitation of certain properties of transmitted light and the light-adaptive behaviour of human vision. Rarely equalled and never surpassed, the great stained-glass windows of the 12th and early 13th cen-

turies actually predate significant technical advances in the glassmaker's craft by more than half a century. And much as these advances undoubtedly contributed to the delicacy and refinement of the stained glass of the later Middle Ages, not only were they unable to arrest the decline of the art, they may rather have hastened it to the extent that they tempted the stained-glass artist to vie with the fresco and easel painter in the naturalistic rendition of his subjects. Neither painting on stained glass nor its assembly with grooved strips of leading is an indispensable feature of the art. Indeed, the leaded window may well have been preceded by windows employing wooden or other forms of assembly such as the cement tracery that has long been traditional in Islāmic architecture; and the single most important technical innovation in 20th-century stained glass, slab glass and concrete, is a variation on the earlier masonry technique.

ELEMENTS AND PRINCIPLES OF DESIGN

Of all the painter's arts, stained glass is probably the most intractable. It is bound not only by the many light-modulating factors that affect its appearance but also by comparatively cumbersome, purely structural demands. And yet no other art seems so little earthbound, so alive, so intrinsically beguiling in its effect. This is because stained glass, far more directly and intensively than other media, exploits the interaction between two highly dynamic phenomena, the one physical and the other organic. The physical factor is light and all of the myriad changes in the general light level and the location and intensity of particular light sources that occur as a matter of course not only from moment to moment but from place to place—a prairie to a forest, a greenhouse to a dungeon. The other phenomenon is the spontaneous light-adaptive process of vision, which seeks to maintain orientation in all luminous environments.

Architecture, by determining the apparent brightness value of the light seen through its window openings, always establishes a definite scale of brightness values with which the stained-glass artist must work. Because the light that penetrated the interior of the 12th- and early-13th-century church took on a brilliance, even harshness, in contrast to the surrounding darkness, the artisans of the period very logically composed their windows with a palette of deep, rich colours. When for doctrinal or economic reasons only clear glass could be used, it was decorated with a fine opaque mesh of grisaille, or monochromatically painted ornament, that effectively broke up and softened the light. Later, as the walls of the churches were opened up to admit more and more light, the difference between the interior and exterior light levels was no longer great enough to illuminate the dense, saturated rubies and blues of the earlier period. In the 14th and 15th centuries, generally higher keyed, drier, and more muted colour harmonies were developed. This reflected a growing preference for lighter, less awesome effects and an actual limitation that the architecture of the time imposed upon the medium of stained glass.

Role of light

The static elements of the glass and its architectural setting are modified by the element of change inherent in natural light. A seemingly endless spectrum of changes in the appearance of stained glass is a result of the changes in the intensity, disposition, atmospheric diffusion, and colour of natural daylight. The luminous life of stained glass, therefore, can best be observed by watching the organic effect of light on the window through the course of a day. If one were to enter the Cathedral of Chartres just after sunrise on the morning of a clear day, it would be to the east windows, especially those in the clerestory, that his eyes would first be drawn. They alone will have come fully to life and all of the others will still seem to half-exist in a kind of hushed twilight. Gradually, as the sun rises in the sky, these windows will become more luminous. Then the east windows will begin to lose their earlier brilliance to those all along the south flank of the cathedral, which by midday will be fairly aglow from the direct rays of the sun. The light streaming through the south windows, however, will have raised the light level inside the north windows opposite them sufficiently to create a distinct, though by no means unpleasant, muting of the radiance of the latter. If the sun at this point disappears behind a cloud and the sky becomes generally overcast, the appearance of all of the windows is immediately and dramatically altered. Because the light, now diffused, comes more or less equally from all directions, the south windows will lose some of their earlier brilliance and vivacity and the north windows will recover theirs. The overall atmosphere of the cathedral is distinctly cooler and graver in its effect, and more than ever before one begins to become aware of absolute differences in the tonality of the various windows themselves. The grisaille windows in the east end of the cathedral, the highly keyed 15th-century window in the Vendôme Chapel in the south aisle of the nave, and the three 12th-century windows over the great west portal all stand out as being substantially more luminous than the rest. If, late in the afternoon, the sun reappears, the viewer is treated to an extraordinary spectacle as the blues in the west windows, by far the most intense in the cathedral, are further emblazoned by the direct rays of the sun. Should the main doors of the cathedral be opened, the direct rays of the late afternoon sun, streaming halfway down the nave of the cathedral, will cast a blinding pall over all the windows within their vicinity until the doors are closed once more. Then as the sky begins to redden with the setting sun, the intense 12th-century blues in the west windows lose their former intensity, and the warmer colours, especially the rubies, become so fiery and assertive that they seem almost to have displaced the blues as the predominant colour in the windows. Finally, when the sun is gone the whole cathedral is plunged once more into a deep twilight, which gradually diminishes until there is no light at all.

Insofar as stained glass may be considered an art of painting, it must be considered an art of painting with light. Whatever techniques or materials it may employ, its own most unique and indispensable effects are always the product of colouring, refracting, obscuring, and fragmenting light.

MATERIALS AND TECHNIQUES

Contrary to popular belief, the glassmaker and the stained-glass artist could seldom have been the same person even in the earliest times; in fact, the two arts were rarely practiced at the same location. The glassmaking works was most readily set up at the edge of a forest, where the tremendous quantities of firewood, ash, and sand that were necessary for the making of glass could be found, whereas the stained-glass-window-making studios were normally set up near the major building sites. The stained-glass artist, thus, has always been dependent upon the glassmaker for his primary material. Coloured with metallic oxides while in a molten state—copper for ruby, cobalt for blue, manganese for purple, antimony for yellow, iron for green—sheets of medieval glass were produced by blowing a bubble of glass, manipulating it into a tubular shape, cutting away the ends to form a cylinder, slitting the cylinder lengthwise down one side, and flattening it into a sheet while the glass was still red hot and in a pliable state. It was then allowed to cool very slowly in a kiln so that it would be properly annealed and not too difficult to cut up into whatever shapes might be required. Since these sheets of glass, with the exception of a type known as flashed glass, were intrinsically coloured with one basic colour throughout, changes from one colour to another in the design of a window could be effected only by introducing separate pieces of glass in each of the requisite colours.

Role of the glassmaker

Whether by accident or by deliberate intent, the glass made in the 12th and 13th centuries had almost the ideal combination of crudity and refinement for stained glass. The sheets, 10 by 12 inches (25 by 30 centimetres) in size, were both flat enough and thin enough to be cut very accurately into the necessary shapes, yet still variable enough in thickness (from less than ⅛ inch [0.317 centimetre] to as much as 5/16 inch [0.793 centimetre]) to have rich transitions in the depth of its colour. With the progress of glass technology in the Middle Ages and Renais-

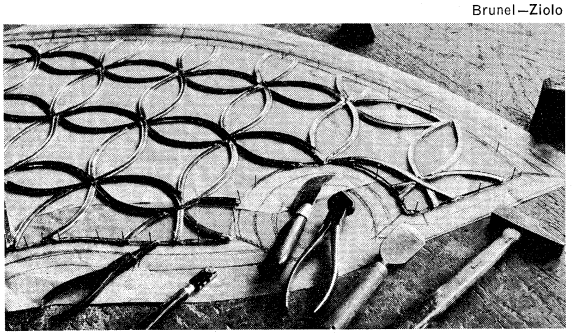
sance came the ability to produce larger, thinner, and flatter sheets of glass in a considerably larger range of colours than had been possible in the 13th century. At each distinguishable stage in this development, however, the glass became less visually interesting as an aesthetic element in its own right. The Gothic Revivalists later realized this, and in the mid-19th century they initiated a return to the earlier methods of producing glass. They developed the so-called "antique" glass, which is remarkably similar in colour, texture, and shading to that used in the 12th- and 13th-century windows. "Antique" glass remains the basic material used in stained-glass windows to this day.

Traditional techniques. The art of stained glass is the translucent offspring of such earlier art forms as mosaic and enamelling. From the mosaicist came the conception of composing monumental images out of many separate pieces of coloured glass. Cloisonné enamelling probably inspired not only the technique of binding these pieces together with metal strips but that for treating the strips themselves as a positive design element. From the enamellers must also have come the near-black vitreous enamel made from rust powder and ground glass that was mixed with a mild water-based glue to form a paint. This could be used to render more or less opaquely onto glass the details of figures, ornaments, and inscriptions.

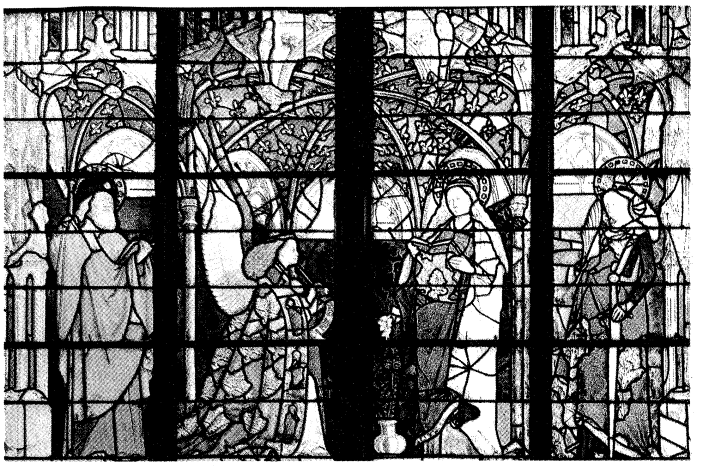
The technique of making stained-glass windows is first described in the *Schedula diversarum artium*, a compendium of craft information probably written between 1110 and 1140 by the monk Theophilus (tentatively identified as the 12th-century goldsmith Rugerus of Helmarshausen). First, a full-sized cartoon, or line drawing, of the window was painted directly onto the top of a whitewashed table, showing the division of the various colour areas into individual pieces of glass. Next, sheets of glass of the appropriate colours were selected and from these pieces were cut, or, more accurately, cracked away with a red hot iron. By applying the hot iron to the edge of the sheet it was possible to start a crack that could then be guided more or less in the direction in which the iron was moved, thus enabling the glazier to break away from the sheet of glass a piece of approximately the right shape and size. This he would then further shape by "grozing," or crumbling away bits of glass from its edges with a notched tool known as a grozing iron. When all of the pieces were thus accurately cut to shape, with due allowance between pieces for the leads that would join them together, the details of the design were painted onto the glass wherever necessary with vitreous enamel. The pieces were then placed in a kiln and fired at a temperature just hot enough to fuse the enamel to the glass. This done, the windows were ready for assembly with grooved strips of lead that look in cross section like the letter H. The glazier would begin by butting together on his workbench two long strips of lead, to form a corner of the panel. He would then set the corner piece of glass in place between these two leads and cut another strip of lead just long enough to surround the rest of the piece. Against this lead he would then be able to set the next piece of glass, and so on across the panel, until it was completely assembled on the glazing bench. The joints between the leads were then soldered, the panel was waterproofed by rubbing a putty compound under the leads, and it was ready for installation.

Cutting and leading techniques

Brunel—Ziolo



Construction of a stained-glass window.
The craftsman has assembled a section of a large window on his glazing bench. Visible in the photograph are the design beneath the window and various tools for cutting glass and working the lead.

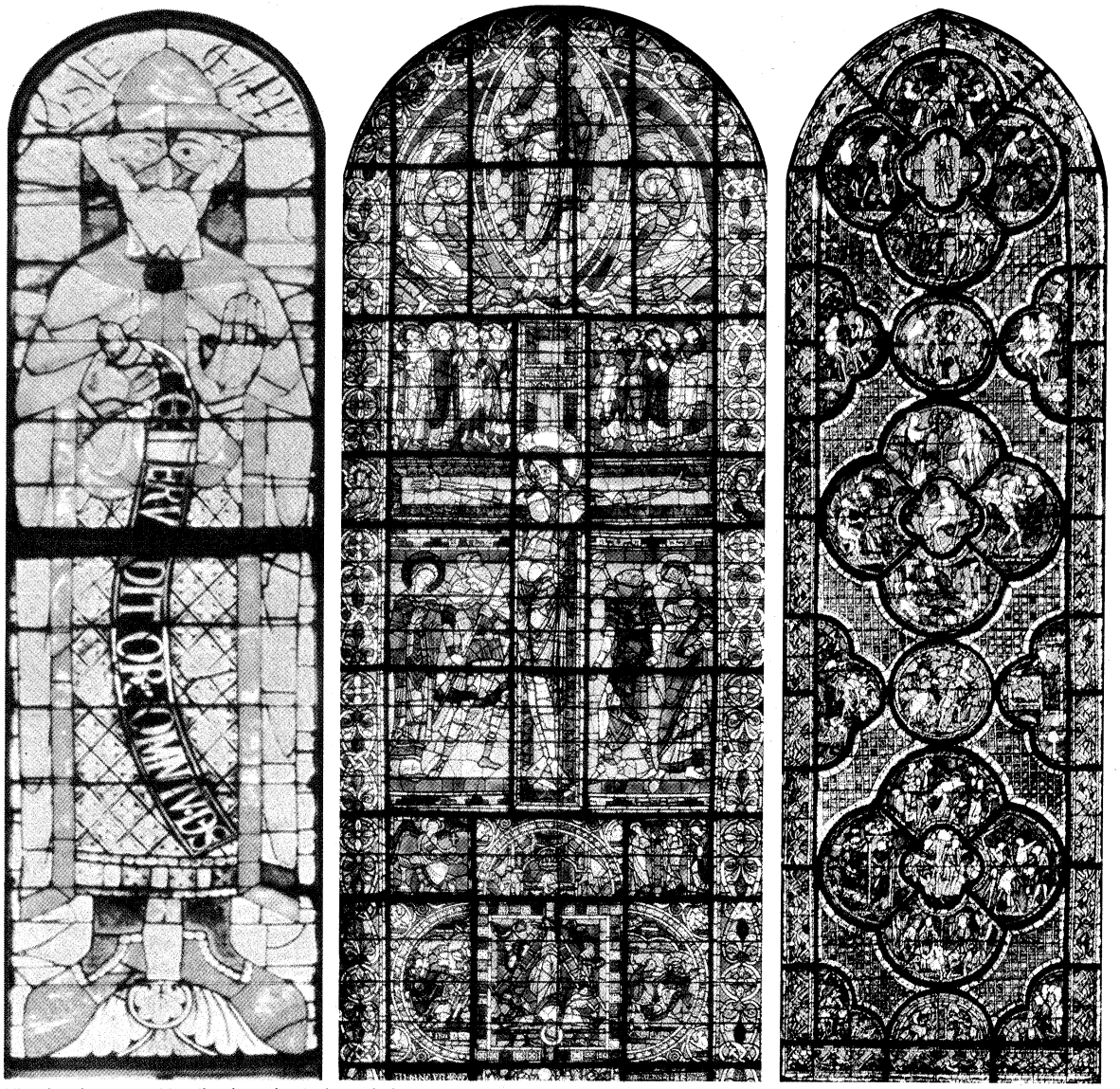


Continuous narrative in four windows.
"The Annunciation," the Jacques Coeur window, *c.* 1450. In Bourges Cathedral, France.
Promophot—Ziolo

Because of the flexibility of the leading it was found necessary to divide all but the very smallest windows into a series of separate leaded panels and to insert iron framing members, or armatures, between the panels. In the earliest single-figure lancet windows, such as the "Prophets" in Augsburg Cathedral, the divisions tend to be purely functional. Very soon, however, more ambitious windows became much too large to be handled in this manner. Whereas the Augsburg "Prophets" measure only about 12 square feet (1.1 square metres) in area, the Poitiers Cathedral "Crucifixion" window contains approximately 175 square feet (16.3 square metres) of stained glass, and the "Life of Christ" in Chartres contains more than 250 square feet (23.2 square metres). A much more elaborate system of subdivisions in the window opening, consisting of vertical as well as horizontal members, was developed. These systems of supports often formed a geometric pattern that was incorporated in the overall design of the window. In fact, it was the ingenious conversion of this structural necessity into a positive design element that set the stage for the creation of the medallion windows of the great Gothic cathedrals. By utilizing these armatures to delineate the principal ornamental subdivisions of the windows, as in the Chartres "Good Samaritan," the glass painters were able to fuse a complex didactic imagery and an austere architecture into one of the most compelling artistic unities of Western art. At the same time, particularly in the upper levels of a church, stone mullions began to be employed for the same purpose. The most spectacular examples are the great rose windows, in which masonry is so literally dissolved into fenestration, and the individual window opening so completely absorbed into the overall pattern, as to defy any meaningful distinction between window and wall. This perfect fusion of image, ornament, and structure, with each deriving strengths from the others that none would ever have alone, was one of the most significant turning points in the history of stained glass. From this point on the relation between stained glass and architecture begins to decline. The aims, techniques, and achievements of the stained-glass artist begin to resemble those of the fresco and easel painters, and it is by the standards applicable to the latter that the stained glass of the 14th, 15th, and 16th centuries must be judged.

Developments in the 14th century. The first significant developments in the glass painter's craft appear to have been made more or less simultaneously in the early years of the 14th century. Glass in a range of previously unavailable secondary colours—smoky ambers, moss greens, and violet—become generally available for the first time. The technique of staining glass yellow by painting it with silver salts is discovered. The glass painters also begin to develop a number of techniques for shading or modelling

Painting and grinding techniques



The development of leading in stained-glass windows.
(Left) "The Prophet Hosea," single-figure window, *c.* 1125. In Augsburg Cathedral, Germany.
(Centre) "The Crucifixion," elaborately divided window, *c.* 1165. In Poitiers Cathedral, France.
(Right) Scenes from the life of the "Good Samaritan," medallion windows, first quarter of the 13th century. In the south aisle of the nave of Chartres Cathedral, France.
(Left) Bavaria-Verlag, (centre) Lauros—Giraudon, (right) Giraudon

model forms with vitreous enamel by applying translucent matts of halftone to the whole surface of the window and delicately brushing it away where highlights are desired. Darker shading is sometimes reinforced by painting on the outer as well as the inner surface of the glass. The uses of line also become increasingly refined and versatile, especially in the 15th century.

To these refinements of the craft was added one wholly new technique, the abrasion of flashed glass. Ruby glass, whose unique composition made this technique possible, was a laminated glass, although it appears to be coloured intrinsically throughout like all of the other glass in the early windows. Because the metallic agent used to produce its colour was so dense, all but the thinnest films of ruby were opaque. To obtain sufficient translucency, either the glassmaker had to suspend striations of ruby in a clear glass, thereby creating the "streaky rubies" of the early 13th century, or the glass was "flashed"; that is, clear glass while still pliant was dipped into molten coloured glass, thus coating its surface with a thin film of colour. Detailed effects, unhindered by intricate leading, could then be achieved by grinding away portions of this coloured film, first on ruby glass and then on other colours deliberately "flashed" for this purpose. To these colours could now also be added the silver stain in tones of yellow ranging from the palest canary tint to a deep fiery amber, depending on how heavily the stain was applied and how thoroughly it was fired. The whole gamut of more or less translucent tonalities that could be created with vitreous enamel were also used. Taken altogether, these techniques when used in combination represented a considerable liberation of stained glass from what was increasingly considered to be the "tyranny" of the lead line.

The technique of grinding flashed glass was first practiced in the late 13th and early 14th centuries; one of the earliest extant examples is in the church at Mussy-sur-Seine in France, where the windows have a blue groundwork covered all over, or diapered, with ruby roses with white centres, each rose being a single piece of glass. This type of work, however, was not common until the 15th and 16th centuries.

Later developments. At the end of the 15th century a whole new range of vitreous enamels was developed, and by the middle of the 16th century the technique of painting in enamel colours on glass began to be of major importance. In this method, granulated coloured glass of the desired colour is mixed with a flux of clear ground glass and fired onto the surface of the glass. Enamel painting was not altogether successful either technically or



Silver salt staining.
"Annunciation to the Shepherds," English 14th-century stained-glass window in which silver salts have been used to stain the glass shades of yellow, and the reds are "streaky ruby" glass. In the Victoria and Albert Museum, London.
By courtesy of the Victoria and Albert Museum, London; photograph, J.R. Freeman & Co. Ltd.

aesthetically, since the colours thus created were translucent rather than transparent, generally pallid, and of uncertain durability. Political disturbances in the mid-17th century created a scarcity of coloured glass throughout Europe, and gradually the traditional use of coloured glass was replaced by the new technique of painting in enamel colours on clear glass.

Modern techniques

Between the 16th and 20th centuries the developments in the craft of making stained-glass windows were purely utilitarian. In the 16th century the diamond glass cutter was invented, and in the 18th century hydrofluoric acid was introduced as a means of etching flashed glass. In the 19th and 20th centuries, gas and electric kilns and soldering irons were used, as were plate-glass easels upon which stained-glass panels could be temporarily mounted for painting before they were leaded. The largest palette of glass—the widest range of colours, textures, and thicknesses that the art has ever known—was also developed in the 20th century. Contemporary technical innovations include the slab glass and concrete windows developed in France around 1930, where glass set in concrete provides an alternative to leading. In the mid-20th century such experimental techniques as bonding glass to glass with transparent resin glues were developed. Measured purely by technical standards, contemporary stained glass has never been rivalled in its versatility as an instrument of artistic expression.

SUBJECT MATTER

In the Middle Ages ecclesiastical art was primarily didactic. The subjects painted in the windows played an important part in the expounding of the Scriptures and the glorification of the church and its saints.

Iconographic programs

The iconographic program of medieval stained-glass windows for ecclesiastical buildings is a product of several factors. To begin with, the cruciform plan of the churches themselves created four focal areas. Each area, by its architectural form and orientation to the sun, tended to elicit the development of certain subjects or types of subject. In Chartres, for example, the five central windows of the choir clerestory and the north rose window are consecrated to the Virgin, the south rose window to the glorification of Christ, and the west rose window to the Last Judgment. In Bourges Cathedral the huge figures of the Apostles in the south clerestory are paired off against the prophets in the north clerestory, the representatives of the New Testament thereby receiving the full light of the sun and their Old Testament counterparts the more crepuscular light of the north sky. The great rose windows, whose circular form is itself cosmological in its implications, are invariably devoted to cosmological themes: the Last Judgment, the Apocalypse, the cycle of the year as expressed in the signs of the zodiac, the glorification of Christ and of the Virgin as the rulers of heaven. On the other hand, one of the reasons that the theme of the Jesse tree remained popular throughout the Middle Ages was that it lent itself to such a rich variety of ornamental treatments. And finally there was the will of the donors of the windows, whose personal preferences determined the subjects of many excellent works that clearly cannot be related to any comprehensive iconographic program. Some idea of the scope of these medieval enterprises can be indicated by the fact that Chartres, by no means the largest of the cathedrals, contains more than 27,000 square feet (2,518.14 square metres) of stained glass, in 176 windows. Of the 64 windows on the lower level, all but a few are medallion windows, which contain anywhere from 20 to 30 or more separate pictorial compositions; and the three rose windows, each more than 40 feet (12 metres) in diameter, are vast composite creations. The work of at least nine separate master designers has been distinguished in the windows of the cathedral, which was completely glazed in less than 40 years, between about 1203 and 1240.

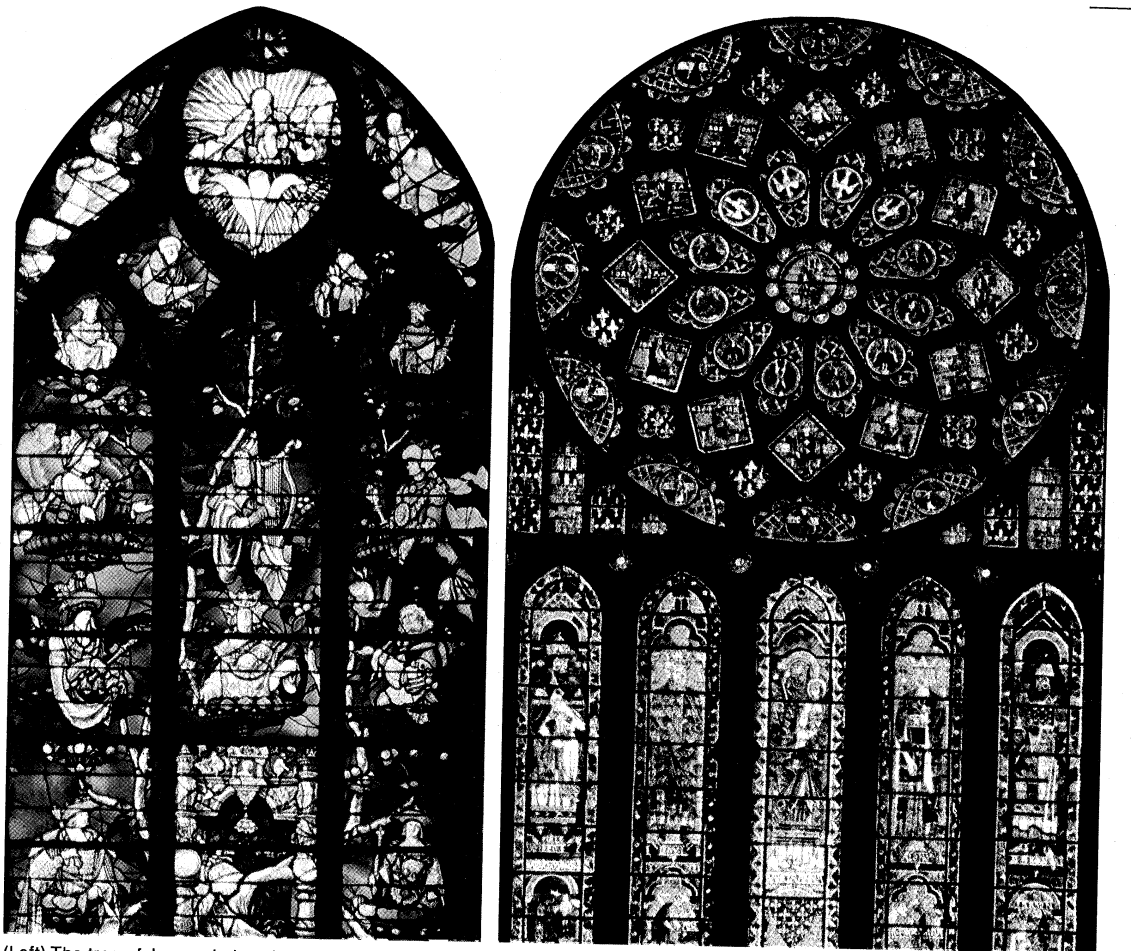
It must be assumed that clerics supplied the master glazier with a program to which he had to conform. A 12th-century manuscript in the British Museum contains a series of circular drawings illustrating the life of St. Guthlac. These drawings might have been intended as a model for a glazier, but the scenes could equally well have been expressed in wall paintings, sculpture, or metalwork. There is more complete knowledge for the later Middle Ages. The glazier was given written instructions from which to prepare provisional sketches that were submitted for the patron's approval before being redrawn in actual size to form the final cartoon. The provisional sketch was known as a *vidimus* (literally, "we have seen"). One example of such written instructions is the program for a window given by Henry VII to the Grey Friars Church at Greenwich, England.

There is ample evidence to show that by the 14th century it was the practice of glaziers to have a stock of finished cartoons, executed on parchment or paper, which could be adapted for different glazing schemes. That these cartoons were used and reused over a long period can be deduced from the will of a York glazier, who died in 1450, in which he bequeathed to his son all his cartoons.

It is evident that in the later Middle Ages the master glazier's workshop was a highly organized enterprise, capable of producing various classes of designs, according to the expense his patrons were prepared to incur. Although the donor, cleric or layman, exercised considerable influence over the choice of subject and its manner of representation, the finished design was essentially the creation of the master glazier. The latter was often an artist in his own right, expressing in the formal language of his own technique the artistic aspirations of his time. (R.So.)

PERIODS AND CENTRES OF ACTIVITY

Early developments. The evolution of the stained-glass window was a slow process. Both texts and excavation testify to the existence of stained-glass windows before the 12th century, but the textual references are too brief and nontechnical to give any clear picture of how the art evolved. The writings of the Fathers of the Latin Church, Lactantius (*c.* AD 240–*c.* 320), Prudentius (AD 348–after 405), and St. Jerome (before 420), mention coloured glass windows in the early Christian basilicas. The 5th-century poet Sidonius Apollinaris described glazed windows in Lyon, France. Pope Leo III (795–816) is recorded to have provided windows of different coloured glass for St. Paul's basilica at Rome. Glazed church windows were widespread in pre-Carolingian Europe in the wealthiest establishments: the Cathedral of York in England was glazed as early as 669. On the site of the Abbey of Monkwearmouth in Sunderland, England, a number of pieces of window glass dating from the late 7th century were found. Coloured green, blue, amber, and red, the edges of several pieces were grozed, or cut for fitting into a window.



(Left) The tree of Jesse window, by the Le Prince family, *c.* 1522–25. In the church of Saint-Étienne, Beauvais, France. (Right) The north rose window, dedicated to the Virgin, and five lancets, with Melchizedek, David, St. Anne holding the baby Mary, Solomon, and Aaron, *c.* 1230–35. In Chartres Cathedral, France.
Left) Jean Roubier—EB Inc., (right) Promophot—Ziolo

In form these early medieval windows varied considerably: the actual window openings were at first filled with thin sheets of marble, alabaster, gypsum, or even wooden boards, which were pierced with holes, coloured glass being inserted into these holes. In addition to glass, other materials were used for the same purpose, thin strips of alabaster set in bronze frame being not uncommon. This form of design, called a "mosaic" window, persisted even in western Europe into the Romanesque period, and 11th-century examples are to be found in Italy in the Cathedral of Torcello near Venice and in the Church of S. Miniato at Florence.

Early leaded windows

Leading was possibly used to hold together the pieces of glass in a window opening contemporaneously with the above early methods. Leading that may have been used in window glazing dating from the 4th century has been uncovered in excavations. The earliest example of a leaded window design was a small panel (destroyed in 1918) in the church at Séry-les-Mézières, northwest of Reims in France, probably of the 9th century. It appears certain that, as at Séry-les-Mézières, many of these early windows contained coloured glass arranged in comparatively simple decorative designs, with little use of the painted design.

There is no documentary evidence even to suggest the existence of pictorial windows until the 9th century, when several rather vague references testify to the appearance of figures in German and French glass. In the 10th-century history of the Church of Saint-Remi at Reims, it is stated that the windows contained various stories.

The fragments of what may be the earliest pictorial window extant were excavated at Lorsch in Germany. It was possible to reconstruct a head of Christ, which shows some stylistic affinity with Carolingian manuscript paintings and probably dates from the 9th, 10th, or 11th century. The earliest complete pictorial windows extant are those containing five figures of prophets in the Cathedral of Augsburg in Germany, belonging to the beginning of the 12th century.

In Carolingian and early Romanesque architecture the window openings, partly for structural reasons, were small and few in number. Polychrome decoration was naturally concentrated on the large mural areas and the vaults rather than on the windows. The development of late Romanesque and Gothic architecture brought a new emphasis on fenestration and openness. It was then that pictorial windows of stained glass became a major art form and in northern Europe the most important single element in church decoration.

Although the pictorial stained-glass window is normally regarded as the invention of and indigenous to western Europe, where its development can be followed with reasonable coherency from the beginning of the 12th century onward, there is still much that is obscure about its earlier evolution. The discovery in Istanbul of stained-glass windows, apparently deriving from a tradition independent of that in the West and datable before 1136, adds to the complexity of the fragmentary evidence already cited.

12th century. It is probable that many of the early stained-glass windows of the 12th century were of the single figure, monumental in character like those examples in the cathedrals of Augsburg and Canterbury or the well-known Virgin and Child known as "La Belle Verrière" at Chartres. The most important feature of the 12th century, however, was the development of the narrative window, consisting of a series of medallions painted with pictorial subjects. This type of window was, so far as is known, first used extensively between 1140 and 1144 at the Abbey of Saint-Denis near Paris. A secondary but

significant development of the second half of the century was the use of allover decorative patterns, or diapers, on the groundwork adjacent to the figures. This design device was probably more common at first in Germany than elsewhere, and an early example is in the Jesse window (*c.* 1170–80) at Frankfurt am Main, now in the city's Städelsches Kunstinstitut.

France. By the 12th century the production of stained-glass windows in northern Europe was considerable, and regional schools begin to be discernible, especially in France, Germany, and England. In France, a number of important regional schools of glass painting emerged, one of the earliest of which was in the west. The most important works of this group include the Ascension window (*c.* 1145) in Le Mans Cathedral and the Crucifixion window (*c.* 1165) in Poitiers Cathedral. In the northeastern region of Champagne appeared another quite distinct group, whose best work is found in the Redemption and St. Stephen windows (*c.* 1150–60) in the cathedral at Châlons-sur-Marne, together with the very important later windows (*c.* 1190) at Saint-Remi in nearby Reims, whose stately figures indicate that Romanesque monumentality has already begun to be tempered by the less austere, less rigorously formal mode of the Gothic.

Regional schools of glass painting in France

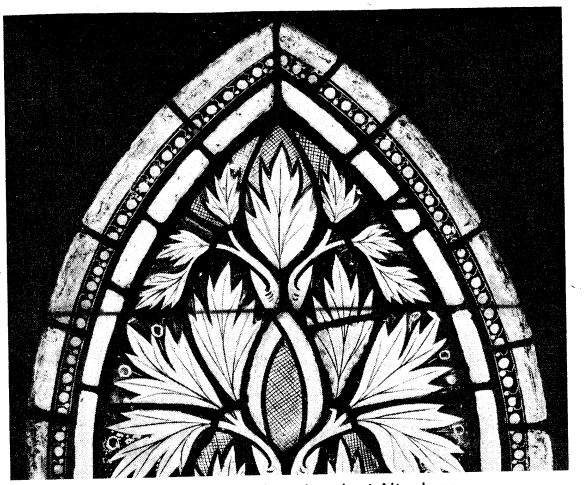
The most important workshop in the Île-de-France region around Paris was connected with the rebuilding of the choir of the Abbey of Saint-Denis. Only fragments of these windows are left, but the three windows (*c.* 1150–55) of the west facade at Chartres are later products of the Saint-Denis workshop and are a summation of all that is most uniquely Romanesque in stained glass.

The stylistic antecedents of these schools are difficult to pinpoint. The Saint-Denis–Chartres group has certain similarities to north French manuscript paintings that are not precisely dated, and the problem is further complicated by Abbot Suger's recording that the glaziers employed at Saint-Denis came from many different regions. The strongly Romanesque character of the Le Mans Ascension window, its general composition, and the particular stylization of drapery forms is similar to earlier manuscript paintings from western France. The Crucifixion window at Châlons-sur-Marne, on the other hand, has precedents in general arrangement in Ottonian manuscript painting and is also closely related in style and composition to the contemporary Mosan school of metalwork centred in the valley of the Meuse River. The similarities between the two are so marked that it is not impossible that the artist worked in both mediums.

Germany. There is less 12th-century glass extant in Germany than in France. The outstanding example of German stained glass of the first half of the century is the series of five prophets (*c.* 1125) in the Cathedral of Augsburg. These hieratic figures have the monumentality of design, rigidly frontal and schematic, characteristic of Romanesque art. The bold use of ruby, green, yellow, and violet glass is completely alien to contemporary French developments. In the second half of the century, art in northern Europe generally, and perhaps more so in Germany, was influenced by Byzantine models. An example is the "Moses and the Burning Bush" window now in the Städelsches Kunstinstitut at Frankfurt am Main or the Magdalen (*c.* 1170) from the church at Weitensfeld, near Klagenfurt, in Austria.

England. England has only fragmentary remains of 12th-century glass. The nave clerestory windows in York Minster contain some reused panels from a series of narrative windows, one of which depicted the life of St. Benedict (*c.* 1140–60). Another panel, a single figure of a king from a tree of Jesse, shows some affinity in style with the glass at Saint-Denis and Chartres but is probably later in date (*c.* 1190). The outstanding survival from the end of the century is the splendid series of figures representing the descent of Christ from Adam, made for the choir clerestory windows (*c.* 1178–1200) of Canterbury Cathedral, which resemble the "Prophet" windows in Saint-Remi at Reims. Their features show a new humanism, and there is a sense of movement, even tension, in their bodies and draperies, comparable to contemporary English manuscript painting.

12th-century English glass



Grisaille panel from the Cistercian church at Altenberg, Germany, 13th century. In the Hessisches Landesmuseum, Darmstadt, West Germany.
By courtesy of the Hessisches Landesmuseum, Darmstadt, West Germany

13th century. A significant feature of the 13th century was the development of the grisaille window, composed largely of white glass, generally painted with foliage designs, and leaded into a more or less complicated geometric pattern. This type of design was employed partly as a means of introducing a larger amount of light and partly because it was considerably cheaper than coloured glass. The combination of grisaille glass and coloured subject medallions, or figures, however, disrupted the monumental overall unity, which is a feature of a window composed entirely of coloured glass, by allowing the penetration of pure light. This change had an important effect on style; the painted design became more linear and refined, the scale more broken and delicate. Although the combination of grisaille and medallions, or figures, is not unknown in the early part of the century, it is more common in the second half, particularly in France and England.

The movement toward humanism, partly inspired by St. Francis of Assisi and his teaching, was accompanied by a related tendency toward naturalism discernible in the visual arts in the later 13th century. The conventional formalized foliage designs of the 12th and earlier 13th centuries gave place to more natural plant motives of oak, vine, and maple, and these break out of the formal patterns and coil with a more organic natural movement.

France. In France, where surviving material is most extensive, the various regional schools are mostly a natural development of their immediate predecessors. There is no radical change in style or technique during the first quarter of the century. In western France the severe Romanesque style was softened and refined, as seen in the Saint-Vital window (*c.* 1200) at Le Mans Cathedral, the Saint-Martin window (*c.* 1210) at the Cathedral of Angers, and the slightly later windows of Abraham, Lot, and Joseph at Poitiers Cathedral. Another distinct workshop, centred at Lyon and responsible (*c.* 1215–20) for the apse windows of Lyon Cathedral, is characterized by a strong Byzantine influence, particularly in iconography. An important workshop in Champagne had already produced in the late 12th century the clerestory windows of Saint-Remi at Reims that foreshadowed the mature Gothic style, while later works of this atelier can be found in the clerestory windows (*c.* 1235) of Reims Cathedral and the choir clerestory windows in the Cathedral of Troyes. In the Île-de-France and eastern France the situation is complicated by the almost complete loss of the later 12th-century works. The north rose window (*c.* 1200–05) of Laon Cathedral is stylistically related to the contemporary sculptures of the facade and to manuscript painting such as the Ingeborg Psalter (Musée Condé, Chantilly). The work of this atelier is extremely distinguished, with an elegance and purity of style and a knowledge of classical art that transcend most of its contemporaries. The rose window (1231–35) of Lausanne Cathedral in Switzerland

Styles in 13th-century French glass

was made by a wandering artist from Picardy, Peter of Arras, and is related in style and iconography to the Laon workshop.

The most extensive glazing program of the first half of the century was at Chartres Cathedral, a tremendous enterprise that brought together various workshops from different regions. The stylistic interactions between the different workshops resulted, particularly in the second quarter of the century, in a more general similarity of style between the various regional workshops. Contemporary with Chartres are the windows (*c.* 1214–35) of the east end of Bourges Cathedral, while four windows (before 1225) at Sens Cathedral are related in composition and ornament to Chartres and Bourges and also have certain stylistic affinities with the contemporary windows at Canterbury Cathedral.

Considerable activity was also centred in the Paris area during the second quarter of the century. The major monument of the period is the Sainte-Chapelle, which was built in Paris between 1243 and 1248. Forming what amounts to a continuous wall of 50-foot- (15-metre-) high stained glass around three sides of the chapel, it contains the most extensive narrative cycle ever produced in this medium, numbering 1,134 scenes in 15 windows. Stylistically, these windows are closely related to Parisian court art of the same date, and their influence can be seen in the later windows at Le Mans Cathedral (about 1254–60) and in the Cathedral of Tours (1245/55–76).

England. The only extensive remains of 13th-century glass in England are found at Canterbury Cathedral, where the 12 Theological windows were produced in about 1200 and the windows relating to St. Thomas Becket about 1200–30. Lincoln Cathedral retains impressive fragments of a series of windows made between 1200 and 1220, but it is impossible fully to appreciate either of these series without making comparison with French work. There are important similarities between Canterbury and Sens, on the one hand, and probably between Lincoln and Paris, on the other. The glaziers, however, were probably English with a close acquaintance with French models.

Germanic countries. Thirteenth-century stained glass in Germanic countries, however, was comparatively uninfluenced by French models. It is more turbulent in design, with agitated draperies, expressive faces, and a complicated ornamented character, particularly in the backgrounds. There were many distinct regional schools, among which Cologne was an important centre. The full-length figures of saints and the Legend of St. Kunibert window (*c.* 1220–30) in the Church of St. Kunibert at Cologne have elaborate geometrical frames around the figures and scenes that are without parallel in French art. These frames are a typical feature of German work; they occur again in later work of this school in the St. Nicholas window (*c.* 1240–50) at Büchen, west of Hamburg, and elsewhere. Another workshop, which produced the Jesse window (*c.* 1225) in the Cathedral of Freiburg im Breisgau, shows a marked affinity with the slightly earlier enamel and metalwork produced by Nicholas of Verdun and his circle (active 12th–13th century), while the remains of a group of windows (*c.* 1230) made for the Franciscan church at Erfurt constitute a particular group strongly indebted to Byzantine models. It appears that after the completion of the Erfurt windows, this workshop, or at least some of its members, went to Assisi in Italy and also to Gotland in Sweden. The most outstanding glass of the mid-century is a related series of windows in the cathedrals of Naumburg, Strasbourg, and Frankfurt. The Naumburg window of Holy Knights and Virgins can be contrasted with the delicate mannered style of Parisian court art.

Features of 13th-century German glass

The earliest Italian examples of stained glass were the three windows executed by German craftsmen of the Erfurt school in the apse of the Upper Church at Assisi between 1230 and 1240. At the end of the 13th century, native designers and craftsmen produced windows. The oculus window (*c.* 1288) in the apse of Siena Cathedral has been attributed to Duccio di Buoninsegna (died 1318 or 1319), and the style of the painter Guido da Siena (active *c.* 1250–75) seems evident in the Madonna and Child panel of the Sanctuary of the Madonna della Grotta near Siena.

Early examples of stained glass in Italy

Early 14th century. Stained glass of the first half of the 14th century is everywhere distinguished by an insouciant fairy tale quality and a langorous charm sometimes tinged with pathos. Regional differences, however, persisted—the gentle reserve and earthy lyricism of the English; the virtuoso painting and exquisite drolleries of the Norman-French; and the full green-, gold-, and russet-dominated palettes of the German windows. The full flowering of the Gothic style side by side with the beginnings of stylistic developments that were to culminate in the Renaissance characterized the aesthetic nature of the early 14th century. The new movement toward the representation of volume and spatial depth, by means of modelling and perspective, had its origins in Flemish and Italian painting. That the glass painter was quickly influenced by this new style is seen, for example, in the St. Anthony window in the Lower Church at Assisi. North of the Alps the earliest extant manifestation of this new interest in perspective and modelling, based on Italian models, occurs in the chancel windows (1325–30) of the Habsburg expiatory church at Königsfelden, near Brugg, Switzerland. The knowledge of Italian models spread quickly and extensively and can be seen in France in the windows (1325–39) at Saint-Ouen in Rouen and those made about 1330 at Évreux Cathedral. In the Germanic lands proto-Renaissance spatial illusionism influenced the transept windows at Augsburg Cathedral and the east window (*c.* 1340) of Vienna Cathedral, while the earliest remaining example in English glass painting is probably the nave windows (*c.* 1330–35) of Stanford-on-Avon in Northamptonshire.

In the early 14th century, the third dimension in canopies was still highly untheoretical and largely governed by considerations of pure design. The practice of representing a figure beneath an architectural canopy was an established convention of the 13th century, particularly used in clerestory windows. In earlier examples the canopy plays a comparatively unimportant part in the total design, but by the end of the 13th century, although still two-dimensional, it had become more elaborate and is an important ornamental feature of the windows of Merton College, Oxford. In German and Austrian windows, the canopy work is often very elaborate and complex in its spatial organization; examples are found at Vienna Cathedral (*c.* 1340) and Erfurt Cathedral (*c.* 1360–70).

Currents in France and England

The art of glass painting, however, did not respond equally to these new influences emanating from Italy. It subdivides itself into two groups, of which France and England together make up one, characterized by its resistance to Italian influences. The use of perspective was purposefully restrained, so that the essential overall surface unity of the design was not violently upset, for the use of flat, patterned diaper grounds effectively counterbalanced the suggestion of spatial effect. In the first half of the century the most important work in France is found in the region of Normandy, especially in the choir windows (*c.* 1330) of Évreux Cathedral and those dating around 1325 to 1339 at Saint-Ouen in Rouen. The English glaziers of this period were extremely prolific, with Oxford, Coventry, and York as important regional centres. The nave windows of York Minster were made during the period 1300 to 1338 and are the largest single enterprise of this period in England. It appears probable that some of the later glass at York (*c.* 1350–70), now distributed over the windows of the choir clerestory, was the work of an imported French glazier, probably from Rouen. A flourishing school in western England, whose best work is found at Wells Cathedral and Eaton Bishop in Herefordshire, shows some affinity with German work. The best French and English work, however, has a lightness of colour and graphic refinement that is enhanced by an extensive use of yellow stain. After about 1300 the geometric grisaille glass gave way to simpler diamond-shaped pieces, painted with delicate trails of foliage and leaded together to give the effect of trellis work.

The second group, which might be termed "Germanic," as it embraces Germany, Bohemia, and Austria, displays

Three-dimensional style of the Germanic countries

a much more three-dimensional style; the colours are deeper and more saturated, the compositions are more complex, both on the surface and in depth, and the canopy designs particularly are often very complicated essays in perspective; the figures are shorter in proportion and their volume more accentuated. It is brusque, almost harsh, contrasting strongly with the elegance of French and English work. All of these traits can be seen, for example, in the panels (*c.* 1350) from Strassengel, near Graz, Austria, now in the Victoria and Albert Museum, London; the earlier windows (*c.* 1350–60) at St. Maria-am-Gestade in Vienna; and at Erfurt Cathedral (*c.* 1360–70). A particular trait of this Germanic group, of which Erfurt is a good example, is a tendency to extend a single composition across the main lights of a window, ignoring the natural divisions of the stonework. The reason for this is partly architectural: the window lights are comparatively much narrower and taller than those in French or English windows.

The International Style

Late 14th and early 15th centuries. The arts from about 1380 to about 1430, which are conveniently grouped together under the title of the International Style, belong essentially to an era of court art inspired by the patronage of kings, the nobility, and the higher orders of ecclesiastics. Surviving windows are extensive, and interactions between the various centres of patronage, complicated by family alliances and the exchange of works of art and artists, are particularly complex. The various national and regional styles can still be distinguished in glass painting, but there is a general tendency toward a mannered, extremely sophisticated elegance of style, sometimes verging on the precious, combined with an interest in portrait realism.

If the genius of the 13th-century stained glass lay in its epic sense of monumentality and that of the early 14th century in its warmth of human feeling, that of the late 14th and early 15th centuries is far more difficult to characterize in a phrase. The style of glass painting became at once more corporeal and more introspective. Figures are rendered with far more attention to individual human traits yet are akin in their majesty to the great prophets in the 13th-century windows. The figures painted between 1384 and 1392 by Hermann von Münster, for the west window of Metz Cathedral in France, and those painted around 1400 by Thomas of Oxford for the chapel of William of Wykeham's College at Winchester, England, seem both newer and older than their immediate predecessors—newer in that they are beginning to bear the signature of a personal style, older in that they recall the grandeur of an earlier, essentially collective expression that now seemed everywhere to be giving way to arid cameo-like refinements on the one side and an earnest rusticity on the other.

The glass painting of this period is of high quality. The most significant examples in France are the "Royal" windows (*c.* 1395) at Évreux Cathedral. One of the most ambitious works of the period is the great east window of York Minster, made between 1405 and 1408 by John Thornton of Coventry.

In Germany and Austria the International Style brought about a softening and refining of the "plastic style" of the earlier part of the century. In Austria an important atelier associated with the court produced the window (*c.* 1386–95) in the church of St. Erhard in der Breitenau and later the series of windows (*c.* 1400) at Viktring, and for the Freisingerkapella at Klosterneuburg (*c.* 1410). Germany contains a large amount of work of this period. At Erfurt Cathedral the window given by Johann von Tiefengruben (*c.* 1400) is lighter in tone and has a more refined pictorial style than the earlier windows there. The series of windows at Rothenburg, which were probably made about 1400, should be noted, together with the cycle of windows (*c.* 1420–30), for the Bessererkapelle at Ulm Cathedral, which show an admirably refined technique.

15th and 16th centuries. The period 1430–1550 saw not only the decline of the Gothic style and the establishment of the new Renaissance style but also the beginning of the transformation of the art of glass painting from a significant means of artistic expression into a hybrid art form: the translucent emulation of fresco and easel painting.

The International Style continued to influence glass painting during the first half of the 15th century but to a lesser degree. Its mannered elegance and extravagant costumes can still be seen in France in the two rose windows (*c.* 1440) at Le Mans Cathedral and also the rose windows (1441–42) at Angers Cathedral. In England this aesthetic is continued in the east window (*c.* 1423–39) of the priory in Great Malvern (Worcestershire), the chapel windows (1441–47) of All Souls College at Oxford, and in the windows made by John Prudde, the king's glazier, for the Beauchamp Chapel of St. Mary's in Warwick, which were commissioned in 1447. The work of Germanic glass painters provides outstanding examples, particularly the window (after 1424) from St. Lambrecht, now in the museum at Graz, Austria, and the charming Alsatian window of St. Katherine (*c.* 1425–50) at Sélestat, France.

The Flemish style

There were, however, important new influences affecting the styles of glass painting. In northern Europe during the first half of the 15th century a flourishing school of painting emerged in Flanders. Although the origin of the Flemish style is partly to be found in the International Style, a more realistic manner of representation and a detailed awareness of actuality developed that is almost the antithesis of the mannered sophistication and the essentially unrealistic world of the International Gothic.

It was impossible for the glass painter to participate fully in the new realism. Glass is a translucent material; the passage of light through it is alone sufficient to create a feeling of unreality. Furthermore, although the panel or mural painter could use the new discoveries of linear perspective to heighten the sense of reality, the glass painter had to contend with the presence of the leading line, which emphasizes the surface plane. This creates a tension and a sense of ambiguity between the actual surface and the illusion of depth. Attempts were made to resolve the problems, but with little success; one result was that the lead line became increasingly divorced from the design instead of being an integral part of it. The increasing use, from the mid-16th century onward, of vitreous enamel pigments had the effect of accelerating this process.

The full flowering of the Renaissance style in Italy and the intense interest in classical art that it stimulated resulted, around 1500, in a profusion of ornamental details, borrowed from the formal language of classical art, in contemporary window designs. In addition, the glaziers at this time often drew inspiration from contemporary engravings, particularly those of Albrecht Dürer.

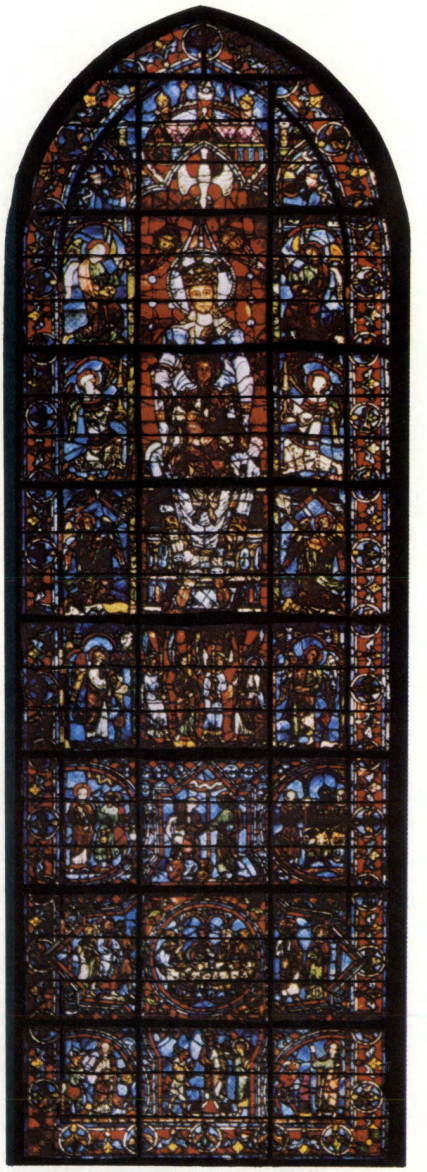
The period 1450–1550 saw no decline in demand for stained glass. One of the most productive and influential areas at this time was Flanders. The six windows (*c.* 1475–1500) of the church of St. Gommaire at Lier and the Virgin window (*c.* 1482) at Anderlecht, although restored, have original portions of excellent quality and close affinity with contemporary Flemish painting.

English stained glass in the 15th century becomes more intimate, more anecdotal. It is less a cathedral art than an art of parish churches and is addressed less to an assembled ecclesia than to the individual believer. At its best it achieves a quiet intensity in the woodcut-like figures in East Harling church (Norfolk), a pathos in the Long Melford church (Suffolk) "Pieta," and a sheer power of expression in the Clavering church (Essex) "Martyrdom of St. Catherine."

The trading and political links between Flanders and England in the second half of the 15th century encouraged the influx of Flemish works of art and artists into England. A number of Flemish glaziers established themselves in Southwark, London, and some even assumed English names. The early 16th century was marked by a series of disputes between the London Guild of Glaziers and the foreigners. The latter were particularly patronized by the court and the more wealthy merchants. The two outstanding monuments of this imported style are the windows (*c.* 1480) of Fairford church (Gloucestershire) and the windows (1515–31) at King's College Chapel,



Sainte-Chapelle, the Upper Church, 1243–48, Paris.



"Virgin and Child," known as "La Belle Verrière," mid-12th century. In Chartres Cathedral, France.

Stained glass of the 12th and early 13th centuries

"The Assumption of the Virgin," *c.* 1145. In Le Mans Cathedral, France.



"Moses and the Burning Bush," by Master Gerlachus, *c.* 1170–80. In the Count Albrecht von Kantz Collection, Lünen, West Germany.



Plate 1: (Top left) Ciccione—Rapho Guillumette, (top right) Promophot—Ziolo, (bottom left) Giraudon, (bottom right) Kalle Aktiengesellschaft, Wiesbaden-Biebrich, West Germany



"Christ Risen," by Paolo Uccello, 15th century. In Sta. Maria del Fiore, the cathedral of Florence.

"Jesus and the Centurion," by Guglielmo de Marcillat, *c.* 1520. In Arezzo Cathedral, Italy.





"Nativity," 1325–30. In the Hapsburg Expiatory Church, Königsfelden, Switzerland.

Stained glass of the late 13th through the 16th centuries

"St. Stephen," 1289–96. In Merton College, Oxford.



Plate 2: (Top left, bottom left) SCALA, N.Y., (top right) Hans Hinz, Basel (bottom right) Giraudon

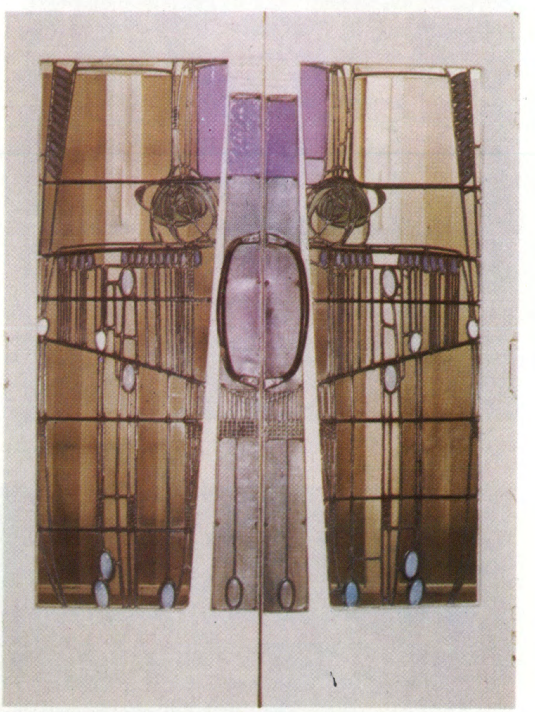


"David," detail from the Vyner window designed by Sir Edward Burne-Jones and executed by William Morris, *c.* 1870. In Christ Church, Oxford.



"Jonah Window," by Abraham van Linge, 1630–40. In Christ Church, Oxford.

Stained glass of the 17th through the early 20th centuries

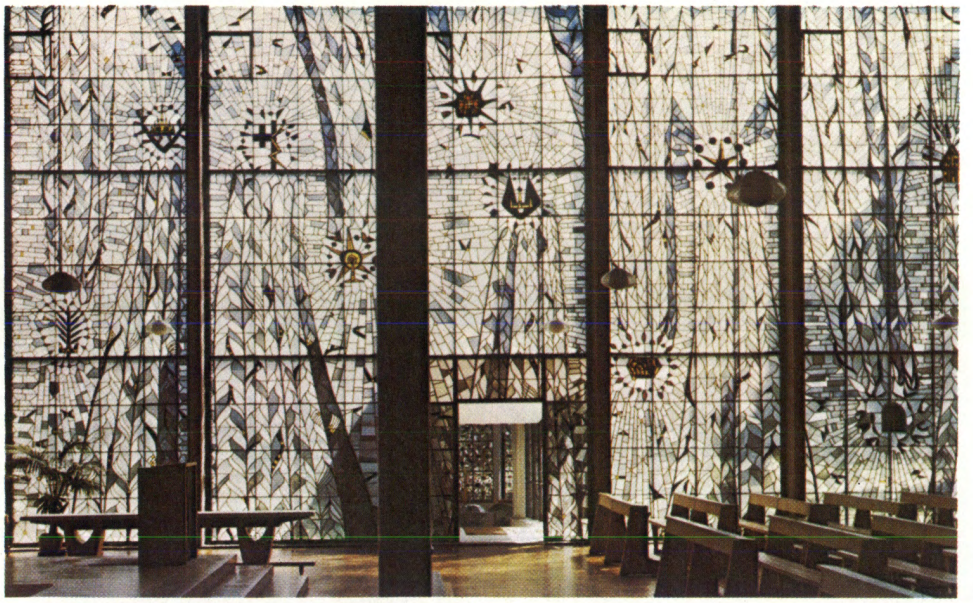


Doors of the Willow Tea Rooms, by Charles Rennie Mackintosh, 1904, Glasgow.



Heckscher House window, by Louis Comfort Tiffany, early 20th century. In the collection of Mr. and Mrs. Hugh F. McKean, Winter Park, Florida.

Plate 3: By courtesy of (bottom left) Mr. and Mrs. Hugh F. McKean, Winter Park, Florida, (bottom right) the House of Fraser, Glasgow; photographs, (top left, top right) Thomas-Photos, Oxford, (bottom right) Studio Swain

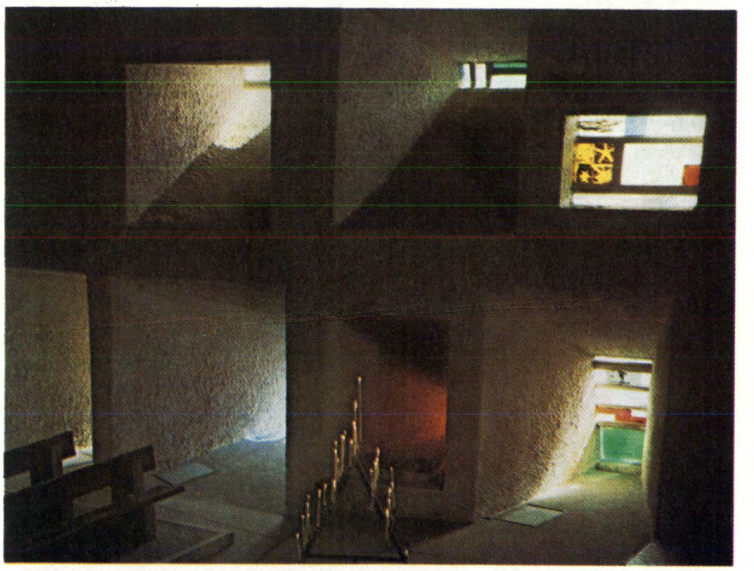


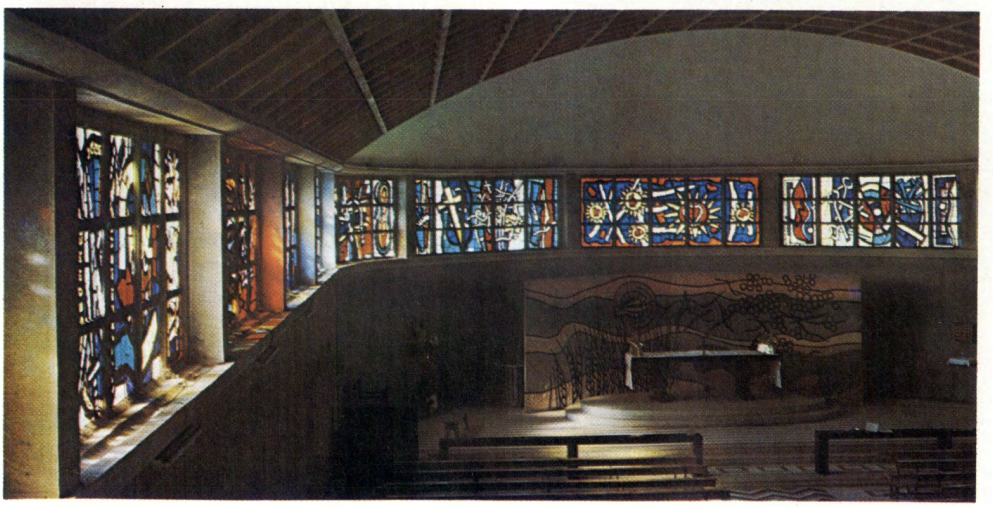
Stained glass of the second half of the 20th century

Maria Königin, Cologne-Marienburg, West Germany, windows by Dominikus Böhm and Heinz Bienefeld, 1954.

(Left) The Tribe of Benjamin, one of the windows of "The Twelve Tribes of Israel," by Marc Chagall, 1960–61. In the Synagogue of the Hadassah-Hebrew University Medical Center, Jerusalem. (Right) Notre-Dame-du-Haut, Ronchamp, France, by Le Corbusier, 1955.







Sacré-Coeur, Audicourt, France, windows by Fernand Léger, 1951.

Plate 4: (Top, centre right, bottom) GEKS, (centre left) © Hadassah Medical Relief Association, Inc., 1961; all rights reserved for Marc Chagall/Andre Sauret

Cambridge, many of which were probably designed by the Antwerp painter and engraver Dirk Vellert.

New realism in French glass painting

The spread of the new realism can be traced in French glass painting. The interactions with Flemish glaziers can be seen, for example, in the realism of the windows (*c.* 1430–40) from the Duke of Burgundy's chapel at Dijon, now in the Victoria and Albert Museum, London, which are the work of Flemish glaziers, and also the Jacques Coeur window (*c.* 1450) at Bourges Cathedral, which is markedly Flemish in style. Native French glaziers were extremely prolific at this time, and Normandy and particularly the city of Rouen contain an incomparable display of windows produced by a large number of distinctive workshops. The leading figures of the first part of the 16th century were Arnoult of Nijmegen (*c.* 1470–1540) and the Le Prince family at Beauvais. Arnoult of Nijmegen worked in both Flanders and France. His most important works are the windows he executed between 1490 and 1500 in Flanders for Tournai Cathedral and the Jesse window (1506) in Saint-Godard at Rouen, which is one of the most impressive examples of glass painting of the period. The works by the Le Prince family are equally impressive, particularly the tree of Jesse window (*c.* 1522–25) in Saint-Étienne at Beauvais.

In Germany the later 15th century is dominated by the prolific activity of Peter Hemmel von Andlau (active 1430–1500) and his workshop. Examples of his work were done for St. Wilhelm's Church in Strasbourg.

In Italy a remarkable ensemble of stained-glass windows was created in the first half of the 15th century for the cupola of Sta. Maria del Fiore in Florence by some of the greatest masters of early Renaissance art: Ghiberti, Donatello, Castagno, and Uccello. Domenico Ghirlandajo created in the second half of the century windows for the Florentine church of Sta. Maria Novella. The late 15th and early 16th centuries are mainly associated in Italy with the name of Guglielmo de Marcillat (1467–1529), a Frenchman whose works display a thorough mastery of technique. His finest windows are at Arezzo Cathedral. The building of Milan Cathedral caused an important school of glass painting to develop there, and the work of Conrad Munch, a German from Cologne, and Nicolo da Varallo, carried out during the second half of the 15th century, is noteworthy. The Milan school continued in full activity during the 16th and 17th centuries.

In Spain, especially at Seville, Leon, and Avila, there are some good 16th-century windows; they are in all cases the work of imported Flemish glass painters.

17th and 18th centuries. The spatial illusions of Baroque paintings were beyond the limitations imposed by the stained-glass medium. The glass painter of the 17th and 18th centuries found himself reduced to completing the cycles of stained-glass windows in medieval churches or to creating contemporary art for an architecture with no artistic affinity with traditional stained glass. The most interesting development in the late 16th and early 17th centuries was the intimate and portable heraldic panel, which became fashionable to hang in domestic windows, particularly in Switzerland, the Low Countries, and Germany. These panels, seldom more than two feet high, are the glass painter's showpieces; they complete the divorce between stained glass and architecture.

The fashionable heraldic panel

Painting glass with vitreous enamels in the 17th and 18th centuries led to the final decline of the art of stained glass. In the St. Janskerk windows at Gouda, Holland, painted by the brothers Wouter and Dirk Crabeth at the end of the 16th century, and in the works of Abraham and Bernard van Linge, made between 1620 and 1640, the realization of the window as a translucent canvas painting is complete. Abraham van Linge's windows painted in 1630 to 1640 for Christ Church Cathedral at Oxford are an excellent example of the destruction of the lead line as an integral part of the design. The leading simply holds together the square sheets of glass: the effect is the same as looking at a picture set behind a rectangular grid. This type of design was continued by English glass painters such as Henry Gyles and the Price and Peckitt families, all of York, and by such as Francis Eginton and Sir Joshua Reynolds, who designed in 1778 the west window for New College Chapel, Oxford.

19th century. The Gothic revival that came as an offspring of the Romantic movement of the late 18th and early 19th centuries represents the beginning of a revitalization of the art of stained glass. The revival of interest in Gothic art stimulated an interest in both the technique and history of medieval glass painting. The pioneer figures in this field were E. Viollet-Le-Duc in France and Charles Winston in England. Winston was a lawyer and antiquarian who associated with various London glaziers and, with the technical help of James Powell and Sons, brought about a considerable improvement in the technical quality of coloured glass; and in 1847 he wrote the first comprehensive study of the medium. The experiments were continued by W.E. Chance, who first successfully produced "antique" glass in 1863.

Revival of medieval glass painting

In the first half of the 19th century the styles and methods of the early Gothic period were reconstructed, but without much aesthetic appreciation of medieval art. Much of the work was stereotyped and mass-produced, particularly in Germany, and varied considerably in technical quality. The latter part of the century is dominated in England by Edward Burne-Jones and William Morris. Burne-Jones provided the designs and Morris adapted them to the medium of stained glass. In windows by them the lead line is once again treated as an integral part of the design, as seen, for example, in the windows for Christ Church at Oxford (1874–75 and 1878), Salisbury Cathedral (1879), and Birmingham Cathedral (1897). In the U.S. the works of John La Farge and Louis Comfort Tiffany were influential in creating an American interest in stained glass. Although the style and sentiment of 19th-century work has not been much in favour in the 20th century, the period had great historical significance in the revival of the basic technique of making stained glass.

Art Nouveau designers used stained glass decoratively for making such objects as lampshades and light fixtures, and turn-of-the-century architects increasingly employed stained glass as an integral element in wholly modern architectural settings: Victor Horta in his Hotel Solvay (1895–1910), Brussels; Antonio Gaudí in his Chapel of Santa Coloma de Cervelló (1898–1914) in the Güell Colony (1898–1914) near Barcelona; Charles Rennie Mackintosh in the Willow Tea Rooms (1904), Glasgow; and Frank Lloyd Wright in the Coonley House (1908), Riverside, Illinois, and the Unity Church (1906), Oak Park, Illinois. These windows and panels clearly mark the beginnings of an authentically modern stained glass, despite their strictly ornamental intent.

20th century. Three interrelated creative currents can be discerned in the development of 20th-century stained glass. First, a significant number of architects, following the lead of their turn-of-the-century predecessors and taking advantage of the new systems of fenestration made possible by modern structural engineering, have continued to discover many new ways of using stained glass. Second, especially in post-World War II France, several major easel painters turned their attention to stained glass, infusing it with many new and powerful images. Third, during the 1950s and 1960s Germany produced the first authentic school of stained glass since the Middle Ages, dedicated to exploiting the unique technical and expressive resources of the medium.

Although the bulk of significant 20th-century stained glass belongs to the period after World War II, earlier experiments, especially in France and Germany, suggested the range of possibilities that could be creatively explored. In Auguste Perret's church of Norte-Dame (1922–23) in Le Raincy, near Paris, the entire wall surface of the church becomes a geometric grillwork of coloured glass designed by the Symbolist painter Maurice Denis. In 1930, the Dutch-born artist Johan Thorn Prikker completed a cycle of windows for the Romanesque Church of St. George in Cologne in which lead lines are used with a graphic eloquence and deep smoldering colours with a monumental gravity that have no parallel even in the greatest medieval windows.

Work after World War II

After 1946 there was an unusual burst of activity. In Le Corbusier's Notre-Dame-du-Haut (1952–55) at Ronchamp, northwest of Dijon in France, the massive south wall, 12 feet (2.72 metres) thick at the base and five feet at the top, is dramatically punctuated with a series of crude, yet remarkably effective, stained-glass windows through which shafts of light fairly explode into the church. Simultaneously, in Dominikus Böhm's Church of Maria Königin (1953–54) in Cologne-Marienburg an entire sidewall of the church is conceived as a diaphanous veil of silvery gray stained glass that half reveals and half conceals the parklike grounds outside with equally dazzling effect. In Wallace K. Harrison's First Presbyterian Church (1958) in Stamford, Connecticut, the whole central section of the church is very nearly engirdled with slab glass and concrete. In all of these structures, different as they are in nearly every other respect, stained glass is seen boldly exploited once again not merely for its local ornamental quality but as a major, integral atmosphere-creating element.

The most seminal contributions of the School of Paris painters to the art of stained glass were Henri Matisse's Chapel of the Rosary (1948–52) in Vence and Fernand Léger's windows for the Church of the Sacred Heart (1950–52) in Audincourt. Both are the creation of artists whose manner of working was rather directly translatable into stained glass. It was but a comparatively short step from Matisse's large coloured-paper collages to the disarmingly simple decorative windows in Vence, but the way Matisse used them to create an enchanting play of colour in the chaste white space of the chapel is masterful. And it took the boldly emblematic style of Léger to reveal the true expressive potentialities of slab glass and concrete. A third important work of this period is the long friezelike window created by the sculptor Léon Zack for the Church of Notre-Dame-des-Pauvres (1955) in Issy-les-Moulineaux, remarkable for its daring sequence of colour harmonies and delicate lead line motifs reminiscent of the art of Paul Klee. The stained-glass windows of Braque, Villon, Rouault, Chagall, and Manessier are also noteworthy if less authoritative in their handling of the medium.

In Germany such distinguished prewar church architects as Dominikus Böhm and Rudolf Schwarz and the stained-glass artist Anton Wendling were able to resume careers that had been interrupted by the Hitler era and to set the course for a whole new generation of stained-glass artists, especially in the Rhineland. Inspired by the example of Thorn Prikker, these artists have continued to explore the unique qualities of stained glass—the special refractory properties of opal-flashed antique glass, the graphic potentialities of the lead line, the bold effects of texture and relief that had become possible with slab glass and concrete—and to create a whole gamut of strange brooding colour harmonies the like of which had not been seen in stained glass since the Augsburg prophets. Among the more important works of this Rhenish school of stained glass are Georg Meistermann's windows for the Dom Sepulchur (1957) in Würzburg and his complete ensemble of windows for the 15th-century church of St. Matthew (1964) in Sobernheim; Ludwig Schaffrath's cycle of modern grisaille windows for the cloister (1962–65) in Aachen, his high triple-gabled window walls for the transepts of St. Peter's Church (1964) in Birkesdorf, near Düren, and his powerfully iconic and technically innovative slab- and rod-glass sanctuary window in St. Matthew's Church (1966–67) in Leverkusen; Wilhelm Buschulte's unusually rich colour harmonies in his cycle of nave windows for the Cathedral of Essen (1964) and the choir of the Church of SS. Peter and Paul (1967) in Wegsburg, near Mönchengladbach; and Johannes Schreiter's almost monochromatic Abstract Expressionist windows for the Church of St. Margaret (1961) in Bürgstadt. Trained once again to work of the scale of the cathedral windows and to develop their art in accordance with its own intrinsic potentialities, such artists have been collaborating with some of the best architects in Germany to create the most impressive body of stained-glass windows since the Middle Ages.

BIBLIOGRAPHY. Illustrated monographs include: COMITE INTERNATIONAL D'HISTOIRE DE L'ART, *Corpus Vitrearum Medii Aevi* (1956–), a series intended when complete to document all medieval stained glass extant; J. BAKER and A. LAMMER, *English Stained Glass* (1960), probably the best collection of high-quality detail photographs of medieval stained glass ever published; H.E. READ, *English Stained Glass* (1926); A.C. SEWTER, *The Stained Glass of William Morris and His Circle* (1972); M. AUBERT *et al.*, *Le Vitrail française* (1958), the standard work on French stained glass, although marred by many poor reproductions. This fault is remedied in part by the following works: M. AUBERT, *Stained Glass of the XIIth and XIIIth Centuries from French Cathedrals* (1947); and E. VON WITZLEBEN, *French Stained Glass* (1966). H. WENTZEL, *Meisterwerke der Glasmalerei*, 2nd ed. (1954), is the standard work on German medieval stained glass; E. VON WITZLEBEN, *Farbwunder Deutscher Glasmalerei aus dem Mittelalter* (1965) is the most copiously illustrated recent monograph on German medieval stained glass. S. BEEH-LUSTENBERGER, *Glasmalerei um 800–1900 im Hessischen Landesmuseum in Darmstadt* (1967), an excellent guide to one of the largest museum collections of medieval stained glass; P. WEMBER, *Johan Thorn Prikker: Glasfenster, Wandbilder, Ornamente 1891–1932* (1966); G. MARCHINI, *Italian Stained Glass Windows* (1957); M. STETTLER, *Swiss Stained Glass of the Fourteenth Century from the Church of Koenigsfelden* (1949); and F. ZSCHOKKE, *Medieval Stained Glass in Switzerland* (Eng. trans., 1947).

Books on aesthetic analysis include J.R. JOHNSON, *The Radiance of Chartres: Studies in the Early Stained Glass of the Cathedral* (1965), valuable not only for its somewhat too sweeping criticisms of Viollet-le-Duc but for its original analysis of the effect of the twilight atmosphere of the cathedral upon our perception of its stained-glass windows; and R. SOWERS, *Stained Glass: An Architectural Art* (1965), a thoroughly illustrated analysis of the relation between stained glass and architecture, with numerous photographs of contemporary stained glass.

Stained-glass techniques are discussed in THEOPHILUS (RUGERUS), *Diversarum Artium Schedula* (1847; Eng. trans., 1963), the earliest account of stained-glass window-making techniques, now believed to have been written in the early 12th century; C. WINSTON, *Hints on Glass Painting* (1847), and his *Memoirs* (1865), two books containing some of the most thorough and perceptive analyses of medieval glass-painting techniques ever written; E. VIOLET-LE-DUC, "Vitrail," in *Dictionnaire Raisonné de l'architecture française*, vol. 9 (1868; Eng. trans., "Mediaeval Stained Glass," 1946), still valuable for its many acute observations; and C.W. WHALL, *Stained Glass Work* (1905), a thorough craft manual by one of the leading turn-of-the-century stained-glass artists in England. Recent manuals include E.L. ARMITAGE, *Stained Glass* (1959); P. REYNTIENS, *The Technique of Stained Glass* (1967); and R. and G. MECALF, *Techniques of Stained Glass* (1971).

(R.So./Ed.)

Stalin, Joseph

During the quarter of a century preceding his death in 1953, the Soviet dictator Joseph Stalin probably exercised greater political power than any other figure in history. Stalin industrialized the Union of Soviet Socialist Republics, forcibly collectivized its agriculture, consolidated his position by intensive police terror, helped to defeat Germany in 1941–45, and extended Soviet controls to include a belt of eastern European states. Chief architect of Soviet totalitarianism and a skilled but phenomenally ruthless organizer, he destroyed the remnants of individual freedom and failed to promote individual prosperity, yet he created a mighty military–industrial complex and led the Soviet Union into the nuclear age.

Stalin's biography was long obscured by a mendacious Soviet-propagated "legend" exaggerating his prowess as a heroic Bolshevik boy-conspirator and faithful follower of Lenin, the founder of the Soviet Union. In his prime, Stalin was hailed as a universal genius, as a "shining sun," or "the staff of life," and also as a "great teacher and friend" (especially of those communities he most savagely persecuted); once he was even publicly invoked as "Our Father" by a metropolitan of the Russian Orthodox Church. Achieving wide visual promotion through busts, statues, and icons of himself, the dictator became the object of a fanatical cult that, in private, he probably regarded with cynicism.



Stalin, 1950.
Sovfoto

The young revolutionary. Of Georgian—not Russian—origin, Stalin was born on December 21 (December 9, old style), 1879, as Joseph Vissarionovich Dzhugashvili, son of a poor cobbler in the provincial Georgian town of Gori in the Caucasus, then an imperial Russian colony. The drunken father savagely beat his son. Speaking only Georgian at home, Joseph learned Russian—which he always spoke with a guttural Georgian accent—while attending the church school at Gori (1888–94). He then moved to the Tiflis Theological Seminary, where he secretly read Karl Marx, the chief theoretician of international Communism, and other forbidden texts, being expelled in 1899 for revolutionary activity, according to the "legend"—or leaving through ill health, according to his doting mother. The mother, a devout washerwoman, had dreamed of her son becoming a priest, but Joseph Dzhugashvili was more ruffianly than clerical in appearance and outlook. He was short, stocky, black haired, fierce eyed, with one arm longer than the other, his swarthy face being scarred by smallpox contracted in infancy. Physically strong and endowed with prodigious willpower, he early learned to disguise his true feelings and to bide his time; in accordance with the Caucasian blood-feud tradition, he was implacable in plotting long-term revenge against those who offended him.

In December 1899, Dzhugashvili became, briefly, a clerk in the Tiflis Observatory, the only paid employment that he is recorded as having taken outside politics; there is no record of his ever having done manual labour. In 1900 he joined the political underground, fomenting labour demonstrations and strikes in the main industrial centres of the Caucasus; but his excessive zeal in pushing duped workers into bloody clashes with the police antagonized his fellow conspirators. After the Social Democrats (Marxist revolutionaries) of the Russian Empire had split into their two competing wings—Menshevik and Bolshevik—in 1903, Dzhugashvili joined the second, more militant, of these factions and became a disciple of its leader, Lenin. Between April 1902 and March 1913, Dzhugashvili was seven times arrested for revolutionary activity, undergoing repeated imprisonment and exile. The mildness of the sentences and the ease with which the young conspirator effected his frequent escapes lend colour to the unproved speculation that Dzhugashvili was for a time an *agent provocateur* in the pay of the imperial political police.

Work with Lenin

Rise to power. Dzhugashvili made slow progress in the party hierarchy. He attended three policy-making conclaves of the Russian Social Democrats—in Tammerfors (now Tampere, Finland; 1905), Stockholm (1906), and London (1907)—without making much impression. But he was active behind the scenes, helping to plot a spectacular holdup in Tiflis (now Tbilisi) on June 25 (June 12, O.S.), 1907, in order to "expropriate" funds for the party. His first big political promotion came in February (January, O.S.) 1912, when Lenin—now in emigration—co-opted him to serve on the first Central Committee of the Bolshevik Party, which had finally broken with the other Social Democrats. In the following year, Dzhugashvili published, at Lenin's behest, an important article on Marxism and the national question. By now he had adopted the name Stalin, deriving from Russian *stal* ("steel"); he also briefly edited the newly founded Bolshevik newspaper *Pravda* before undergoing his longest period of exile: in Siberia from July 1913 to March 1917.

In about 1904 Stalin had married a pious Georgian girl, Ekaterina Svanidze. She died some three years later and left a son, Jacob, whom his father treated with contempt, calling him a weakling after an unsuccessful suicide attempt in the late 1920s; when Jacob was taken prisoner by the Germans during World War II, Stalin refused a German offer to exchange his son.

Reaching Petrograd from Siberia on March 25 (March 12, O.S.), 1917, Stalin resumed editorship of *Pravda*. He briefly advocated Bolshevik cooperation with the provisional government of middle class liberals that had succeeded to uneasy power on the last tsar's abdication during the February Revolution. But under Lenin's influence, Stalin soon switched to the more militant policy of armed seizure of power by the Bolsheviks. When their coup d'etat occurred in November (October, old style) 1917, he played an important role, but one less prominent than that of his chief rival, Leon Trotsky.

Active as a politico-military leader on various fronts during the Civil War of 1918–20, Stalin also held two ministerial posts in the new Bolshevik government, being commissar for nationalities (1917–23) and for state control (or workers' and peasants' inspection; 1919–23). But it was his position as secretary general of the party's Central Committee, from 1922 until his death, that provided the power base for his dictatorship. Besides heading the secretariat, he was also member of the powerful Politburo and of many other interlocking and overlapping committees—an arch-bureaucrat engaged in quietly outmanoeuvring brilliant rivals, including Trotsky and Grigory Zinoviev, who despised such mundane organizational work. Because the pockmarked Georgian was so obviously unintellectual, they thought him unintelligent: a gross error, and one literally fatal in their case.

Ministerial posts

From 1921 onward Stalin flouted the ailing Lenin's wishes, until, a year before his death, Lenin wrote a political "testament," since widely publicized, calling for Stalin's removal from the secretary generalship; coming from Lenin, this document was potentially ruinous to Stalin's career, but his usual luck and skill enabled him to have it discounted during his lifetime.

Lenin's successor. After Lenin's death, in January 1924, Stalin promoted an extravagant, quasi-Byzantine cult of the deceased leader. Arch-priest of Leninism, Stalin also promoted his own cult in the following year by having the city of Tsaritsyn renamed Stalingrad (now Volgograd). His main rival, Trotsky (once Lenin's heir apparent), was now in eclipse, ousted by the ruling triumvirate of Zinoviev, Lev Kamenev, and Stalin. Soon afterward Stalin joined with the rightist leaders Nikolay Bukharin and Aleksey Rykov in an alliance directed against his former co-triumvirs. Pinning his faith in the ability of the Soviet Union to establish a viable political system without waiting for the support hitherto expected from worldwide revolution, the Secretary General advocated a policy of "Socialism in one country"; this was popular with the hardheaded party managers whom he was promoting to influential positions in the middle hierarchy. His most powerful rivals were all dismissed, Bukharin and Rykov soon following Zinoviev and Kamenev into disgrace and political limbo pending execution. Stalin expelled Trotsky from the Soviet Union in 1929 and had him assassinated in Mexico in 1940.

The policy of "Socialism in one country"

Collectivization of agriculture and industrialization

In 1928 Stalin abandoned Lenin's quasi-capitalist New Economic Policy in favour of headlong state-organized industrialization under a succession of five-year plans. This was, in effect, a new Russian revolution more devastating in its effects than those of 1917. The dictator's blows fell most heavily on the peasantry, some 25,000,000 rustic households being compelled to amalgamate in collective or state farms within a few years. Resisting desperately, the reluctant muzhiks were attacked by troops and OGPU (political police) units. Uncooperative peasants, termed kulaks, were arrested en masse, being shot, exiled, or absorbed into the rapidly expanding network of Stalinist concentration camps and worked to death under atrocious conditions. Collectivization also caused a great famine in the Ukraine. Yet Stalin continued to export the grain stocks that a less cruel leader would have rushed to the famine-stricken areas. Some 10,000,000 peasants may have perished through his policies during these years.

Crash industrialization was less disastrous in its effects, but it, too, numbered its grandiose failures, to which Stalin responded by arraigning industrial managers in a succession of show trials. Intimidated into confessing imaginary crimes, the accused served as self-denounced scapegoats for catastrophes arising from the Secretary General's policies. Yet Stalin was successful in rapidly industrializing a backward country—as was widely acknowledged by enthusiastic contemporary foreign witnesses, including Adolf Hitler and such well-known writers as H.G. Wells and George Bernard Shaw.

Among those who vainly sought to moderate Stalin's policies was his young second wife, Nadezhda Alliluyeva, whom he had married in 1919 and who committed suicide in 1932. They had two children. The son, Vasily, perished as an alcoholic after rising to unmerited high rank in the Soviet Air Force. The daughter, Svetlana, became the object for her father's alternating affection and bad temper; emigrating after his death, she wrote memoirs that illuminate Stalin's well-camouflaged private life.

The great purges. In late 1934—just when the worst excesses of Stalinism seemed to have spent themselves—the Secretary General launched a new campaign of political terror against the very Communist Party members who had brought him to power; his pretext was the assassination, in Leningrad on December 1, of his leading colleague and potential rival, Sergey Kirov. That Stalin himself had arranged Kirov's murder—as an excuse for the promotion of mass bloodshed—was strongly hinted by Nikita Khrushchev, first secretary of the party, in a speech denouncing Stalin at the 20th Party Congress in 1956.

Stalin used the show trial of leading Communists as a means for expanding the new terror. In August 1936, Zinoviev and Kamenev were paraded in court to repeat fabricated confessions, sentenced to death, and shot; two more major trials followed, in January 1937 and March 1938. In June 1937, Marshal Mikhail Tukhachevsky, at the time the most influential military personality, and other leading generals were reported as court-martialled on charges of treason and executed.

Such were the main publicly acknowledged persecutions that empowered Stalin to tame the Soviet Communist Party and the Soviet elite as a whole. He not only "liquidated" veteran semi-independent Bolsheviks but also many party bosses, military leaders, industrial managers, and high government officials totally subservient to himself. Other victims included foreign Communists on Soviet territory and members of the very political police organization, now called the NKVD. All other sections of the Soviet elite—the arts, the academic world, the legal and diplomatic professions—also lost a high proportion of victims, as did the population at large, to a semi-haphazard, galloping persecution that fed on extorted denunciations and confessions. These implicated even more victims until Stalin himself reduced the terror, though he never abandoned it. Stalin's political victims were numbered in tens of millions. His main motive was, presumably, to maximize his personal power.

Role in World War II. During World War II Stalin emerged, after an unpromising start, as the most successful of the supreme leaders thrown up by the belligerent nations. In August 1939, after first attempting to form an anti-Hitler alliance with the Western powers, he concluded a pact with Hitler, which encouraged the German dictator to attack Poland and begin World War II. Anxious to strengthen his western frontiers while his new but palpably treacherous German ally was still engaged in the West, Stalin annexed eastern Poland, Estonia, Latvia, Lithuania, and parts of Romania; he also attacked Finland and extorted territorial concessions. In May 1941 Stalin recognized the growing danger of German attack on the Soviet Union by appointing himself chairman of the Council of People's Commissars (head of the government): his first governmental office since 1923.

Stalin's prewar defensive measures were exposed as incompetent by the German blitzkrieg that surged deep into Soviet territory after Hitler's unprovoked attack on the Soviet Union of June 22, 1941. Khrushchev claimed that Stalin was shocked into temporary inactivity by the onslaught, but, if so, he soon rallied and appointed himself supreme commander in chief. When the Germans menaced Moscow in the winter of 1941, he remained in the threatened capital, helping to organize a great counter-offensive. The battle of Stalingrad (in the following winter) and the Battle of Kursk (in the summer of 1943) were also won by the Soviet Army under Stalin's supreme direction, turning the tide of invasion against the retreating Germans, who capitulated in May 1945. As war leader, Stalin maintained close personal control over the Soviet battlefronts, military reserves, and war economy. At first overinclined to intervene with inept telephoned instructions a la Hitler, the Soviet generalissimo gradually learned to delegate military decisions.

Role as supreme commander in chief

Stalin participated in high-level allied meetings, including those of the "Big Three" with Churchill and Roosevelt at Teheran (1943) and Yalta (1945). A formidable negotiator, he outwitted these foreign statesmen; his superior skill has been acclaimed by Anthony Eden, then British foreign secretary.

Last years. After the war Stalin imposed on eastern Europe a new kind of colonial control based on native Communist regimes nominally independent but in fact subservient to himself. He thus increased the number of his subjects by about a hundred million. But in 1948 the defection of Titoist Yugoslavia from the Soviet camp struck a severe blow to world Communism as a Stalin-dominated monolith. To prevent other client states from following Tito's example, Stalin instigated local show trials, manipulated like those of the Great Purge of the 1930s in Russia, in which satellite Communist leaders confessed to Titoism, many being executed.

Postwar foreign policy

Far from continuing his wartime alliance with the United States and Britain, Stalin now regarded these countries —and especially the United States—as the archenemies that he needed after Hitler's death. At home, the primacy of Marxist ideology was harshly reasserted, Stalin's chief ideological hatchet man, Andrey Zhdanov, the secretary of the Central Committee, mounting a reign of terror in the Soviet artistic and intellectual world; foreign achievements were derided, and the primacy of Russians as inventors and pioneers in practically every field was asserted. Hopes for domestic relaxation, widely aroused in the Soviet Union during the war, were thus sadly disappointed.

Increasingly suspicious and paranoid in his later years, Stalin ordered the arrest, announced in January 1953, of certain—mostly Jewish—Kremlin doctors on charges of medically murdering various Soviet leaders, including Zhdanov. The dictator was evidently preparing to make this "Doctors' Plot" the pretext for yet another great terror menacing all his senior associates, but he died suddenly on March 5, 1953, according to the official report; so convenient was this death to his entourage that suspicions of foul play were voiced.

Assessment. A politician to the marrow of his bones, Stalin had little private or family life, finding his main relaxation in impromptu buffet suppers, to which he

Personal life

would invite high party officials, generals, visiting foreign potentates, and the like. Drinking little himself on these occasions, the dictator would encourage excessive indulgence in others, thus revealing weak points that he could exploit. He would also tease his guests, jocularity and malice being nicely balanced in his manner; for such bluff banter Stalin's main henchman, Vyacheslav Molotov, the stuttering foreign minister, was often a target. Stalin had a keen, ironical sense of humour, usually devoted to deflating his guests rather than to amusing them.

Achievements

Foremost among Stalin's accomplishments was the industrialization of a country which, when he assumed complete control in 1928, was still notably backward by comparison with the leading industrial nations of the world. By 1937, after less than a decade's rule as totalitarian dictator, he had increased the Soviet Union's total industrial output to the point where it was surpassed only by that of the United States. The extent of this achievement may best be appreciated if one remembers that Russia had held only fifth place for overall industrial output in 1913, and that it thereafter suffered many years of even greater devastation—through world war, civil war, famine, and pestilence—than afflicted any of the world's other chief industrial countries during the same period. Yet more appallingly ravaged during World War II, the Soviet Union was nevertheless able, under Stalin's leadership, to play a major part in defeating Hitler while maintaining its position as the world's second most powerful industrial—and now military—complex after the United States. In 1949 Stalinist Russia signalized its arrival as the world's second nuclear power by exploding an atomic bomb.

Against these formidable achievements must be set one major disadvantage. Though a high industrial output was indeed achieved under Stalin, very little of it ever became available to the ordinary Soviet citizen in the form of consumer goods or amenities of life. A considerable proportion of the national wealth—a proportion wholly unparalleled in the history of any peacetime capitalist country—was appropriated by the state to cover military expenditure, the police apparatus, and further industrialization. It is also arguable that a comparable degree of industrialization would have come about in any case—and surely by means less savage—under almost any conceivable regime that might have evolved as an alternative to Stalinism.

Stalin's collectivization of agriculture did not produce—either during his lifetime or during the decades immediately following his death—positive economic results remotely comparable to those attained by Soviet industry. Considered as a means of asserting control over the politically recalcitrant peasantry, however, collectivization triumphantly justified itself and continues to do so, remaining one of the dictator's most durable achievements. Moreover, the process of intensive urbanization, as instituted by Stalin, continued after his death in what still remained a population more predominantly rural than that of any other major industrial country. In 1937, 56 percent of the population was recorded as engaged in agriculture or forestry; by 1958 that proportion had dropped to 42 percent, very largely as a result of Stalin's policies.

The most lasting of all the dictator's achievements may prove to be his elaborately bureaucratized administrative machinery based on the interlinking of the Communist Party, ministries, legislative bodies, trade unions, political police, and armed forces, as also on a host of other meshing control devices. During the two decades following the dictator's death, these continued to supply the essential management levers of Soviet society, often remaining under the control of individuals who had risen to prominence during the years of the Stalinist terror. But the element of total personal dictatorship did not survive Stalin in its most extreme form. One result of his death was the resurgence of the Communist Party as the primary centre of power, after years during which that organization, along with all other Soviet institutions, had been subordinated to a single man's whim. Yet, despite the great power wielded by Stalin's successors as party leaders, they became no more than dominant figures within the framework of a ruling oligarchy. They did not develop into potentates responsible to themselves alone, such as Stalin was during his quarter of a century's virtually unchallenged rule.

That Stalin's system persists, in all its major essentials, after the death of its creator is partly due to the very excess of severity practiced by the great tyrant. Not only did his methods crush initiative among Soviet administrators, physically destroying many, but they also left a legacy of remembered fear so extreme as to render continuing post-Stalin restrictions tolerable to the population; the people would have more bitterly resented—might even, perhaps, have rejected—such rigours, had it not been for their vivid recollection of repressions immeasurably harsher. Just as Hitler's wartime cruelty toward the Soviet population turned Stalin into a genuine national hero—making him the Soviet Union's champion against an alien terror even worse than his own—so too Stalin's successors owed the stability of their system in part to the comparison, still fresh in many minds, with the far worse conditions that obtained during the despot's sway.

Stalin has arguably made a greater impact on the lives of more individuals than any other figure in history. But the evaluation of his overall achievement still remains, nearly two decades after his death, a highly controversial matter. Historians have not yet reached any definitive consensus of opinion on the worth of his accomplishments, and it is unlikely that they ever will. To the United States scholar George F. Kennan, Stalin is a great man, but one great in his "incredible criminality . . . a criminality effectively without limits," while Robert C. Tucker, a United States specialist on Soviet affairs, has described Stalin as a 20th-century Ivan the Terrible. To the British historian E.H. Carr, the Georgian dictator appears as a ruthless, vigorous figure, but one lacking in originality—a comparative nonentity thrust into greatness by the inexorable march of the great revolution that he found himself leading. To the late Isaac Deutscher, the author of biographies of Trotsky and Stalin—who, like Carr, broadly accepts Trotsky's version of Stalin as a somewhat mediocre personage—Stalin represents a lamentably deviant element in the evolution of Marxism. Neither Deutscher nor Carr has found Stalin's truly appalling record sufficiently impressive to raise doubts about the ultimate value of the Russian October Revolution's historic achievements.

Reputation

To such views, each illuminating a facet of Stalin's personality, may be added the suggestion that Stalin was anything but a plodding mediocrity, being rather a man of superlative, all-transcending talent. His special brilliance was, however, narrowly specialized and confined within the single crucial area of creative political manipulation. Outside the realm of pioneering political manoeuvre, in which Stalin's flair remains unsurpassed, his qualities appear in no way superhuman, apart perhaps from the dogged persistence with which he was accustomed to pursue to the death both his manifold goals and his innumerable enemies.

MAJOR WORKS

BOOKS: *Anarkhizm ili sotsializm?* (1949, Russian version of articles published serially in Georgian periodicals, 1906–07; *Anarchism or Socialism?*, 1950); *Natsionalny vopros i marksizm* (1914, based on article "Natsionalny vopros i sotsial-demokratiya" in periodical *Prosveshchenie*, 1913, and reissued as *Marksizm i natsionalnokolonialny vopros*, 1934; *Marxism and the National and Colonial Question*, 1935); *Na putyakh k Oktyabryu* (1925; selections translated as *The Road to Power*, 1937); *Voprosy leninizma* (1926, 11th ed., much enlarged, 1939; *Leninism*, 1940, or *Problems of Leninism*, 1940 and 1953), a monumental collection of treatises, articles, reports, and speeches including "Ob osnovakh leninizma" (1924; *Theory and Practice of Leninism*, 1925, and *Foundations of Leninism*, 1932); "K voprosam agrarnoy politiki v SSSR" (1929); "O proyekte konstitutsii Soyuza SSR" (1936; *On the Draft Constitution of the U.S.S.R.*, 1936, and *The New Democracy*, 1937); and "O dialekticheskom i istoricheskom materializme" (1938; *Dialectical and Historical Materialism*, 1941); *Ob oppozitsii* (1928), a massive collection of writings of the period 1921–27; *O pravom*

uklone v VKP(b) (speech of 1929 published integrally as a separate book in 1954); *O nedostatkakh partiynoy raboty i merakh likvidatsy Trotskistskikh i inykh dvurushnikov* (1937); *Istoriya Vsesoyuznoy Kommunisticheskoy Partii (Bolshevikov): Kratky kurs* (1938, ostensibly the work of a party committee, but wholly Stalinian in inspiration; *Short History of the Communist Party of the Soviet Union*, 1939); *Iosif Vissarionovich Stalin: Kratkaya biografiya* (1940, 2nd ed. 1947, likewise ostensibly a collective production; *Stalin*, 1943); *O Velikoy Otechestvennoy Voyne Sovetskogo Soyuza* (1942, 5th ed., much enlarged, 1946–52; *The Great Patriotic War of the Soviet Union*, 1945, or *War Speeches, Orders of the Day, and Answers to Foreign Press Correspondents during the Great Patriotic War, July 3, 1941–June 22nd, 1945*, 1946); *Marksizm i voprosy yazykoznaniya* (1950; *Marxism and Linguistics*, 1951); *Ekonomicheskie problemy sotsializma v SSSR* (1952; *Economic Problems of Socialism in the U.S.S.R.*, 1952); *Perepiska Predsedatelya Sovieta Ministrov SSSR s prezidentami SShA i premier-ministrami Velikobritanii vo vremya Velikoy Otechestvennoy Voyny* (1957; *Correspondence with Churchill, Attlee, Roosevelt and Truman, 1941–45*, 2 pt., 1958).

BIBLIOGRAPHY. The standard Soviet edition of Stalin's works in Russian is his *Сочинения*, 13 vol. (1946–51), covering publications up to January 1934; an American-published continuation of this edition and containing later works (vol. 14–16) in Russian has been issued in similar format by the Hoover Institution (1967), ed. by ROBERT H. MCNEAL.

An annotated bibliography by ROBERT H. MCNEAL, *Stalin's Works* (1967), considers the authenticity of material in Russian attributed to Stalin. Bibliographies more useful to the general reader will be found in the works of Deutscher, Payne, Souvarine, and Smith cited below.

There is no definitive biography of Stalin. The most useful among scores of published studies include: ROBERT PAYNE, *Rise and Fall of Stalin* (1965), readable, but verbose; ISAAC DEUTSCHER, *Stalin: A Political Biography*, 2nd ed. (1966), one-sidedly Marxist in its approach; LEON TROTSKY, *Stalin: An Appraisal of the Man and His Influence*, 2nd ed. (1946), an eloquent and witty but biassed and uncompleted study; also denunciatory, BORIS SOUVARINE, *Stalin: A Critical Survey of Bolshevism* (1939), still commands some respect, as does BERTRAM D. WOLFE, *Three Who Made a Revolution* (1956). *Stalin*, ed. by T.H. RIGBY (1966), is an excellent short anthology of biographical and critical material.

Reliable, detailed firsthand accounts of Stalin's domestic background are few, the only family memoirs not subject to Stalinist censorship being those published by the dictator's daughter, SVETLANA ALLILUYEVA, in emigration: *Twenty Letters to a Friend* (Eng. trans. 1967) and *Only One Year* (Eng. trans. 1969); for Soviet-censored memoirs by Svetlana Alliluyeva's grandfather and aunt, see *The Alliluyev Memoirs*, trans. and ed. by DAVID TUTAEV (1968); the most valuable full account of Stalin in his role as dinner-party host is that of MILOVAN DJILAS, *Conversations with Stalin* (Eng. trans. 1962); on this subject the memoirs of Winston Churchill, Anthony Eden, Charles de Gaulle, and many others may be consulted (see also the bibliographies in works by Payne and Deutscher cited above).

On Stalin's pre-revolutionary career, EDWARD ELLIS SMITH, *The Young Stalin* (1967), unsuccessfully but informatively attempts to prove that the subject was an agent of the Tsarist political police; erring far more in the opposite direction, LAVRENTY BERIA, *On the History of the Bolshevik Organizations in Transcaucasia* (1949), mendaciously glorifies Stalin's role as an anti-Tsarist political conspirator and is the chief classic of Stalinist legend-building.

For the political background of Stalin's mature career, LEONARD SCHAPIRO, *The Communist Party of the Soviet Union*, 2nd ed. (1970), is especially valuable, as is JOHN A. ARMSTRONG, *The Politics of Totalitarianism* (1961); a more subjective version of the same events by a former participant is ABDURAKHMAN AVTORKHANOV, *Stalin and the Soviet Communist Party* (1959). BORIS I. NICOLAYEVSKY, *Power and the Soviet Elite* (1965), is an illuminating collection of essays bearing on Stalin's activities from 1934 onward by the leading Russian émigré authority on Soviet conspiratorial politics. ROBERT CONQUEST, *The Great Terror* (1968), is the fullest and liveliest account of the persecutions of 1937–38, on which ZBIGNIEW K. BRZEZINSKI, *Permanent Purge* (1956), is also helpful; see further *The Great Purge Trial*, ed. by ROBERT C. TUCKER and STEPHEN F. COHEN (1965). F. BECK and W. GODIN, *Russian Purge and the Extraction of Confession* (1951), remains a classic account of Stalin's terror; on this see also W.G. KRIVITSKY, *I Was Stalin's Agent* (1939); and ALEXANDER ORLOV, *Secret History of Stalin's Crimes* (1953), informative primary sources, but they cannot always be checked against other material.

On Stalin's role as wartime leader, SEWERYN BIALER'S anthology of Soviet memoir material, *Stalin and His Generals* (1969), is outstandingly useful; so too is HERBERT FEIS, *Churchill, Roosevelt, Stalin* (1957); see further, ALEXANDER WERTH, *Russia at War, 1941–1945* (1964), and *The Memoirs of Marshal Zhukov* (1971).

On post-war and posthumous Stalin, ROBERT CONQUEST, *Power and Policy in the U.S.S.R.* (1961), plausibly unravels many mysteries. On the controversy surrounding Stalin's image after his death, see *The Anti-Stalin Campaign and International Communism: A Selection of Documents*, ed. by the RUSSIAN INSTITUTE, COLUMBIA UNIVERSITY (1956); BERTRAM D. WOLFE, *Khrushchev and Stalin's Ghost* (1957); and *The Stalin Dictatorship: Khrushchev's "Secret Speech" and Other Documents*, ed. by T.H. RIGBY (1968).

(R.F.Hi.)

Stanislavsky, Konstantin

With legitimate claims to historic significance as an actor, as a director, and as a founder of the Moscow Art Theatre, Konstantin Stanislavsky (also spelled Stanislavski) is best known as the innovator of what may have been the most influential theory of acting of the 20th century—his celebrated system, the basis of what became widely known as the "method." Beginning as a Realist opposed to the highly stylized theatrical conventions of the late 19th century, he remained preoccupied throughout his career with the complex relationships between inner truths felt by the performer and the outer truths his performance conveys to the audience, that is, between the actor's subjective psychological states and the external physical expression of those states. Seeking a means of commanding the resources that are evident in a performer's moments of "inspiration," Stanislavsky came close to creating a science of theatre.

By courtesy of the Moscow Art Theatre Museum; photograph, Sovfoto



Stanislavsky as Vershinin in Chekhov's *Three Sisters*, 1901.

Early influences. Stanislavsky was born Konstantin Sergeyevich Alekseyev in Moscow on January 17 (new style; 5, old style), 1863. His father, a manufacturer, and mother, the daughter of a French actress, were intimate with the outstanding art and literary personalities of Moscow. As a child Stanislavsky attended the circus, opera, and ballet, and his early years were influenced by an atmosphere of family affection and a fascination with the theatre. When he first appeared on his parents' amateur stage, at the age of 14, he believed his performance was inspired and was distressed that it was not praised. In the dramatic group that subsequently was organized by his family and called the Alekseyev Circle, he was teased for his clumsiness while performing comedy and vaudeville. At first hiding his desire to become an actor with characteristic self-doubts and self-criticism, Stanislavsky obsessively worked on his shortcomings of voice and diction

The Alekseyev Circle

and on the awkwardness of his body. His thoroughness and his preoccupation with all aspects of a production came to distinguish him from other members of the Alexeyev Circle, and he gradually became its central figure. Stanislavsky also performed in other groups as theatre came to absorb his life. He adopted the pseudonym Stanislavsky in 1885, and in 1888 he married Maria Perevoshchikova, a schoolteacher, who became his devoted disciple and lifelong companion, as well as an outstanding actress under the name Lilina.

Stanislavsky's love for theatre developed into an allegiance to what he saw as an art of social significance. Theatre, a powerful influence on people, he believed, must be on the highest level, and the actor must be the people's educator. Evaluating his own and his colleagues' acting from this viewpoint, he found it unsatisfactory. Moreover, he found that it could not be improved through sporadic productions but required instead a permanent company.

Building a people's theatre. In 1888 Stanislavsky and others established the Society of Art and Literature with a permanent amateur company. Endowed with great talent, musicality, a striking appearance, a vivid imagination, and a subtle intuition, Stanislavsky began to develop the plasticity of his body and a greater range of voice. Praise came from famous foreign actors, and great Russian actresses invited him to perform with them. Thus encouraged, Stanislavsky staged his first independent production, Tolstoy's *Fruits of Enlightenment*, in 1891, a major Moscow theatrical event. Most significantly, it impressed a promising writer and director, Vladimir Nemirovich-Danchenko (1858–1943), whose later association with Stanislavsky was to have a paramount influence on the theatre.

Collaboration with Nemirovich-Danchenko

Nemirovich-Danchenko followed Stanislavsky's activities until their historic meeting in 1897, when they outlined a plan for a people's theatre. It was to consist of the most talented amateurs of Stanislavsky's society and of the students of the Philharmonic Music and Drama School, which Nemirovich-Danchenko directed. As the Moscow Art Theatre, it became the arena for Stanislavsky's reforms. Despite differences in personality, the two men, both of whom were vital to the artistic development of the theatre, worked closely until Stanislavsky's death. Nemirovich-Danchenko was a man of the world, who enjoyed gaiety: Stanislavsky's tastes were puritanical. When Nemirovich-Danchenko reprimanded an actor, he did so privately in his office; Stanislavsky, demanding respect toward theatre, admonished actors in a voice that echoed through the building. A major artistic disagreement occurred in 1905, when Stanislavsky suggested rehearsing a play by having the actors first improvise their parts. Nemirovich-Danchenko and the director Vsevolod Meyerhold (1874–1942) vehemently objected to the proposal. Stanislavsky thus abandoned the idea, but 30 years later he formulated it as a law for the analysis of a play and a role.

Although Stanislavsky was a compassionate man, he was uncompromising in educating actors. He might occasionally ban actors from rehearsal for days as punishment for a few minutes' tardiness. He insisted on integrity of performance on stage, repeating for hours his dreaded "I do not believe you." His dedication was total: when an actor asked Stanislavsky to demonstrate and then interrupted his demonstration with "I do not believe you," the concentrating master replied "Really? Then I shall have to work on this at home." Throughout his career, Stanislavsky went to the theatre hours before curtain time to put on what he called his physical and spiritual makeup. Addressing the cast for Aleksey Tolstoy's *Tsar Fyodor Ioannovich* before the opening of the Moscow Art Theatre on October 14 (26, O.S.), 1898, Stanislavsky said: "We strive to brighten the dark life of poor people, to give them happy aesthetic moments. We are striving for the first sensible, moral, popular theatre, and to this high goal we dedicate our lives."

Developing the system. Stanislavsky was disappointed in the acting that night. He found it to be merely imitative of the gestures, intonations, and conceptions of the director. To affect the spectators, he reflected, there must be alive people on stage, and the mere external behaviour of the actors is insufficient to create a man's unique inner world. To learn about inner behaviour, Stanislavsky turned to psychologists and at the same time sharply changed his approach—from the purely external to the purely psychological. He began developing the first elements of his "grammar of dramatic art." A play was studied around a table for months; the actors were required to achieve the experience they were to portray before entering on stage.

Relations with Chekhov

Stanislavsky's experience with Anton Chekhov's *Sea Gull* confirmed his convictions about the need for what he called "the life of the human spirit" on stage. With difficulty, Stanislavsky obtained Chekhov's permission to restage *The Sea Gull* after its original production in St. Petersburg in 1896 had been a failure. Directed by Stanislavsky and Nemirovich-Danchenko in 1898, *The Sea Gull* became a triumph, heralding the birth of the Moscow Art Theatre as a new force in world theatre. Chekhov, who had resolved never to write another play after his initial failure, was acclaimed a great playwright. His writings significantly influenced the development of the system that was to evolve over decades in Stanislavsky's search for the best means to be used by actors and directors.

Staging Chekhov's play, Stanislavsky and Nemirovich-Danchenko discovered a new manner of performing, of revealing a man's inner world: they emphasized ensemble and the subordination of each individual actor to the whole. All technical means were directed toward creating one image. Seeking a theatrical language through which the play's life would be relived by the audience, Stanislavsky subordinated the director's and actors' interpretation to the dramatist's "superobjective," that is, to his intent. Actors, he felt, had to have a common training and be capable of "reincarnation." That reincarnation, however, did not imply the complete identification of an actor with the character; rather, some part of the actor's personality must remain independent of the role in order to subordinate it to the needs of the play as a whole. This point was clarified by Stanislavsky's teachings of conscious control and of the actor's "I" and "not I." Total "melting"—that is, the total identification of an actor with a particular role—he said, occurred only in pathological cases.

Stanislavsky was haunted by the actor's need for a means to stir his inner life, to awaken his own initiative. Endowed with an analytical mind, Stanislavsky generalized and systematized his own experience and that of his colleagues and students. He combined all this with his knowledge of the performances and statements of major artists. Fighting against the artificiality of stage, dilettantism, and acting "by instinct," he sought a natural state based on the laws of man's organic nature. Stanislavsky was building a theatre of "living experience" that demanded not an external demonstration of emotions, as was found in what he called the "theatre of representation," but the reproduction of authentic emotions at every performance: "Here, today, now." Actors, he said, must learn to use the natural laws through which geniuses sometimes spontaneously channel their creativity. This is the essence of the system: to enable the capable actor to achieve the creative state that was hitherto the privilege of genius alone.

The "super-super-objective"

The basis of Stanislavsky's educational program is expressed in his "Love art in you, not you in art." He stressed friendship in the group and ensemble on stage, and he cited as the actors' "super-superobjective" the striving to contribute to society. His dedication and powerful personality had great influence. Emphasizing dignity in the theatre, Stanislavsky demanded harmony between the actor's life and professional behaviour. He warned that the artist's spiritual formation would be sensed in his creations. Stanislavsky's ethical principles, which were the foundation of his aesthetics, were always indivisible from his methodology; he believed them to be essential to an atmosphere in which collective creativity was possible.

In 1902 Stanislavsky successfully staged both Gorky's *Petty Bourgeois* and *The Lower Depths*, codirecting the latter with Nemirovich-Danchenko. Among the numerous powerful roles performed by Stanislavsky were Astrov in *Uncle Vanya* in 1899 and Gayev in *The Cherry Orchard* in 1904, by Chekhov; Doctor Stockman in Ibsen's *Enemy of the People* in 1900; and Satin in *The Lower Depths*. Both as an actor and as a director, Stanislavsky demonstrated a remarkable subtlety in rendering psychological patterns and an exceptional talent for satirical characterization.

Commanding respect from followers and adversaries alike, Stanislavsky became a dominant influence on the Russian intellectuals of the time. He formed the First Studio in 1912, where his innovations were verified through the work of young actors, avoiding the older, more conservative actors' resistance. In 1918 he undertook the guidance of the Bolshoi Opera Studio, which was later named for him. There he staged Tchaikovsky's *Eugene Onegin* in 1922, which was acclaimed as a major reform in opera.

Revising the system. In 1922-24 the Moscow Art Theatre toured Europe and the United States with Stanislavsky as its administrator, director, and leading actor. A great interest was stirred in his system. During this period he wrote his famous autobiography, *My Life in Art*. Ever preoccupied in it with content and form, Stanislavsky acknowledged that the "theatre of representation," which he had disparaged, nonetheless produced brilliant actors. In the best of them he perceived true emotions; the great opera singer Fyodor Chaliapin combined depth and grand theatrical form, and other celebrated Realist actors would not have been so impressive, Stanislavsky recognized, if they had not attached importance to external expression.

Return to theatricality

Recognizing that theatre was at its best when deep content harmonized with vivid theatrical form, Stanislavsky supervised the First Studio's production of Shakespeare's *Twelfth Night* in 1917 and Gogol's *Government Inspector* in 1921, encouraging the actor Michael Chekhov in a brilliantly grotesque characterization. His staging of Aleksandr Ostrovsky's *Ardent Heart* (1926) and of Beaumarchais's *Marriage of Figaro* (1927) demonstrated increasingly bold attempts at theatricality. His monumental *Armored Train 14–69*, a play about the Russian Revolution, was a milestone in Soviet theatre in 1927, and his *Dead Souls* was a brilliant incarnation of Gogol's masterpiece.

Stanislavsky continued to establish, to destroy, and to establish again. Although his techniques changed, his method for many years remained purely psychological. Always critical of his own means, however, Stanislavsky noted that after prolonged mental work, actors went on stage with "swollen heads" and that they could not act as characters but were simply dwelling in their own experience. He witnessed the transformation of his own precepts into a new theatrical cliché: actors presented a demonstration of an experience that only had the look of truth. He concluded that mental analysis alone harboured the danger of a calculated rather than a creative approach, paralyzing the actor's subconscious. Stanislavsky's efforts to create an alive man on stage seemed to be threatened. With increased zeal, he pursued the means to "organized" emotions; that is, a means by which the emotions implicit in a role could be aroused at will.

Last years. While acting in *The Three Sisters* during the Moscow Art Theatre's 30th anniversary presentation on October 29, 1928, Stanislavsky suffered a heart attack. Abandoning acting, he concentrated on directing and educating actors and directors. In 1935 he founded the Opera Drama Studio.

Pavlov's influence

In the early 1930s Stanislavsky studied the writings of physiologist Ivan Pavlov and attempted to analyze the actor's creativity on the basis of the brain's higher nervous activity. Pavlov, too, was interested in Stanislavsky's system, an association interrupted by Pavlov's death in 1936.

Stanislavsky now followed Pavlov's road in the science of the brain. Seeking a means of synthesizing content and form, Stanislavsky in 1935 fulfilled a 40-year quest with the discovery of what he regarded as the unbreakable psychophysical unity of the human experience and its external expression. Using this discovery, Stanislavsky developed his "method of physical actions," an analysis leading to the synthesis of content (or experience) and form (or expression) and the means to "organized" emotions. He immediately began an experimental production of Molière's *Tartuffe*.

Although his teachings influenced theatre throughout the world, Stanislavsky was reluctant to publish his discoveries until his theses had been verified in practice. Thus, many of his writings were prepared for publication only after his death in Moscow on August 7, 1938, by the scholars who compiled and edited the eight volumes of his *Complete Works*.

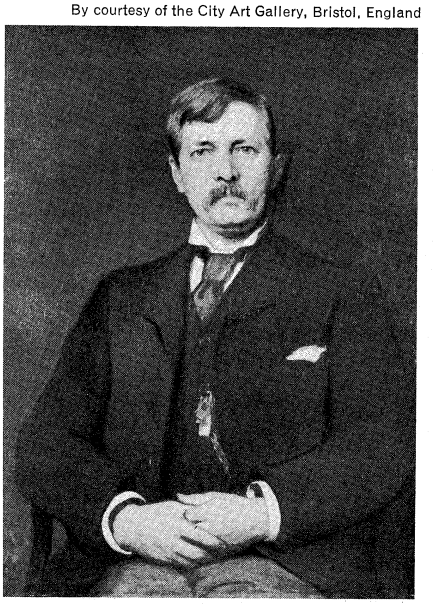
BIBLIOGRAPHY. DAVID MAGARSHACK, *Stanislavsky: A Life* (1950); K.S. STANISLAVSKY, *My Life in Art*, trans. from the Russian by J.J. ROBBINS (1924, reprinted 1956), an outstanding autobiography; *An Actor Prepares*, trans. from the Russian by E.R. HAPGOOD (1936, reprinted 1956), Stanislavsky's first book on his technique and the only one published during his lifetime; and *Building a Character*, trans. from the Russian by E.R. HAPGOOD (1949), a text prepared for publication by Stanislavsky's disciples; SONIA MOORE, *The Stanislavski System*, rev. ed., preface by SIR JOHN GIELGUD, foreword by JOSHUA LOGAN (1965), an updated discussion of the Stanislavsky acting technique, *Training an Actor: The Stanislavski System in Class* (1968), a work based on class tape recordings demonstrating the Stanislavsky technique in practice.

(S.M.)

Stanley, Sir Henry Morton

Though Sir Henry Morton Stanley first achieved fame for his rescue of David Livingstone, in 1871, his true importance lies in his discovery and development of the Congo. To him belongs the credit of having assembled the geographical discoveries and guesses of an earlier generation into a coherent whole which approximates the present map of central Africa.

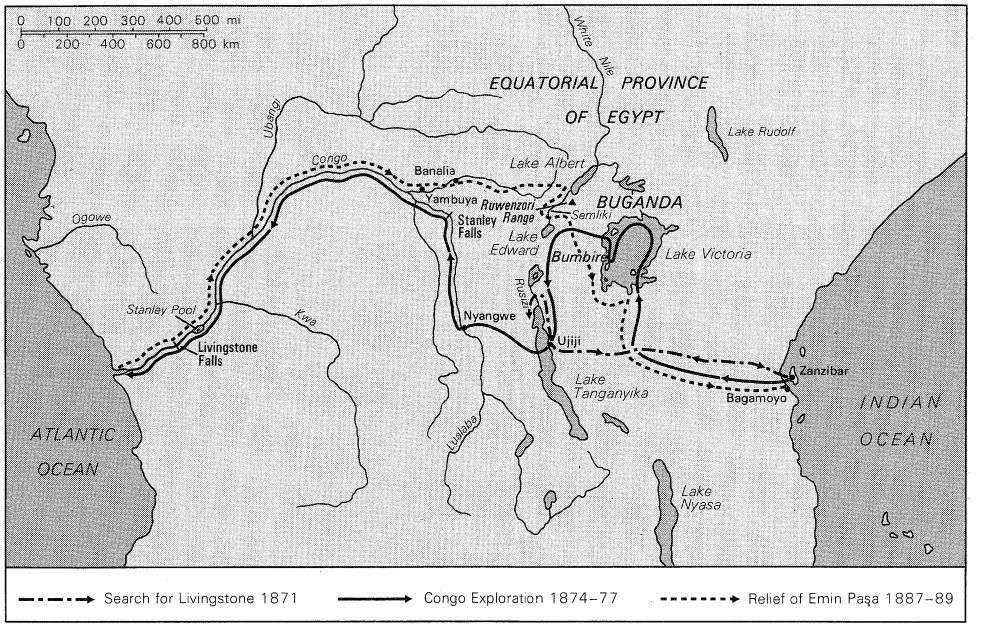
By courtesy of the City Art Gallery, Bristol, England



Stanley, portrait by Sir Hubert von Herkomer (1849–1914). In the City Art Gallery, Bristol, England.

Early life in Wales

He was born on January 28, 1841, at Denbigh, North Wales, the illegitimate son of John Rowlands and Elizabeth Parry, and was registered as John Rowlands. His mother left him to the care of her father, Moses Parry, and he grew up partly in the charge of reluctant relatives and partly in the St. Asaph workhouse, where he suffered from the severity of the workhouse master. He acquired a reasonable education at St. Asaph, developed a strong religious faith, and learned to value order and cleanliness; but the humiliations of institutional life left deep marks



The expeditions of Sir Henry Morton Stanley.

on his character, as did his mother's continued neglect. At the age of 15 he ran away from the workhouse, and his escape was followed by an unhappy interlude of dependence on his mother's relatives which ended only when he sailed from Liverpool as a cabin boy to New Orleans, Louisiana, where he landed in 1859.

There Rowlands met Henry Morton Stanley, a kind-hearted merchant who took charge of him, gave him his own name, and declared his intention of providing for him. Although the merchant died abroad soon afterward, leaving his protégé once more alone in the world, his kindness gave a new direction to the boy's life, and henceforth young Stanley felt that at least one person had loved him and believed in him. In his own prosperous middle age he, too, adopted a son.

For some years Stanley led a roving life: a soldier in the American Civil War, a seaman on merchant ships and in the Federal navy, a journalist in the early days of frontier expansion—he even managed a trip to Turkey, recorded in *My Early Travels and Adventures in America and Asia* (1895). Twice he visited his mother in Wales, only to encounter the same indifference she had shown since his birth.

In 1867 Stanley offered his services to James Gordon Bennett of the *New York Herald* as a special correspondent with the British expeditionary force sent against Theodore II of Ethiopia, and he was first to report the news of the fall of Magdala in 1868. He was in Madrid covering the Spanish Civil War for the *New York Herald* in 1869, when James Gordon Bennett, Jr., summoned him to Paris and commissioned him to "find Livingstone."

Relief of Livingstone. Little had been heard of Livingstone after he left for the interior of Africa in 1866 to explore the central African lakes and to ascertain the source of the Nile, but he was thought to be in the neighbourhood of Ujiji on Lake Tanganyika. Bennett instructed Stanley to travel to Africa by way of Egypt (for the opening of the Suez Canal), Syria, and Palestine, and it was not until January 6, 1871, that he reached Zanzibar.

Intent on a "scoop," Stanley was secretive with the Zanzibar authorities, including Sir John Kirk the British consul, who was responsible for Livingstone's supplies. Stanley set out for the interior on March 21, and with a well-equipped caravan forced his way through a countryside disturbed by fighting and stricken with sickness. He reached Ujiji in November, where he found Livingstone ailing and short of food and medicine, greeting him with the famous words "Dr. Livingstone, I presume?" Together they explored Lake Tanganyika and proved that since the Rusizi River flowed into and not out of the northern end of the lake, it could not be the source of the Nile as had been suggested by some geographers who doubted that Lake Victoria was the main source. Stanley arrived back in Zanzibar in May 1872, leaving Livingstone to pursue his search for the Nile further west in the region of what, as Stanley himself was later to prove, was the Congo. Stanley was not well received in England. In the same year (1872) he published his story in *How I Found Livingstone*.

Expedition with Livingstone

The Royal Geographical Society (Livingstone's chief sponsor) awarded him its Patron's Medal for 1873. He addressed the British association at Brighton in October 1872 and was received by Queen Victoria who gave him a gold snuffbox and her thanks for his services. In 1873 Stanley went to Ashanti (in modern Ghana) as war correspondent for the *New York Herald*, and in 1874 he published *Coomassie and Magdala: the Story of Two British Campaigns in Africa*, describing his experiences there and in Ethiopia.

Discovery and development of the Congo. When Livingstone died in 1873, Stanley resolved to take up the exploration of Africa where he had left off. The problem of the Nile sources and of the position and size of the central African lakes had been only partly solved by Livingstone, Sir Richard Burton, John H. Speke, and Sir Samuel Baker. Stanley secured financial backing from the *Daily Telegraph* and the *New York Herald* for an expedition to Africa. Well manned and provisioned and with three companions, Stanley left Zanzibar on November 12, 1874, and headed for Lake Victoria. His visit to King Mtesa of Buganda during 1875 led to the admission of Christian missionaries in 1876 and for the eventual establishment of a British protectorate. Circumnavigating Lake Victoria in his boat, the "Lady Alice," Stanley confirmed Speke's estimate of its size and importance. Skirmishes with suspicious natives on the lake shore and, especially, a fight at Bumbireh that resulted in a number of casualties gave rise in England to much criticism of this new kind of traveller with his journalistic outlook and forceful methods. On this voyage were sown many of the seeds of Stanley's later unpopularity among the English and of the rough parliamentary campaigns in which he was heckled as a "man of blood." His claim that he had been acting in self-defense was itself put forward in aggressive terms that did not reassure public opinion.

Stanley was prevented by tribal unrest from investigating Lake Albert, and he turned south to Lake Tanganyika, which he also circumnavigated, finding no outlet to connect it with Lake Albert and the Nile system.

By the end of 1876 Stanley was ready for what he called "the grandest task of all"—the identification of the Lualaba River, which Livingstone had hoped might be the Nile

but thought more likely to be the Congo. In 1876 the expedition arrived at Nyangwe, the northernmost point reached by Livingstone on the Lualaba, and set off down the river with Tippu Tib, the leading Arab trader of the region.

Because of the extreme difficulties and danger of the journey, the Arabs turned back at the limit of their trading area. Undaunted, Stanley launched the "Lady Alice" and sailed down the unexplored river to the lake he called Stanley Pool; further on, he encountered a 280-foot drop in the river and 32 cataracts, which he named Livingstone Falls. Unable to navigate further, he was forced to proceed overland on foot to the Atlantic Ocean, which he reached on August 12, 1877. His three companions and half his Africans were dead, but he had traced the Congo. Stanley described this tremendous journey in *Through the Dark Continent* (1878).

Founding of Congo Free State

Failing to enlist British interests in the development of the Congo, Stanley accepted the invitation of King Leopold II of the Belgians to lead an expedition there under the auspices of the Comité d'Études du Haut Congo. This offshoot of Leopold's International African Association (founded 1876) was transformed later into the Association Internationale du Congo, out of which grew the Congo Free State. From August 1879 to June 1884 Stanley was in the Congo basin, where he made a road from the lower Congo up to Stanley Pool, launched steamers on the upper river, and built a chain of stations. His perserverance in the face of difficulties earned him, from his men, the nickname of Bula Matari—"the rock breaker." The Berlin West Africa Conference of 1884–85, which Stanley attended, recognized the Congo Free State under the sovereignty of King Leopold.

Relief of Emin Paşa. Stanley's last expedition in Africa was for the relief of Mehmed Emin Paşa, governor of the Equatorial Province of Egypt. Emin had been cut off by the Mahdist revolt of 1882 in the neighbourhood of Lake Albert. Sir William Mackinnon, chairman of the British India Steam Navigation Company, launched a rescue fund and invited Stanley to lead the rescue expedition. Stanley chose to travel by way of the Congo; he knew the route and relied on Tippu Tib to supply porters. He left England in January 1887 and, stopping to recruit men in Zanzibar, arrived at the mouth of the Congo in March. The expedition reached the navigable head of the river in June, and there, at Yambuya, Stanley left a rear column with orders to await Tippu Tib's porters and to follow him later if these failed to arrive.

Stanley and four associates led the advance guard out of the camp on June 28, reaching Lake Albert five months later after a gruelling march through dense forest. They did not contact Emin until April 1888 when Stanley found, to his astonishment, that Emin was unwilling to evacuate his province. Leaving him to assemble his troops, Stanley then returned to look for the missing rear column. He reached Banalia, just short of Yambuya, on August 17, where he found the remains of the column, depleted and demoralized. Tippu Tib had not sent the porters, and Stanley's officers were dead—one had been murdered, and the other died of fever soon after.

Stanley blamed the dead officers for the disaster, and on his return to England a controversy arose about the fate of the rear guard. This was the most bitter of the many quarrels that heavily damaged Stanley's reputation.

Back at Lake Albert in January 1889, Stanley heard that Emin had been detained by disaffected troops, but eventually the party was reunited, and on April 10, 1889, some 1,500 persons set out, arriving at Bagamoyo on the east coast on December 4. Stanley cleared up some other geographical points on this expedition. The mist-covered Ruwenzori Range, which had been concealed when Lake Albert had been discovered in 1864, and which Stanley himself had passed unnoticed in 1876, was identified as Ptolemy's Mountains of the Moon; the size of Lake Albert was correctly estimated, and its tributary, the Semliki River, traced south to the lake that Stanley named Albert Edward (now Lake Edward). *In Darkest Africa* (1890) is Stanley's account of this arduous and unlucky expedition.

Parliament. Stanley married Dorothy Tennant on July 12, 1890, and they adopted a son, Denzil. Stanley was renaturalized a British subject in 1892 (he had become a United States citizen on May 15, 1885) and sat in Parliament as Liberal Unionist member for North Lambeth from 1895 to 1900.

In 1897 he visited South Africa and wrote *Through South Africa* (1898). He was awarded a Grand Cross of the Bath in 1899, becoming Sir Henry Morton Stanley. His last years were spent mainly at Furze Hill near Pirbright, Surrey, a small estate bought in 1898.

Stanley died in London on May 10, 1904, and was buried at Pirbright after a service in Westminster Abbey.

Significance. From a modern viewpoint, much contemporary criticism of Stanley appears unjust. To the select band of travellers whose centre was the Royal Geographical Society, dedicated to the pursuit of knowledge, Stanley appeared as an outsider, an American journalist with huge funds at his disposal. His motives in exploring were suspect, and his methods were considered to be rough. His humble origin probably mattered less than he supposed to men who came, in fact, from a variety of social backgrounds, but his uncouth and uncooperative manners and his tendencies toward self-righteousness and boastfulness ensured him an unpopularity that long coloured his reputation. Harsh and humourless, touchy and intolerant, Stanley antagonized many who should have been his friends, and the controversies that arose around him have tended to obscure his considerable achievements. He was the bridge between the old and the new; at first the intrepid explorer, the friend of Livingstone, and the discoverer of the Congo; then the pioneer who made possible the Congo Free State; finally, the British elder statesman who lived into the age of the great colonialist Cecil Rhodes and observed the opening up of modern Africa.

BIBLIOGRAPHY. *The Autobiography of Henry Morton Stanley*, ed. by DOROTHY STANLEY (1909), was completed by his wife after Stanley's death; it contains the only firsthand account of his early years, otherwise scantily documented. F. HIRD, *H.M. Stanley: The Authorized Life* (1935), is a reliable account of his career; usefully supplemented by IAN ANSTRUTHER, *I Presume* (1956); and BYRON FARWELL, *The Man Who Presumed* (1957). *Stanley's Despatches to the New York Herald, 1871–72, 1874–77*, ed. with an illuminating introduction by NORMAN R. BENNETT (1970), is a reassessment of Stanley and his place in the history of 19th-century exploration. The books in which Stanley described his adventures can be compared usefully with complementary sources. A partial check on *Through the Dark Continent* (1878) is provided by *The Exploration Diaries of H.M. Stanley*, ed. by RICHARD STANLEY and ALAN NEAME (1961), containing a hitherto unpublished record of part of the transcontinental expedition of 1874–77. *The Congo and the Founding of Its Free State*, 2 vol. (1885), should be read with reference to *Stanley: lettres inédites*, ed. by ALBERT MAURICE (1955; Eng. trans., *H.M. Stanley: Unpublished Letters*, 1957), covering the years 1881–84. Stanley's *In Darkest Africa*, 2 vol. (1890), should be supplemented by the following: W.G. BARTTELOT (ed.), *The Life of Edmund Musgrave Barttelot* (1890); DOROTHY MIDDLETON (ed.), *The Diary of A.J. Mounteney Jephson* (1969); and T.H. PARKE, *My Personal Experiences in Equatorial Africa* (1891). A good general account is A.J. WAUTERS, *Stanley au secours d'Émin-Pacha* (1890; Eng. trans., *Stanley's Emin Pasha Expedition*, 1890), by the Belgian geographer who was much concerned in the preparations for the venture. A modern reassessment based on exhaustive research into published and unpublished sources is IAIN R. SMITH, *The Emin Pasha Relief Expedition, 1886–90* (1971). MARCEL LUWEL, *Stanley* (1959, in French), includes a good general bibliography.

(D.Mi.)

Star

Stars are massive, self-luminous objects, shining by radiation derived from internal energy sources. With respect to mass, radius, and luminosity, the Sun is a typical star. Its mass is 1.99×10^{30} kilograms (more than 300,000 times that of Earth); its radius is 696,000 kilometres; its power output is 3.8×10^{23} kilowatts. The masses, radii, and luminosities (total intrinsic brightness) of other stars are often expressed in terms of those of the Sun.

Table 1: The Twenty Brightest Stars

name	visual magnitude* and spectrum		distance in light-years†	visual luminosity ⊙	
	A‡	B‡		A‡	B‡
Sirius	−1.50 A1 V	+7.20 dA5	8.6	23.0	0.0080
Canopus	−0.73 F0 Ib		98.0	1,450.0	
Alpha Centauri	+0.10 G2 V	+1.50 K5 V	4.3	1.3	0.3600
Vega	+0.04 A0 V		26.0	52.0	
Arcturus	0.00 K2 III		36.0	110.0	
Capella	+0.05 G III	+10.20 M1 V	45.0	160.0	
Rigel	+0.08 B8 Ia	+6.60 B9	600.0	25,000.0	70.0000
Procyon	+0.34 F5 IV	+10.80 d	11.4	7.6	0.0005
Betelgeuse	+0.41 M2 I (var.)		600.0	21,000.0	
Archernar	+0.47 B5 IV-V		65.0	210.0	
Beta Centauri	+0.61 B1 II	B2(?)	300.0	3,600.0	?
Altair	+0.77 B7 V		16.6	10.0	
Alpha Crucis	+1.58 B1 IV-V	+2.09 B3 V	390.0	2,760.0	1,600.0000
Aldebaran	+0.86 K5 III	+13.00 M2 V	52.0	100.0	0.0013
Spica	+1.12 B1.5 V	+3.10 B3 V	274.0	2,000.0	330.0000
Antares	+0.90 M1 Ib (var.)	+5.10 B4e V	420.0	6,000.0	115.0000
Pollux	+1.15 K0 III		37.0	170.0	
Fomalhaut	+1.18 A3 V	+6.50 K4 V	22.6	13.0	0.1000
Deneb	+1.26 A2 Ia		1,400.0	47,000.0	
Beta Crucis	+1.24 B0.5 V		500.0	5,700.0	

*Negative magnitudes are brightest, and one magnitude difference corresponds to a difference in brightness of 2.5 times; *e.g.*, a star of magnitude −1 is ten times brighter than one of magnitude +1.5. †One light-year equals about 5.88×10^{12} miles (9.46×10^{12} kilometres). ‡A and B are brighter and fainter components, respectively, of star.
Source: From H.C. van de Kamp; C.W. Allen; R. Hanbury Brown; J. Davis, L.R. Allen and J. Rome; G. Abell.

Table 1 gives approximate data pertaining to the 20 brightest stars, or, more precisely, stellar systems, since some of them are double (binary) or even triple stars. Successive columns give the name of the star, its brightness expressed in units called visual magnitudes and in the quality of the light radiated (that is, the spectral class or classes to which the star [or its components] belongs), the distance in light-years, and the visual luminosity in terms of the Sun, symbol ⊙ in Table 1 and the rest of the article. All of the primary stars in Table 1 are intrinsically as bright as or brighter than the Sun; some of the companion stars are fainter. Many stars vary in the amount of light they radiate. Stars such as Altair, Alpha Centauri, and Procyon are called dwarf stars; their dimensions are roughly comparable to those of the Sun. Sirius and Vega, though much brighter, are also dwarf stars; their higher temperatures yield a large rate of emission per unit area. Aldebaran, Arcturus, and Capella are examples of giant stars, whose dimensions are much larger than those of the Sun. Thus, measurements with an interferometer (an instrument that measures the star's angular diameter; *i.e.*, the angle subtended by the diameter at the observer's position) combined with parallax measurements, which give the star's distance, give diameters of 24 and 45 solar diameters for Arcturus and Aldebaran, respectively. Betelgeuse and Antares are examples of supergiant stars; the latter has a diameter 560 times that of the Sun, whereas the diameter of the variable star Betelgeuse (at a distance of 600 light-years) varies between 600 and 900 solar diameters. Other supergiants are Alpha Herculis and the variable star Mira Ceti, whose diameters are 480 and 420 times that of the Sun. Several of the stellar class of white dwarf stars, which have low luminosities and high densities, are also listed.

Range of stellar sizes

The 20 nearest stars. Table 2 gives information about the 20 nearest known stars. Only Alpha Centauri, Procyon, and Sirius appear in both lists. All other nearby stars are intrinsically fainter than the Sun, and most of them can be seen only with the telescope. The very luminous stars can be seen at great distances; intrinsically faint stars can be seen only if they are very near the Sun. Several of the nearby stars have planet-like companions.

Tables 1 and 2, which pertain to only a small number of stars, illustrate some important points. The stars listed fall roughly into three groups: (1) dwarfs, with radii that range between one-tenth solar radius and five solar radii and masses from 0.07 to about 20 solar masses; (2) giant and supergiant stars, with radii tens or even hundreds of solar radii and extremely low average densities; and, finally, (3) white dwarf stars, which may have masses comparable to that of the Sun but dimensions appropriate to planets. Hence, they have average densities hundreds of thousands of times greater than that of water.

Table 2: The Twenty Nearest Stars

name	visual apparent magnitude and spectrum		distance in light-years	visual luminosity ⊙	
	A*	B*		A*	B*
Alpha Centauri†	0.1 G2 V	1.5 K5 V	4.3	1.30	0.36000
Barnard's Star‡	9.5 M5 V		5.9	0.00044	
Wolf 359	13.5 M6e		7.6	0.00002	
Lalande 21185‡	7.5 M2 V		8.1	0.0052	
Sirius§	−1.5 A1 V	7.2 dA5	8.6	23	0.00800
Luyten 726–8‖	12.5 M6e V	13.0 M6e	8.9	0.00006	0.00004
Ross 154	10.6 M5e V		9.4	0.0004	
Ross 248	12.2 M6e V		10.3	0.00011	
Epsilon Eridani	3.7 K2 V		10.7	0.30	
Luyten 789–6	12.2 M6 V		10.8	0.00012	
Ross 128	11.1 M5 V		10.8	0.00033	
61 Cygni¶	5.2 K5 V	6.0 K7 V	11.2	0.083	0.04000
Epsilon Indi	4.7 K5 V		11.2	1–87 = 0.13♀	
Procyon♁	0.3 F5I V	10.8 d	11.4	7.6	0.00050
Σ 2298	8.9 M3.5 V	9.7 M4 V	11.5	0.0028	0.00130
Groombridge 34	8.1 M1 V	11.0 M6 V	11.6	0.0058	0.00040
Lacaille 9352	7.4 M2 V		11.7	0.012	
Tau Ceti	3.5 G8 V		11.9	0.44	
BD+5°1668	9.8 M4 V		12.2	0.0014	
Lacaille 8760	6.7 M1 V		12.5	0.025	

*A and B are brighter and fainter components, respectively, of star. †The two components of Alpha Centauri have masses of $M_1 = 1.0$ and $M_2 = 0.9$ solar masses, respectively. They are separated by 23.1 astronomical units and revolve around one another with a period $P = 79.9$ years. A third component of the system is an 11th magnitude M5 dwarf with a luminosity $L = 0.00006\ L_{Sun}$. ‡These stars have unseen, planet-like companions whose presence is revealed by gravitational attraction on the visible star. §The two components of Sirius are separated by 19.9 astronomical units, $P = 50.1$ years. The bright component has 2.2 solar masses; the faint component is a white dwarf of 0.9 solar masses. ‖The mean separation of these stars is 15.3 astronomical units, $P = 200$ years, but the masses are 0.05 and 0.04 Suns; they probably do not shine by nuclear energy generation but are radiating away gravitational energy. ¶Binary with mean separation $\alpha = 83.2$ astronomical units, $P = 653$ years, $M_1 = 0.7$, $M_2 = 0.7$ solar masses. The brighter component has a planet-like companion circling it with a period $P = 4.8$ years and a mass of about $0.008\ M_{Sun}$. ♀1–87 = fainter. ♁Binary separation = 15.8 astronomical units, $P = 40.6$ years, $M_1 = 1.8 M\odot$, $M_2 = 0.6\ M\odot$. The fainter component is a white dwarf star.
Source: From H.C. van de Kamp.

These rough groupings of the stars correspond to stages in their life histories (see below *Stellar evolution*). The first category, which will be identified with what is called the main sequence (see below), contains stars that obtain their energy mainly by converting hydrogen into helium in their cores. The second category comprises stars that have exhausted the hydrogen in their cores and are burning hydrogen in small shells surrounding the core. The white dwarfs represent the final stage in the life of a typical star, at a point when most available sources of energy are used up and the star becomes relatively dim.

The large number of binary and even multiple systems is notable. These systems are built on scales comparable with that of the solar system. Some of the nearby single stars have invisible companions, which can be detected only by their gravitational effects upon the primary star, which cause it to move through space not along a straight line but following a sinusoidal (wavelike) curve. These invisible companions are not self-luminous; they have masses of the order of 0.01 solar mass or even less and, properly, may be considered as planets. The implication that may reasonably be drawn from these data is that double stars and solar systems are formed by similar evolutionary processes.

Invisible companion stars

Before there can be any adequate discussion of the masses, dimensions, and luminosities of various types of stars, it is necessary to know their distances. For making this determination, there are several methods, some of them of a statistical nature, permitting only the average distances of various types of objects to be ascertained. The quantity and character of the electromagnetic radiation the stars emit must also be measured.

Distances of the stars. This subject is treated in detail in the article PARALLAX, ASTRONOMICAL. Summarized here are a few highlights and essential definitions.

The basic method is determination of trigonometric parallax. If the position of a nearby star is measured from two points on opposite sides of the Earth's orbit, a small angular displacement will be observed. Using the Earth's orbit as the base line, the distance of the star can be found from the angular size of the parallax p (see Figure 3; PARALLAX, ASTRONOMICAL). If $p = 1''$ (one second) of arc, the distance of the star would be 206,265 times the Earth's distance from the Sun, that is, 3.26 light years. This unit of distance is called the parsec, derived from *par*allax of one *sec*ond of arc. One light-year, the distance through which light waves travel in one year, is a popularly used unit; 3.26 light-years are equivalent to one parsec.

Parallax

Parallax is inversely proportional to distance. Thus, a star at ten parsecs has a parallax of 0.1″ of arc. Stars at greater distances have smaller parallaxes, and these are more difficult to measure directly. Hence, indirect methods depending on the absolute brightness of different classes and other classifications are used for the majority of stars. The nearest star, the triple system of Alpha Centauri, has a parallax of 0.760″ of arc. Hence, its distance is 1/0.760, or 1.31 parsecs, or 4.3 light-years. Probable errors of such parallaxes are typically 0.005″ of arc, which means that there is a 50:50 chance that a star whose parallax is 0.01″ of arc lies between 67 and 200 parsecs (corresponding to parallaxes of 0.015 and 0.005, respectively) and an equal chance that it lies outside that range. Thus, measurement of trigonometric parallaxes are of little value beyond the nearby stars.

Identification of individual stars: nomenclature. The earliest designations of the stars still in common use are their Arabic names. These follow no particular system, are often fanciful in character, and are rarely used except by navigators and amateurs. The brighter stars in each constellation are now denoted by a system introduced in 1603 by the German astronomer Johann Bayer. The star appearing brightest in a constellation is denoted as α (alpha), the next brightest as β (beta), etc. When stars are of comparable brightness, as in the Big Dipper, they are lettered according to their position in the constellation. Later, the first British Astronomer Royal, John Flamsteed, introduced a system of numbers assigned in order of the stars' position coordinate, right ascension.

Star catalogs

Stars too faint to be given a Bayer letter or Flamsteed number are listed in other ways, often by number, in star catalogs, one of the most extensive of which is the *Bonner Durchmusterung* (BD). It has been extended to cover the Southern Hemisphere by the *Córdoba Durchmusterung* (CD) and the *Cape Photographic Durchmusterung* (CPD). Star charts are provided for each of these catalogs, so that the identification of a given star is not difficult. Many stars are known only by the number listed for them in one or more of the major catalogs. Perhaps the most important is the *Henry Draper Catalogue* (HD), which with its extensions contains data for 225,000 stars (see also ASTRONOMICAL MAPS).

Variable stars

Variable stars—*i.e.*, stars variable in brightness—may be listed in one or more of such lists and, if bright enough, may have been assigned a name, a Flamsteed number or a Bayer letter, or all three. When known to be variable, however, they are also assigned a capital letter or letters, written with the possessive or genitive case of the constellation. RR Lyrae, R Coronae Borealis, and U Cephei, for example, are such variables and are important as prototypes, being the first discovered with the particular characteristics of a whole class of variables now known by their name. RR Lyrae stars, for example, are periodic variables and a subclass of the larger group called Cepheids (after the prototype, Delta Cephei).

After single letters and combinations of double letters have been exhausted in a given constellation, the next discovered variable star in that constellation is assigned a number, and after that consecutively increasing numbers are assigned to subsequent discoveries as they occur.

This article is divided into the following sections:

I. Observable data for stars
  Stellar positions
    Fundamental measurements
    Stellar motions
  Brightness of the stars
    Stellar colours
    Magnitude systems
  Stellar spectra
    Spectral classes
    Effects of stellar density and luminosity
II. Derived, or calculated, stellar characteristics
  Intrinsic stellar brightness
    Absolute magnitudes
    Total luminosities; bolometric magnitudes
  Stellar masses
  Stellar diameters
  Stellar temperatures
  Average values of temperatures, radii, masses, and luminosities for dwarf stars
III. Stellar variability
  Geometric variables
  Intrinsic variables
    Pulsating stars
    Explosive variables
IV. Statistics of stars
  Correlations among luminosity, spectrum, mass, and radius
    Relation between the number of stars and their luminosities
    Mass–luminosity correlation
  Statistics of binary systems
  Statistics of special types of stars
V. Stellar structure
  Stellar atmospheres
  Internal structure of stars
    Relations obtained from physical laws
    Energy generation in stars
VI. Stellar evolution
  The life history of a typical star
    Protostars
    The star reaches the main sequence
    Evolution away from the main sequence
    Estimates of stellar ages
  Formation of chemical elements in stars
  Probable fate of a star: white dwarfs, pulsars, and black holes
    White dwarfs and novae
    Neutron stars
    Black holes

I. Observable data for stars

STELLAR POSITIONS

Fundamental measurements. Accurate measurements of stellar positions are essential to many problems of stellar astronomy. Positions of the brighter stars can be measured very accurately in the equatorial coordinate system (the coordinates are called right ascension [α, or RA] and declination [δ] and are given for some epoch). Fainter stars are measured upon photographic plates with respect to these brighter stars, and finally the entire group is referred to the positions of known external galaxies. These external galaxies are far enough away to define an essentially fixed, or immovable, system, whereas

positions of both the bright and faint stars are affected over quite short periods of time by galactic rotation and by their own motions through the Galaxy.

The *General Catalogue* gives precise positions for about 30,000 stars. Even more accurate data are found in the *Dritter Fundamental-Katalog des Berliner astronomischen Jahrbuchs* (FK 3) or its revision (the FK 4).

Stellar motions. Accurate measurements of position (for example α, δ) make it possible to determine the motion of a star across the line of sight (proper motion). The amount of the proper motion denoted by μ (in seconds of arc per year) divided by the parallax of the star and multiplied by 4.74 equals the tangential velocity, T, in kilometres per second in the plane of the celestial sphere. Thus, tangential velocity = (annual proper motion/parallax) $\times$ 4.74 km per sec, or $T = 4.74$ (μsec/parallax).

Tangential and radial velocity

The motion in the line of sight (called radial velocity) is obtained directly from spectroscopic observations (Figure 1). If λ is the wavelength of a characteristic spectrum



Figure 1: Components of stellar motion. V represents radial velocity; T is tangential velocity of the star S with respect to the Sun; ν is the total velocity (see text).

line of some atom or ion present in the star, and λ_L is the wavelength of the same line on the Earth, then the difference $\Delta\lambda$, or $\lambda - \lambda_L$, divided by λ_L is equal to the radial velocity V divided by the velocity of light, c: $\Delta\lambda/\lambda_L = V/c$. Shifts of the spectral line toward red (positive V) indicate recession. If the parallax is known, measurements of μ and V make it possible to determine the space motion of the star. Radial velocities are corrected for the Earth's rotation and for its motion around the Sun so that they will refer to the line-of-sight motion of the star with respect to the Sun.

The observed motion of a given star is made up of two parts: the motion of the star with respect to the average of the nearby stars (the local standard of rest) and the motion of the Sun with respect to this same standard of rest. The first problem to be solved is the determination of the Sun's motion, both in velocity and in direction, with respect to the nearby stars. This directional, or vector, quantity is established most reliably from radial velocity data. Stars in the region of the sky toward which the Sun is moving (called the apex of the Sun's way) show a predominantly negative velocity (approach), whereas those in the opposite region of the sky (the antapex) show a positive velocity (recession). Stars at a position 90° away from the apex or antapex show a random mixture of positive and negative radial velocities. A statistical analysis of the observed radial velocities permits the determination of the apex and the velocity of the Sun, $V_\odot$, with respect to the average of the nearby stars. For the naked-eye stars, the apex is not far from the star Mu Herculis. With respect to fainter stars or special types, such as RR Lyrae variables, the position of the apex will be different, since these objects have their own systematic motion.

Components of proper motion

Once the direction of the solar motion is known, the proper motion of any star can be divided into two components, one along the great circle passing through the apex, the star, and the antapex, and the other perpendicular thereto. The first, or upsilon, component contains the effects of both the star's own peculiar motion and the reflection of the Sun's motion. A backward drift caused by the Sun's motion will be a maximum for stars 90° from the apex and zero at both the apex and antapex. The second, or tau, component is unaffected by the solar motion.

Special efforts have been made to detect stars of large proper motion, since these tend to be nearby, often being intrinsically very faint stars. Searches for stars of large parallax are often based on lists of stars of large proper motion. Although proper motions can be measured for very faint stars, radial velocities can be determined only for stars whose spectra can be photographed with sufficient dispersion.

BRIGHTNESS OF THE STARS

Brightness and magnitude scales

Stellar brightnesses are usually expressed by means of their magnitudes, a usage inherited from classical antiquity. A star of the first magnitude is about 2.5 times as bright as one of the second magnitude, which in turn is about 2.5 times as bright as one of the third magnitude, and so on, so that a star of the first magnitude is 100 times as bright as one of the sixth magnitude. The magnitude of Sirius, apparently the brightest star in the sky, is −1.5. Canopus, the second-brightest, has a magnitude of −0.73, while the faintest star that can normally be seen without telescopic aid is of the sixth magnitude. Stars as faint as the 23rd magnitude have been measured with the 200-inch telescope on Palomar Mountain in California.

The scale of magnitudes constitutes a geometrical progression of brightness. Magnitudes may be converted to light ratios by letting l_n and l_m be the brightnesses of stars of magnitudes n and m, respectively; the logarithm of the ratio of the two brightnesses is then equal to 0.4 times the difference between them—*i.e.*, $\log (l_m/l_n) = 0.4\ (n - m)$. Magnitudes are actually defined in terms of observed brightness, a quantity that depends on the light-detecting device employed. Visual magnitudes are measured with the eye, photographic magnitudes from images on a photographic plate; and there are various types of photoelectric magnitudes. A magnitude system must have a reference, or zero, point. In practice this is fixed arbitrarily and is essentially based on the star catalogs of Ptolemy and Hipparchus.

Absolute magnitude

The actually measured brightnesses give apparent magnitudes. These cannot be converted to intrinsic brightnesses until the distances of the objects concerned are known. The absolute magnitude of a star is defined as the magnitude it would have if it were viewed at the standard distance of ten parsecs (32.6 light-years, or 2,062,650 astronomical units; an astronomical unit is the average Earth–Sun distance). The apparent visual magnitude of the Sun is −26.9, that of the Full Moon is −12.7, and that of the brightest star is −1.5. The visual absolute magnitude of the Sun corresponds to a diminution in brightness by a factor $(2{,}062{,}650)^2$ and is, using logarithms, $-26.9 + 2.5 \times \log (2{,}062{,}650)^2$, or $-26.9 + 31.6$; that is, 4.7. The absolute magnitude of the Full Moon is about +32, and to see it at ten parsecs would require a telescope with an aperture about a quarter-mile wide. The very brightest stars, such as Rigel (Beta Orionis), have absolute magnitudes of −7 to −9 (see Table 1). The faintest known star, the companion to BD+4°4048, which was discovered by the U.S. astronomer G. Van Biesbroeck and is therefore often called Van Biesbroeck's star, has an absolute visual magnitude of 19—1,000,000 times fainter than the Sun. A large number of such faint stars may exist, but they are extremely difficult to detect.

Stellar colours. Stars differ in colour. Most of the naked-eye stars in the constellation Orion are blue-white, but Betelgeuse (Alpha Orionis) is a deep red. In the telescope, Albireo (Beta Cygni) is seen as two stars, one blue, the other orange. One quantitative means of measuring stellar colours is to compare the visual magnitude of the star with a magnitude measured on a photographic plate. Yellow, the colour to which the eye is most sensitive, contributes most to visual magnitude. The ordinary photographic plate is insensitive to yellow and red light but is sensitive to blue and violet, so that hot blue stars appear bright, and yellow and red stars are recorded as relatively faint. In all magnitude scales, one magnitude step corresponds to a brightness ratio of about 2.5. The zero point is chosen so that white stars with surface temperatures of about 10,000° K (about 17,500° F) have the same visual and photographic magnitudes. The conventional colour index is defined as the photographic magnitude minus the visual magnitude; the colour index of the Sun is $+5.46 - 4.84 = 0.62$.

Colour index

Visual magnitudes have been determined photographically with the aid of orthochromatic or yellow-sensitive

plates and yellow filters. Such photovisual magnitudes differ from visual magnitudes less than naked-eye estimates of visual magnitudes by different observers differ among themselves.

Fundamental measurements of stellar magnitudes and colours are now made photoelectrically, although the photographic plate continues to play a vital role as an interpolation device when very large numbers of stars are to be measured (see PHOTOMETRY, ASTRONOMICAL).

Magnitude systems. The older system of stellar magnitudes, normally employed only for photographic and photovisual (or visual) magnitudes, encountered a number of severe difficulties. Furthermore, when only one colour is observed, if, for instance, the star's colour is found to be red, without further information it is impossible to decide whether the star is red because it is cool or whether it is really a hot star whose colour has been reddened by the passage of the light through material in space. By measuring the magnitudes of the same stars in three or more colours, it is possible to overcome these difficulties.

Modern magnitude systems are based on measures, usually made photoelectrically, of a few standard stars, the measurement accuracy being limited chiefly by the quality of the atmosphere through which the light has to pass. Magnitudes that measure light from only parts of the wavelength range are usually named for the colour of the light. Thus, there are blue magnitudes, M_B, or B, whose range roughly corresponds to that of the (blue) photographic magnitudes; visual magnitudes, M_v, or V, whose range is centred in the yellow-green part of the spectrum, red magnitudes; and so on.

In the 1960s, measurements in infrared light assumed considerable importance (see below). Also, photometric observations of individual stars from space vehicles and rockets made possible the measurement of stellar colours over a large range in wavelength. These data are important for hot stars and for assessing the effects of interstellar extinction.

The measured total of all radiation at all wavelengths from a star is called a bolometric magnitude and is described below.

STELLAR SPECTRA

A star's spectrum contains information about its temperature, chemical composition, and intrinsic luminosity. Spectrograms secured with a slit spectrograph consist of a sequence of images of the slit in the light of the star at successive wavelengths; adequate dispersion may show the star to be a member of a close binary system, in rapid rotation, or to have an extended atmosphere. Quantitative determination of its chemical composition then becomes possible. Inspection of a high-dispersion spectrum of the star may reveal evidence of a strong magnetic field.

Spectral classes. Most stars are grouped into a small number of spectral classes. The *Henry Draper Catalogue* lists spectral classes from the hottest to the coolest stars, designated by the letters O, B, A, F, G, K, M. This group is supplemented by R and N stars (now more frequently referred to as the carbon C stars) and S stars. The R, N, and S stars differ from the others in chemical composition; also, they are invariably giant or supergiant stars (see ASTRONOMICAL SPECTROSCOPY, PRINCIPLES OF).

Conditions in stellar atmospheres

The spectral sequence O–M represents stars of essentially the same chemical composition but of different temperatures and atmospheric pressures. This simple interpretation, put forward in the early 1920s by an Indian physicist, Meghnad Saha, has provided the physical basis for all subsequent interpretations of stellar spectra. The spectral sequence is also a colour sequence; the O and B stars are intrinsically the bluest and hottest; the M, R, N, and S stars are the reddest and coolest. When the effects of space reddening are eliminated, and stars of a given kind, such as dwarf stars, are considered, there is a one-to-one correspondence between colour and spectral class or between spectral class and temperature. These dwarf stars form what is known as the main sequence in the so-called Hertzsprung–Russell diagram (see below).

In the case of the cool stars of class M, the spectra indicate the presence of familiar metals, such as iron, calcium, magnesium, and so on, and also (particularly in the red and green parts of the spectrum) titanium oxide molecules. In the somewhat hotter K stars the titanium oxide bands disappear, and the spectrum exhibits a wealth of metallic lines. A few especially stable fragments of molecules such as cyanogen and the hydroxyl radical persist in these stars and even in G stars, such as the Sun. The spectra of G stars are dominated by the characteristic lines of metals, particularly of iron, calcium, sodium, magnesium, and titanium.

The behaviour of calcium illustrates the phenomenon of thermal ionization. At low temperatures a calcium atom retains all its electrons and radiates a spectrum characteristic of the neutral, or normal, atom; at higher temperatures, collisions between atoms and electrons and the absorption of radiation both tend to detach electrons and to produce singly ionized calcium atoms. At the same time these ions may recombine with electrons to produce neutral calcium atoms. At high temperatures or low electron pressures, or both, most of the atoms are ionized. At low temperatures and high densities the equilibrium favours the neutral state. The concentrations of ions and neutral atoms can be computed from the temperature, the density, and the ionization potential (the energy required to detach an electron from the atom, usually expressed in electron volts).

The line of neutral calcium at 4227 angstroms thus is strong in cool M dwarf stars, in which the pressure is high and the temperature is low; but in the hotter G stars the lines of ionized calcium at 3968 angstroms and 3933 angstroms (called the H and K lines) become much stronger than any other feature in the spectrum.

In stars of spectral class F, lines of neutral atoms become weak relative to those of ionized atoms. The hydrogen lines become stronger, attaining their maximum intensities in A stars, in which the temperature is around 9,000° K. Thereafter they gradually fade as the hydrogen becomes ionized.

Production of spectral lines

Spectral lines are produced by transitions of atoms or ions between energy levels, when energy in the form of light is absorbed or emitted. The yellow D lines of sodium or the H and K lines of ionized calcium are produced by jumps from the lowest energy levels (ground levels) of these atoms. The observable hydrogen lines, however, are produced by jumps from atoms in the so-called second energy level, which lies far above the ground level. Only at high temperatures are sufficient numbers of atoms maintained in this level by collisions, radiations, etc., to permit an appreciable number of absorptions to occur. At the low temperatures of the M or even of the K stars there are few atoms in the second level. On the other hand, at the very high temperatures of the B and O stars, the hydrogen atoms are nearly all ionized and therefore cannot absorb or emit any line radiation whatsoever, so that, at best, only faint hydrogen lines are seen. The lines of ionized metals such as iron in the F stars never attain the strength that the neutral lines do in the K stars. These lines often involve higher energy levels, which are more sparsely populated than the lower levels. Another factor is that the general fogginess or opacity of the atmospheres of these hotter stars becomes greatly increased so that there are fewer and fewer atoms in the visible layers that can produce the observed lines.

Opacity of the star's outer layers

The continuous (as distinct from the line) spectrum of the Sun is produced primarily by the photo-dissociation of negative hydrogen ions; that is, atoms of hydrogen to which an extra electron is loosely attached. In the solar atmosphere, when these negative hydrogen ions are subsequently destroyed by photodissociation, they absorb energy at any of a whole range of wavelengths in the process and thus produce a continuous range of absorption of radiation. The principal source of absorption of light in the hotter stars is the photo-ionization of hydrogen atoms, both from ground level and from higher levels.

Photoelectric absorption by normal hydrogen atoms in a cool environment is confined to the very short wavelength ultraviolet; but if the hydrogen atoms become

excited to higher energy levels, the photo-ionization energies then required are much lower, and the continuous absorptions fall in the visual range. Other usually less important sources of absorption leading to opacity in the stars' atmospheres exist (see below).

The hot B stars, such as Epsilon Orionis, are characterized by lines of helium, singly ionized oxygen, nitrogen, and neon. In the very hot O stars, lines of ionized helium appear. Other prominent lines include those of doubly ionized nitrogen, oxygen, and carbon and trebly ionized silicon, all of which require more energy to produce.

Effects of stellar density and luminosity. The physical processes behind the formation of stellar spectra are well enough understood to permit determinations of temperatures, densities, and quantitative chemical compositions of stellar atmospheres. The star studied in greatest detail is, of course, the Sun, but other stars also have been investigated in detail.

Although the general characteristics of the spectral sequence can be understood in terms of a variation of temperature, many spectral features depend on the densities of the stars' atmospheric gases. These densities depend in turn on surface gravities. Dwarfs, with great surface gravities, tend to have high atmospheric (surface) densities; giants and supergiants, with low surface gravities, have relatively low surface densities.

The hydrogen lines provide an illustration in point. Normally, an undisturbed atom radiates a very narrow line. If its energy levels are disturbed by charged particles passing nearby, it will radiate at a wavelength near its characteristic wavelength. In a hot gas the range of disturbance of the hydrogen lines is very high, so that the spectrum line radiated by the whole mass of gas will be much spread out; the amount of the blurring depends on the density of the gas in a known fashion. Dwarf A stars, such as Sirius, show broad hydrogen lines with extensive "wings" where the line fades slowly out into the background, whereas supergiant A stars, with less dense atmospheres, show relatively narrow hydrogen lines. The relationship between the spectrum of a star and its intrinsic luminosity, called the luminosity effect, has to be established empirically.

M–K spectral classification system

In the modern system of spectral classification called the M–K system, for the American astronomers W.W. Morgan and P.C. Keenan who introduced it, luminosity class is assigned to the star along with the Draper spectral class. Thus, Alpha Persei is classed as F5 Ia, which means that it falls about halfway between the beginning of class F(F0) and of class G(G0). The Ia means that it is a particularly luminous supergiant. The star Pi Cephei, classed as G2 III, is a giant falling between G0 and K0 but much closer to G0. The Sun, a dwarf star of class G2, is classified as G2 V. A star of luminosity class II falls between supergiants and giants; one of class IV is called a subgiant.

Some stars are binaries or multiple stars, and their colours and spectral classes often provide important information relevant to stellar evolution. Other stars are variable in magnitude, colour, and spectrum (see below).

II. Derived, or calculated, stellar characteristics

INTRINSIC STELLAR BRIGHTNESS

Absolute magnitudes. Positions, apparent magnitudes, and colours can all be measured for vast numbers of stars. If the star is close enough, its parallax can be measured directly; for more remote stars the moving-cluster method, which depends on similarity of proper motions, extends the range further. The intrinsic luminosities of many types of bright stars and average luminosities for certain variables have been established from their membership in open clusters. Such clusters contain stars of intrinsic brightnesses and colours similar to those of the Sun and its neighbours. As has been pointed out, stellar brightnesses are expressed in terms of their absolute magnitudes; *i.e.*, the brightness the star would have at a distance of ten parsecs. If the distance is known, the absolute magnitude, *M*, and corresponding luminosity can be found by simple arithmetic from the apparent brightness of the cluster stars.

Total luminosities; bolometric magnitudes. The corrections required to reduce visual magnitudes to bolometric magnitudes (see above) are large for very cool and for very hot stars but are relatively small for stars such as the Sun (see Figure 2). The determination of the true total luminosity of a star affords a measure of its actual energy output. When the energy radiated by a star is observed at the Earth, only that portion to which the energy detector is sensitive and that can be transmitted through the atmosphere is recorded. Most of the energy of stars like the Sun is emitted in spectral regions that can be observed from the Earth's surface; but an M star with a surface temperature of 3,000° K has an energy maximum on a wavelength scale at 10,000 angstroms, far in the infrared, and most of its energy cannot therefore be measured as light. Bright cool stars can be observed at infrared wavelengths, however, with special instruments that measure the amount of heat radiated by the star. Corrections for the heavy absorption of the heat by water molecules and others in the Earth's atmosphere must be made. Such magnitudes are referred to as radiometric magnitudes (see also INFRARED SOURCES, ASTRONOMICAL).

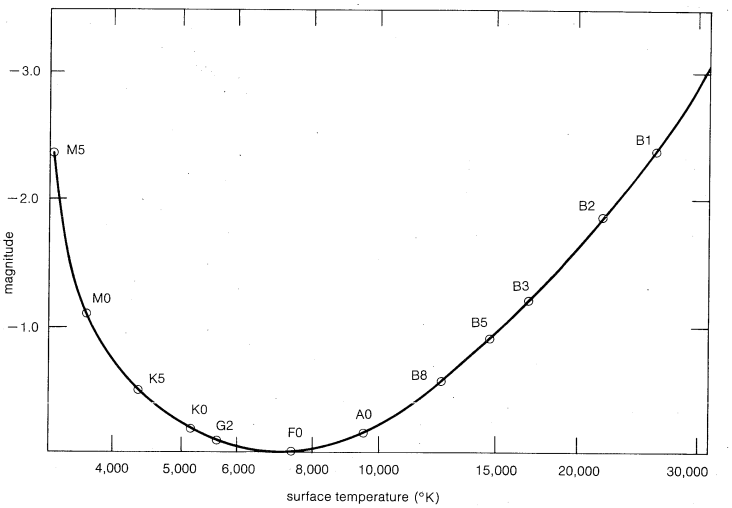


Figure 2: Bolometric correction. Dwarf and main sequence stars of twelve spectral classes are shown on the curve. Cool and hot stars have large corrections. The corrections always increase the stellar luminosity (see text).

Spectra of hot stars

The hotter stars pose more difficult problems, since the Earth's atmosphere extinguishes all radiation at wavelengths shorter than 2900 angstroms. A star whose surface temperature is 20,000° K or above will radiate most of its energy in the inaccessible ultraviolet. Measurements made with detectors flown in rockets or satellites extend the observable wavelength region down to 1000 angstroms or less, although all radiation of distant stars is cut out below 912 angstroms, a region in which absorption by neutral hydrogen atoms in intervening space becomes effective.

To compare the true luminosities of two stars, the appropriate bolometric corrections must first be added to each of their absolute magnitudes; then the ratio of the luminosities can be calculated.

STELLAR MASSES

The masses of stars can be found only from binary systems, if the scale of the orbits of the stars around each other is known. Binary stars are divided into three categories, depending on the mode of observation employed. Visual binaries can be seen as double stars in the telescope. True doubles, as distinguished from optical (line-of-sight) doubles, move through space together and display a common space motion; sometimes a common orbital motion can be measured as well. Spectroscopic binaries are found from radial-velocity observations. A component of such a binary has a continuously changing periodic velocity that alters the wavelengths of its spectral lines in a rhythmic way; the velocity curve repeats itself exactly from one cycle to the next, and the motion

can be interpreted as orbital motion. In a few cases rhythmic changes in both components' lines can be measured. Eclipsing binaries are detected from the light variations they show when one star passes between the observer and the other. The best known example is Algol (Beta Persei); good data are known for many such systems (see ECLIPSE, OCCULTATION, AND TRANSIT).

Masses of visual binaries. Visual-binary systems permit a determination of stellar masses when the parallax, p, and from it the distance are known. The period, P, of the binary system is expressed in years, and the angular radius, a'', of the orbit (more accurately, its semimajor axis), can be measured directly; the distance measure provides a scale, so that the true dimensions of the semimajor axis, a, can be found. If, in astronomical units, $a = a''/p''$, the sum of the masses $m_1 + m_2$ is, by Kepler's law, proportional to the cube of the mean distance, a, between them and inversely proportional to the square of the period, P, in years. In symbols, $m_1 + m_2 = a^3/P^2 = a''^3/(p''^3P^2)$, in units of the Sun's mass. For example, for the binary 70 Ophiuchi, $P = 87.85$ years, $p'' = 0.199''$ of arc, $a'' = 4.55''$ of arc, so that $a = 22.8$ astronomical units, and $m_1 + m_2 = 1.56$ solar masses. From a measurement of the motions of the two components with respect to the background stars, the orbit of each star has been determined with respect to the common centre of gravity. The mass ratio $m_2/(m_1 + m_2)$ is 0.422; the individual masses are 0.90 and 0.66 solar masses.

The star known as 61 Cygni was the first whose distance was measured and the first stellar system found to contain a planet. Visually, it is a double star separated by 83.2 astronomical units. Its components move about one another with a period of 653 years. The brighter component has an invisible companion, with a mass 0.008 of that of the Sun, which moves about it at a distance of 4,200,000 kilometres with a period of about 4.8 years. According to all accepted criteria, this companion must be regarded as a bona fide planet. Careful examination of the motions of other stars in the neighbourhood of the Sun has suggested that other planetary companions may exist.

Planetary companions to stars

Masses of spectroscopic binaries. By 1967, orbital elements for 737 systems had been listed. The semimajor axes or the masses cannot be found for most spectroscopic binaries, since the angle between the orbit plane and the plane of the sky cannot be determined; the possible foreshortening introduces uncertainty. If two spectra can be observed, mass ratios can be found. If one spectrum alone can be observed, only a quantity called the mass function can be found, from which is calculated a lower limit to the stellar masses. If the spectroscopic binary is also observed to be an eclipsing system, the inclination of the orbit and often the masses of the individual stars can be found.

STELLAR DIAMETERS

Eclipsing binary systems provide extensive data on stellar dimensions. These systems include dwarf stars and also supergiant stars such as 31 Cygni and VV Cephei. Components of close binary systems are sometimes subject to evolutionary effects, mass exchange, and other disturbances that change the details of the spectra (see below).

Angular diameters of bright red giant and supergiant stars were first measured directly in the 1920s, using the principle of interference of light. Only bright stars with large angular diameters can be measured by this method.

Interferometer measurements

A more recently developed photon-correlation interferometric method is usable for bright stars of smaller apparent size; in this way the angular diameters of 15 stars ranging in brightness from Sirius (−1.46) to Alpha Pavonis (+1.74) have been measured.

A very recent method, called speckle interferometry, has been developed to reproduce the true disks of red supergiant stars and resolve spectroscopic binaries such as Capella. The speckle phenomenon is a rapidly changing interference diffraction effect seen in a highly magnified diffraction image of a star observed with a large telescope.

If the absolute magnitude of a star and its temperature are known, its size can be computed. The temperature determines the rate at which energy is emitted by each unit of area, and the total luminosity gives the total power output. Thus, the surface area of the star and, from it, its diameter may be estimated. This is the only way available for estimating the dimensions of white dwarf stars. The chief uncertainty here lies in choosing the temperature that represents the rate of energy emission.

STELLAR TEMPERATURES

Stellar temperatures may be defined in a number of ways. From the character of the spectrum and the various degrees of ionization and excitation found from its analysis, an ionization or excitation temperature can be derived.

Meaning of a star's temperature

A comparison of the V and B magnitudes yields a colour index, $m_B - m_V$, related to the colour temperature of the star. The colour temperature is therefore a measure of the relative amounts of radiation in two more or less broad-wavelength regions, while the ionization and excitation temperatures pertain to the temperatures of strata wherein spectral lines are formed.

If the angular diameter of a star can be measured (see above) and if the total energy flux received at the Earth (corrected for atmospheric extinction) is known, the so-called brightness temperature can be found.

The effective temperature, T_{eff}, of a star is defined in terms of its total energy output and radius. Thus, since σT^4_{eff} is the rate of radiation per unit area for a perfectly radiating sphere and if L is the total radiation of a star considered to be a sphere of radius R, such a sphere (called a blackbody) would emit a total amount of energy that is the surface area, $4\pi R^2$, times the energy per unit area just given. In symbols, $L = 4\pi R^2\sigma T^4_{eff}$. This relation defines the star's equivalent black-sphere or effective temperature.

Since the total energy radiated by a star cannot be directly observed (except for the Sun), the effective temperature is a derived quantity, not an observed one. Yet, theoretically, it is the fundamental temperature. If the bolometric corrections (see above) are known, the effective temperature can be found for any star whose absolute visual magnitude and radius are known.

AVERAGE VALUES OF TEMPERATURES, RADII, MASSES, AND LUMINOSITIES FOR DWARF STARS

Table 3 gives surface temperatures, bolometric or total luminosities, masses, and radii in terms of corresponding solar values for main-sequence stars ranging from very luminous objects to faint M dwarf stars. There is a considerable range for each of these quantities. For a given mass (M, the fundamental factor), a fair spread in radius (R), luminosity (L), surface temperature, and spectral class may exist. This spread is produced by stellar evolution effects and causes a broadening of the dwarf or main sequence. Masses are obtained from visual and eclipsing binary systems observed spectroscopically. Radii are found from eclipsing binary systems, from direct measurements in a few favourable cases, by calculations, and from absolute visual magnitudes and temperatures.

Meaningful average values of R, L, and M can be given only for dwarf stars. Giant and subgiant stars all show large ranges in R for a given mass. Conversely, giant stars of very nearly the same radius, surface temperature, and luminosity may have appreciably different masses.

Stars more massive and luminous than Y Cygni (see above) exist, but their diameters and true luminosities are not accurately known. Stars also exist with masses less than 0.2 of the Sun's mass, although, if M is less than 0.07 of the Sun's mass, the object cannot shine by nuclear energy at all and will simply cool down directly as a black dwarf or even a planet.

Minimum stellar mass

III. Stellar variability

Many of the stars in the sky are variable. Some are geometric variables, as in eclipsing binary systems; others are intrinsic variables.

GEOMETRIC VARIABLES

These are stars from which the amount of light reaching the observer alters because of changes in the stars' posi-

Table 3: Temperatures, Luminosities, Masses, and Radii for Main-Sequence Stars

spectral class	temperature (°K)	luminosity	mass	radius
O9.5	33,000	30,000.000	18.00	5.90
B0	32,000	16,000.000	16.00	5.70
B2	22,000	8,300.000	10.50	5.10
B3	17,000	2,500.000	8.30	4.80
B5	15,000	750.000	5.40	3.70
B8	12,500	130.000	3.50	2.70
B9	10,800	100.000	2.90	2.50
A0	9,500	63.000	2.60	2.30
A2	9,000	40.000	2.20	2.00
A5	8,700	24.000	1.90	1.80
A7	8,100	11.000	1.80	1.70
F0	7,400	9.000	1.60	1.50
F2	7,100	6.300	1.50	1.30
F5	6,400	4.000	1.35	1.20
F8	6,100	2.500	1.20	1.10
G0	5,900	1.450	1.08	1.05
G2	5,800	1.100	1.00	1.00
G5	5,600	0.700	0.95	0.91
G8	5,300	0.440	0.85	0.87
K0	5,100	0.360	0.83	0.83
K2	4,830	0.280	0.78	0.79
K5	4,370	0.180	0.68	0.74
K8	3,900	0.120	0.58	0.67
M0	3,670	0.075	0.47	0.63
M1	3,500	0.050	0.40	0.56
M2	3,400	0.030	0.33	0.36
M3	3,300	0.014	0.26	0.29
M4	3,200	0.005	0.20	0.21

Source: Compiled from data given by D.M. Popper, D.L. Harris, K. Strand, and C.E. Worley. The temperature scale is based on work of Hanbury Brown and his associates, G.J. Peters and E.K.L. Upton, S. Wolff, L.V. Kuhi, and D. Hayes, of D.C. Morton and T.A. Adams, of D.M. Popper, J.E. Stebbins and A.E. Whitford, A. Code, and J.B. Oke.

tions. Most geometric variables are eclipsing binaries (see also ECLIPSE, OCCULTATION, AND TRANSIT). An eclipsing binary consists of two close stars moving in an orbit so placed in space that the light of one may at times be hidden behind the other. Depending on the orientation of the orbit and the sizes of the stars, the eclipses may be total or annular (a ring of one star shows behind the other at the climax of the eclipse), or both eclipses may be partial. The best known example of an eclipsing binary is the famous star Beta Persei, or Algol, which has a period (interval between eclipses) of 2.87 days. The brighter (B8) star contributes about 92 percent of the light of the system and the eclipsed star much less than 8 percent; the system contains a third star, which is not eclipsed. About 17 eclipsing binary systems are visible to the naked eye.

Light curves of binaries

The light curve for an eclipsing binary displays magnitude measurements for the system over a complete light cycle. The light of the variable star is usually compared with that of a nearby (comparison) star believed to be fixed in brightness. In Figure 3A, the deep, or primary, minimum is produced when the star of higher surface brightness is eclipsed. It represents the total eclipse and is characterized by a flat bottom. The shallower secondary eclipse occurs when the star of higher surface brightness passes in front of the other; it corresponds to an annular eclipse (or transit). In a partial eclipse (Figure 3B) neither star is ever completely hidden, and the light changes continuously during an eclipse.

The shape of the light-curve during an eclipse gives the ratio of the radii of the two stars, and also one radius in terms of the size of the orbit, the ratio of the luminosities, and the inclination of the orbital plane to the plane of the sky.

If radial-velocity curves are available—*i.e.*, if the binary is spectroscopic as well as eclipsing—additional information is found. When both velocity curves are observable, the size of the orbit and the sizes, masses, and densities of the stars can be found. If the parallax of the system can be found, the brightness temperatures of the individual stars may be estimated from their luminosities and diameters. All of these procedures have been carried out for the faint binary Castor C and for the bright B star Mu Scorpii.

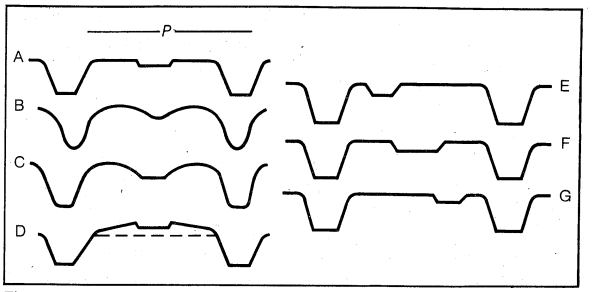


Figure 3: Schematic light curves of eclipsing binaries. *P* is the period (see text).

Light curves of the types shown in Figures 3A and 3B usually occur in systems in which the companions are moderately distant from one another. In a close system, tidal distortion caused by the strong mutual gravitation may cause the stars to become elongated along the line joining their centres so that the light curve is of the type shown in Figure 3C.

Close stars may reflect each others' light noticeably. If a small, high-temperature star is paired with a larger object of low surface brightness and if the distance between the stars is small, the part of the cool star facing the hotter one will be substantially brightened by it. Just before (and just after) secondary eclipse, this illuminated hemisphere is pointed toward the observer, and the total light of the system is a maximum (see Figure 3D).

Light curves need to be corrected for these effects before they can be further analyzed. Another complication is that stars such as the Sun show a darkening to the limb, or edge, of their disk.

One other effect on the light curves must be mentioned. Some stars move around each other in strongly elliptical orbits. If the major axis of the orbit is at right angles to the direction toward the observer, the stars will take different lengths of time to go from primary minimum to secondary minimum and from secondary minimum to primary minimum, because they move faster when close together than when far apart. Hence, the secondary minimum will not fall precisely at the midpoint between primary eclipses (Figure 3E). If the orbit precesses and the system is observed at a later date, the secondary minimum may then be symetrically placed, but now the two eclipses are of unequal duration (Figure 3F); at this time the major axis is pointed toward the observer. One eclipse now occurs when the occulting star is moving most rapidly, the other when it is moving most slowly. At a still later date the secondary minimum is displaced to the opposite side from the midpoint (Figure 3G). Observation of these effects in an eclipsing system shows the rate of rotation of the orbit. A third body might cause this rotation and possibly sometimes does but would produce other effects that could be seen but are not observed; in most instances it must be concluded that only two stars are present. The orbit rotation can then be explained as follows.

Tidal effects

If the stars are distorted by their mutual tidal attractions, the elliptical orbit will slowly rotate; the speed of rotation will depend in a precisely known fashion on their separation, masses, radii, and degree of density increase toward the centre. The orbit would turn at a maximum possible speed for stars that were homogeneous; *i.e.*, of equal density throughout. It would not rotate at all for stars whose masses were essentially all concentrated to the centre (approximately, very small stars with very deep but tenuous atmospheres). All theories of stellar structure predict large mass concentrations toward the centre; for the Sun the central density is about 100 times the mean density. These predictions can be tested by the behaviour of eclipsing binaries; in all instances the central condensations are large, providing a useful check on the theory.

The data from eclipsing binary systems are not necessarily applicable to isolated single stars. Systems in which a smaller, hotter star is accompanied by a larger, cooler object are easier to detect than are systems that contain, for example, two main-sequence stars. In such an unequal

system, at least the cooler star has certainly been affected by evolutionary changes and probably also the brighter one. The evolutionary development of two stars near one another does not exactly parallel that of two well-separated or isolated stars.

Eclipsing binaries include combinations of a variety of stars ranging from white dwarfs to huge supergiants (such as VV Cephei), which would engulf Jupiter and all the inner planets if placed at the position of the Sun.

Some members of eclipsing binary systems are intrinsic variables. In many such systems large clouds of ionized gas swirl between the components. In others, such as Castor C, the faint M dwarf component may be a flare star, one in which the brightness may unpredictably and suddenly increase to many times its normal value (see below).

INTRINSIC VARIABLES

A fair number of stars are intrinsically variable; that is, their total energy output fluctuates with time. Some of these stars were found by accident, but many were detected as a result of carefully planned searches. Variable stars are important in astronomy for a number of reasons. They usually appear to be stars at critical or short-lived phases of their evolution; detailed studies of their light and spectral characteristics, spatial distribution, and association with other types of stars may provide important clues to the life histories of various classes of stars. Certain kinds of variable stars, such as Cepheids (periodic variables) and novae (explosive variables), are extremely important in that they make it possible to establish the distances of remote systems beyond the Galaxy, such as the Magellanic Clouds and some of the nearer external galaxies (see GALAXIES, EXTERNAL). If the intrinsic luminosity of a recognizable variable is known and this kind of variable star can be found in a distant stellar system, the distance of the latter can be estimated from a measurement of the apparent magnitude, m, and the absolute magnitude, M, if the space absorption is known (see above and INTERSTELLAR MEDIUM). Historically, the absolute magnitude turned out to be difficult to establish for the Cepheid variable stars, a very important class of objects named after the prototype, Delta Cephei.

Kinds of intrinsic variables

Variable stars are often classified as behaving like a prototype star, and the class is then named for this star; *e.g.*, RR Lyrae stars are those whose variability follows the pattern of the star RR Lyrae. The most important classes of intrinsically variable stars are the following:

1. Pulsational variables; *i.e.*, stars whose variations in light and colour are believed to arise primarily from stellar pulsations. These include Cepheids, RR Lyrae stars, and Delta Scuti stars, all with short regular periods up to a few days; long-period variables; semiregular variables; irregular red variables; and Beta Canis Majoris stars.
2. Explosive, or catastrophic, variables, in which the variations are produced by the wrenching away of part of the star, usually the outer layers, in some explosive process. They include SS Cygni or U Geminorum stars; novae; and supernovae (which are usually regarded as representing an enormous explosion involving most of the material in a star).
3. Miscellaneous and special types of variables: R Coronae Borealis stars; T Tauri stars; flare stars; pulsars (neutron stars); spectrum and magnetic variables; X-ray variable stars; radio variable stars.

T Tauri and flare stars, spectrum and magnetic variables, X-ray stars, radio stars, and pulsars are considered in a later section.

R Coronae Borealis variables are giant stars of about the Sun's temperature whose atmospheres are characterized by excessive quantities of carbon. Their brightness remains constant until suddenly the star dims by several magnitudes and then slowly recovers its original brightness; the star's colour remains the same during the changes in brightness. The dimmings occur in a random fashion and seem to be due to the huge concentrations of carbon. At times the carbon vapour literally condenses into soot, and the star is hidden until the smog blanket is evaporated. Similar veiling may sometimes occur in other types of low-temperature star, particularly the long-period variables.

Pulsating stars. An impressive body of evidence indicates that stellar pulsations can account for the variability of Cepheids, long-period variables, semiregular variables, Beta Canis Majoris stars (pulsating stars with periods of only a few hours; see below), and even the irregular red variables. Of this group the Cepheids have been studied in greatest detail, both theoretically and by observation. These stars are regular in their behaviour; some repeat their light curves with great faithfulness from one cycle to the next over periods of many years.

Cepheid variables. Much confusion existed in the study of Cepheids until it was recognized that different types of Cepheids were associated with different groups, or population types, of stars. Stars fall into distinct groups as defined by motions, age, and position in the Galaxy. Population I is the name given to the stars associated with the spiral arms of this and other galaxies. This group contains stars of all ages, from those in the process of formation to defunct white dwarfs. They are always associated with the dust and gas of the interstellar medium. The best examples of type I populations are the stars of the galactic, or open, clusters such as the Pleiades or the beautiful double cluster in Perseus (h and χ Persei). Population I is sometimes divided into an "old" Population I (which includes the Sun) and a "young" Population I (stars a few tens of millions of years old or less).

Population types I and II

Population II is often divided into two categories. The oldest stars (extreme type II) are those associated with the central bulge of the Galaxy, globular clusters, the galactic halo, and certain external elliptical galaxies. A second group, called the disk population, comprises stars found between the spiral arms of the Galaxy and at some distance above and below the galactic plane. These stars are also very old, perhaps 10^{10} years, although not quite as old as the extreme Population II. In this article the populations are called simply types I and II.

Separate species of variable stars with different characteristics correspond to these population types. Cepheids belonging to the spiral-arm population (type I) are characterized by regularity in their behaviour. They show continuous velocity curves indicative of a regular pulsation. They exhibit a relation between period and luminosity in the sense that the longer the period of the star, the greater its intrinsic brightness; this period–luminosity relationship has been used to establish the distance of remote stellar systems. Cepheids with different properties are found in the type II population, away from the Milky Way, in globular clusters. They are bluer than classical type I Cepheids of the same period, and their light curves have different shapes. Studies of the light and velocity curves indicate that shells of gas are ejected from the stars as discontinuous layers that later fall back toward the surface. These stars exhibit no well-defined relation between period and luminosity; consequently, they are worthless as distance indicators. The distance of a Cepheid in a distant stellar system can be determined only if it is known to be a classical Cepheid.

Type II Cepheids

RR Lyrae variables. Closely associated with type II Cepheids are the cluster-type, or RR Lyrae, variables. Many of these stars are found in clusters, but some, such as the prototype RR Lyrae, are found far from any cluster or the central galactic bulge. The periods are less than a day, and there is no correlation between period and luminosity. Their absolute magnitudes are about 0.6; that is, they are about 50 times as bright as the Sun. They are useful for determining the distance of star clusters and some of the nearer external galaxies, since their short periods permit them to be detected readily.

Pulsation theory. An impressive amount of evidence supports the pulsation theory, first proposed as a possibility for stellar variation in 1879, applied to Cepheids in 1914, and further developed by the English astronomer Sir Arthur Eddington. He found that if stars have roughly the same kind of internal structure, then the period, P, multiplied by the square root of the density, ρ, is equal to a constant that depends on the internal structure.

The Eddington theory, although a good approximation, encountered some severe difficulties that have been met through modifications made in the 1960s. If the entire star pulsated in synchronism, it should be brightest when compressed and smallest and faintest when expanded and largest. The velocity curve should pass through zero at both maximum and minimum light. Observations contradict these predictions. When the star pulsates, all parts of the main body move in synchronism, but the outer observable strata get out of step or lag behind the pulsation of the inner regions. Running waves are set up within the outer regions (envelope) of the star. Pulsations involve only the outer part of a star; the core, where energy is generated by thermonuclear reactions, is unaffected.

Many years ago, careful measurements of the average magnitudes and colours of RR Lyrae stars in the globular cluster M3 showed that all these stars fell within a narrow range of luminosity and colour (or surface temperature; see Figure 4) or, what amounts to the same thing, luminosity and radius. Also, every star falling in this narrow range of brightness and size was an RR Lyrae variable. Later work has indicated that similar considerations apply to classical Cepheids. Variability is thus a characteristic of any star whose evolution carries it to a certain size and luminosity.

From *Astrophysical Journal*, reproduced by permission of the University of Chicago Press

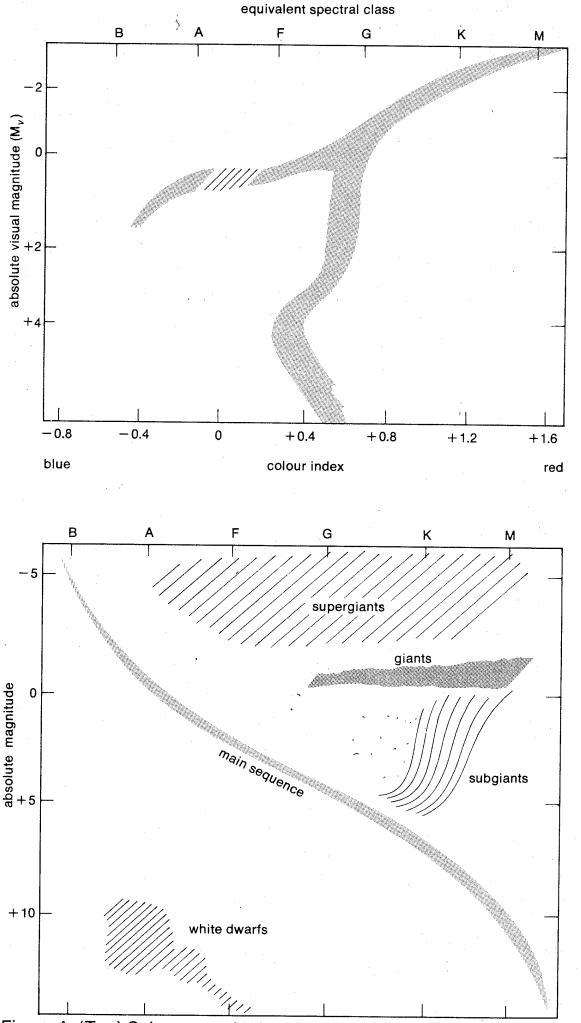


Figure 4: (Top) Colour–magnitude array for a typical type II population in a globular cluster. Absolute magnitude is plotted against the colour index (visual magnitude minus photographic magnitude). It is assumed that RR Lyrae stars (crosshatched area) have an absolute magnitude of 0.6. (Bottom) Schematic spectrum–luminosity correlation (Hertzsprung–Russell diagram) of spiral-arm stars (neighbourhood of the Sun).

In the pulsation theory as now developed, the light and velocity curves of Cepheids can be interpreted not only qualitatively but also quantitatively. For example, the complicated light curves of Population II Cepheids have been precisely predicted by the theory. Stellar pulsation, like other rhythmical actions, may give rise to harmonic phenomena wherein beats reinforce or interfere with one another. Beat and interference phenomena then complicate the light and velocity curves. The RR Lyrae stars supply some of the best examples; but semiregular variables such as the RV Tauri stars, "dwarf" Cepheids such as Delta Scuti, or SX Phoenicis evidently vibrate simultaneously with two or more periods.

Beta Canis Majoris variables. The Beta Canis Majoris variables are high-temperature stars (spectral class B) that often show complicated variations in spectral line shapes and intensities, velocity curves, and light. They often have two periods of variation so similar in duration that complicated interference or beat phenomena are observed, both in radial velocities and in the shapes of spectral lines.

Long-period variables. Long-period variables also probably owe their variations to pulsations. Here the situation is complicated by the vast extent of their atmospheres, so that radiations originating at very different depths in the star are observed at the same time. At certain phases of the variations, bright hydrogen lines are observed, overlain with titanium oxide absorption. The explanation is an outward-moving layer of hot, recombining gas, whose radiation is absorbed by strata of cool gases. These stars are all cool red giant and supergiant stars of spectral classes M (normal composition), R and N (carbon-rich), or S (heavy-metal rich). The range in visual brightness during a pulsation may be a hundredfold, but the range in total energy output is much less, because at very low stellar temperatures (1,500°–3,000° K) most of the energy is radiated in the infrared as heat rather than as light.

Unlike the light curves of classical Cepheids, the light curves of these red variables show considerable variations from one cycle to another. The visual magnitude, M_v, of Mira Ceti (Omicron Ceti) is normally around 9 to 9.5 at minimum light, but at maximum it may lie between 5 and 2. Time intervals between maxima often vary considerably. In such cool objects a very small change in temperature can produce a huge change in the output of visible radiation. At the low temperatures of the red variables, compounds and probably solid particles are formed copiously, so that the visible light may be profoundly affected by a slight change in physical conditions. Random fluctuations from cycle to cycle, which would produce negligible effects in a hotter star, will produce marked light changes in a long-period variable.

Types of long-period variables

Long-period variables appear to fall into two groups: (1) those with periods around 200 days tend to be associated with the type II population; (2) those of periods around a year belong to the type I population.

Red semiregular variables such as the RV Tauri stars show complicated light and spectrum changes. They do not repeat themselves from one cycle to the next; their behaviour suggests a simultaneous operation of two or more modes of oscillation. Betelgeuse is an example of an irregular red variable. In these stars the free period of oscillation does not coincide with the periodicity of the driving mechanism.

Explosive variables. *Novae.* Novae, or temporary stars, are explosive stars. In ordinary novae the explosion seems to involve only the outer layers, as the star later returns to its former brightness; in supernovae the explosion is catastrophic. Normally, novae are small blue stars much fainter than the Sun, although very much hotter. When an outburst occurs, the star may brighten very suddenly, ten magnitudes or even more in a few hours. Thereafter it fades; the rate of fading is connected with the brightness of the nova. The brightest novae, which reach absolute magnitudes of about −8, fade most rapidly, whereas a typical slow nova, which reaches an absolute magnitude of −5, may take ten times as long to decline in brightness. The changes in light are accompanied by pronounced spectroscopic changes that can be interpreted as arising from changes in an ejected shell that dissipates slowly in space. In its earliest phases the

Probable ejection of a shell of gas

expanding shell is opaque; as its area rapidly grows, with a surface temperature near 7,000° K, the nova brightens rapidly. Then, near maximum light, the shell becomes transparent, and its total brightness plummets rapidly, causing the nova to dim.

The mass of the shell is believed to be rather small, comparable with the mass of the Earth. Only the outer layers of the star seem to be affected; the main mass settles down after the outburst into a state much as before until a new outburst appears. The existence of repeating novae such as the star T Coronae Borealis suggests that perhaps all novae repeat at intervals ranging up to thousands or perhaps millions of years; and probably, the larger the explosion, the longer the interval. There is strong evidence that novae are components of close double stars and, in particular, that they have evolved from the most common kind of eclipsing binaries, those of the W Ursae Majoris type, which are pairs of stars apparently similar to the Sun in size but revolving about one another while almost touching.

Nova-like variables. Stars of the SS Cygni type show nova-like outbursts but with a much smaller amplitude. The intervals between outbursts are a few months to about a year. These variables are close binaries. The development of this particular type may be possible only in close binary systems.

Also associated with novae are stars such as Z Andromedae, which show remarkable spectra combining the characteristics of cool M-type stars, gaseous nebulae, and hot stars with bright emissions of hydrogen and helium. They occasionally show nova-like outbursts.

Eta Carinae

A peculiar object with no known counterpart is Eta Carinae, which appears in the telescope as a fuzzy red "star" slightly less than 2″ of arc in diameter. Surrounding it is a roughly elliptical nonuniform shell or halo about 10″ of arc in diameter. First observed as a star of about the fourth magnitude in 1677, it brightened irregularly, undergoing an outburst in 1843. Unlike a proper nova or supernova, it remained bright for several years, fading out of sight to the naked eye about 100 years ago. Since then it has shown irregular variations around the seventh magnitude. The first spectrum obtained of the object resembled that of a somewhat abnormal supergiant, but it now shows very complex emission-line structures dominated by the radiations of ionized iron. If a star was present, it may have been blown to bits. Eta Carinae was brighter than normal novae, yet it resembles no known supernova; it does not appear to be a source of X-rays or nonthermal radio-frequency emission.

Supernovae. Supernovae represent a distinct type of phenomenon: a star suddenly becomes up to several hundred million times as bright as the Sun, then fades away. Three have been observed optically in the Galaxy —Tycho's star of 1572, Kepler's "nova" of 1604, and the outburst in 1054 that produced the Crab Nebula. Hundreds have been found in other galaxies. Several remnants have been found by radio-astronomy methods.

The Crab Nebula itself is an emitter of nonthermal radiation. The residue of the supernova, now identified as a pulsar, believed to be a rotating neutron star, is a remarkable variable in optical, X-ray, and radio-frequency ranges (see below). Supernovae are believed to be stars that actually explode as a consequence of runaway thermonuclear reactions in their interiors (see below).

Probably all variable stars represent more or less ephemeral phases in the evolution of a star. Aside from such catastrophic events as those that produce a supernova, some phases of stellar variability may be of such brief duration as to permit recognizable changes during an interval of 50 or 100 years. Other stages may require many thousands of years; the period of Delta Cephei, for example, has not changed by a detectable amount since its variability was discovered.

IV. Statistics of stars

CORRELATIONS AMONG LUMINOSITY, SPECTRUM, MASS, AND RADIUS

Some of the most important generalizations concerning the nature and evolution of stars can be made from correlations between observable properties and from certain statistical results. One of the most important of these correlations concerns the spectrum and luminosity.

When the absolute magnitudes of the stars (or their intrinsic luminosities on a logarithmic scale) are plotted in a diagram against the temperature or, what is equivalent, the spectral types, the stars do not fall at random on the diagram but tend to congregate in certain restricted domains (see Figure 4, bottom). Such a diagram is usually called a Hertzsprung–Russell diagram, for a Danish astronomer, Ejnar Hertzsprung, and an American astronomer, Henry Norris Russell, who independently discovered the relations shown in it. Most of the congregated stars are dwarfs lying closely around a diagonal line called the main sequence. These stars range from hot blue objects 10,000 times (or more) brighter than the Sun, down through white A stars such as Sirius, yellow-white F stars such as Procyon, through the Sun, to orange K stars such as Epsilon Eridani, and finally to faint red dwarfs thousands of times fainter than the Sun. The sequence is continuous, the luminosities fall off smoothly with decreasing surface temperature, the masses and radii decrease but at a much slower rate, and the densities gradually increase (see Table 3).

Hertzsprung–Russell diagram

The second group of stars to be recognized was a group of giants, objects such as Capella, Arcturus, and Aldebaran, yellow, orange, or red stars about 100 times as bright as the Sun and with diameters of the order of 10,000,000 or 20,000,000 miles (20,000,000 or 30,000,000 kilometres). The class of supergiants includes stars of all spectral classes; they show a big spread in intrinsic brightness; some even approach absolute magnitudes —7 or —8. Some red supergiants, such as the variable VV Cephei, exceed in size the orbit of Jupiter or even that of Saturn, although most of them are smaller. Supergiants are short-lived and rare in space, though their great luminosity enables them to be seen at great distances.

Giant stars

Subgiants are stars that are redder and larger than main-sequence stars of the same luminosity. Many of the best known examples are found in close binary systems, particularly in eclipsing systems where conditions favour their detection.

The white-dwarf sequence lies about ten magnitudes below the main sequence. These stars are in the last stages of their evolution (see below *Stellar evolution*).

The spectrum–luminosity diagram has numerous gaps. There are very few stars above the white dwarfs and to the left of the main sequence. The main sequence and the giants are separated by a gap named after Hertzsprung, who in 1911 was the first to recognize the difference between main-sequence and giant stars. The actual concentration of stars differs considerably in different parts of the diagram. Highly luminous stars are rare; those of low luminosity are very numerous.

Figure 4 (bottom) applies to the spiral arm in the neighbourhood of the Sun and represents what would be obtained if a composite Hertzsprung–Russell diagram were constructed combining data for a large number of the star groups called open clusters, such as the double cluster h and χ Persei, the Pleiades, the Coma cluster, and the Hyades. It includes very young stars, a few million years old, as well as very ancient stars, 8,000,000,000 or 9,000,000,000 years old.

Figure 4 (top) exhibits the type of spectrum–luminosity relation characteristic of globular clusters, the central bulge of the Galaxy, and elliptical external galaxies; that is, of stellar Population II. Since these systems are very remote from the observer, the stars are faint, and their spectra can be observed only with difficulty. Hence, it is necessary to measure their colours rather than the spectra. Since the colours are closely related to surface temperatures and therefore to spectral classes, equivalent spectral classes may be used; but it is to be emphasized that stellar colours, not spectral classes, are observed in this instance.

The difference between the two diagrams of Figure 4 is striking. In the top diagram there are no supergiants, and, instead of falling at $M_v = 0.0$, the giant branch starts around —3.5 for very red stars and forms a continuous sequence until an absolute magnitude of about 0 is

reached. Then there is a band of stars, all about the same colour, proceeding downward (*i.e.*, to fainter stars) to about $M_v = +3$. The giant branch is connected to the main sequence at about +4 by a narrow band. The main sequence extends to fainter redder stars in much the same way as in the spiral-arm population. It ends just above the Sun, however, and does not extend up through the A, B, and O spectral classes, though occasionally a few such stars are found in the region normally occupied by the main sequence.

The horizontal branch falls near magnitude +0.6 and fills the Hertzsprung gap, extending to increasingly blue stars beyond the RR Lyrae stars, which are indicated by the crosshatched area. Among these hot blue stars are found novae and the nuclei of planetary nebulae, so called because their photographic image is disk-shaped and resembles that of a distant planet. Not all globular clusters show identical colour–magnitude arrays, and the differences may be due to differences in the cluster ages. Some clusters have many RR Lyrae stars; others have very few. Some have numerous giants; others have numerous blue stars on the horizontal branch. In the cluster M92 the whole giant branch is bluer than that in the cluster M13.

Globular-cluster stars

Spectroscopic observations indicate that stars in globular clusters do not all have the same chemical compositions; in particular, the same metal–hydrogen ratios differ from cluster to cluster. In some globular clusters, such as 47 Tucanae, the metal–hydrogen ratio seems to be about the same as in the Sun. In Omega Centauri it is substantially less, whereas in M92 the ratio may be of the order of 1 percent of the solar value. Globular clusters may look very similar to one another and yet have substantially different metal–hydrogen ratios.

Not all metal-deficient stars are confined to globular clusters. Some so-called subdwarfs with metal–hydrogen ratios of less than 1 to 10 percent of the solar value are found in the neighbourhood of the Sun, as are also less extreme examples of Population II. High-velocity stars are of type II, but not all type II stars display a high velocity. Numerically, the ancient type II population and the so-called disk population probably greatly exceed the type I population in the Galaxy.

Relation between the number of stars and their luminosities. Of great statistical interest is the relationship between the luminosities of the stars and their frequency of occurrence. The naked-eye stars are nearly all intrinsically brighter than the Sun, but, for the known stars within 20 light-years of the Sun, the opposite is true. The bright stars are easily seen at great distances; the faint ones can be detected only if they are close. Only if stars of magnitude +11 were 1,000,000,000 times more abundant than stars of magnitude −4 could they be seen in equal numbers, provided all stars could be observed to some fixed limit of apparent brightness. Table 4 gives the relation between the numbers of stars and their luminosities for stars in the neighbourhood of the Sun. For a sphere of radius ten parsecs, about seven stars between the fourth and fifth magnitudes, 30 stars between the tenth and 11th, and 67 stars between the 14th and 15th magnitude could be expected in such a volume. There would be less than one chance in a thousand of finding a star brighter than magnitude −5. The luminosity function has a maximum near magnitude 15.5.

Table 4: Relation Between Numbers of Stars and Their Luminosities

absolute magnitude	numbers of stars	absolute magnitude	numbers of stars	absolute magnitude	numbers of stars
−0.5	0.05	+6.5	12.9	13.5	54
+0.5	0.52	7.5	16.1	14.5	67
+1.5	1.30	8.5	20.0	15.5	80
+2.5	3.20	9.5	24.5	16.5	60
+3.5	4.80	10.5	30.0	17.5	37
+4.5	7.00	11.5	35.0	18.5	23
+5.5	9.50	12.5	42.0	19.5	13

Source: After W.J. Luyten, *Monthly Notices*, Royal Astronomical Society, vol. 139 (1968).

The luminosity function depends on population type. For a pure type II population, the luminosity function will differ substantially from that of a pure type I population. There will be a small peak near absolute magnitude +0.6 corresponding to the horizontal branch for type II, and no stars as bright as absolute magnitude −5. The luminosity function for pure type I is evaluated best from galactic star clusters, the stars in such a cluster being at about the same distance. The neighbourhood of the Sun includes examples of both Populations I and II.

Mass–luminosity correlation. A plot of mass against bolometric luminosity for visual binaries for which good parallaxes and masses are available show that, for stars with masses comparable to that of the Sun, the luminosity, L, varies as a power, $3 + \beta$, of the mass M. This relation can be expressed as $L = (M)^{3+\beta}$. The power differs for substantially fainter or for much brighter stars (see Table 3).

The mass–luminosity correlation applies only to unevolved main-sequence stars. It fails for giants and supergiants and for the subgiant (dimmer) components of eclipsing binaries, all of which have changed considerably during their lifetime. It does not apply to any of the stars in a globular cluster that do not lie on the main sequence, nor to the white dwarfs, which are abnormally faint for their masses.

The mass–luminosity correlation, predicted theoretically by Eddington, is a general relationship that holds for all stars having essentially the same internal density and temperature distributions; that is, for what are termed the same stellar models.

STATISTICS OF BINARY SYSTEMS

Near the Sun, most stars are members of binary systems, and some of the nearest single stars are suspected of having companions.

Although some binary components are separated by hundreds of astronomical units while others are contact binaries, binary systems are most frequently built on the same scale as the solar system. The average separation between components is of the order of ten astronomical units. The division in mass between two components of a binary seems to be nearly random. A mass ratio of about 1:20 could occur about 5 percent of the time, and under these circumstances a solar system might form.

The formation of double and multiple stars on the one hand and that of solar systems on the other seem to be different facets of the same process. Planets may be produced by a process like that depicted in the nebular hypothesis, except that most of the mass and angular momentum is swept out of the system. Only a small fraction of the original material is retained in the planets. Solar systems may be frequent in the Earth's Galaxy; perhaps 100,000,000 stars have bona fide planets (see SOLAR SYSTEM).

STATISTICS OF SPECIAL TYPES OF STARS

White dwarfs. White dwarfs comprise a group of stars falling about ten magnitudes below the main sequence. They are more numerous than stars of any group except those of the main sequence. The masses of white dwarfs are comparable to that of the Sun; their densities far exceed anything on Earth. The first to be recognized was the companion to Sirius, originally detected by its gravitational attraction upon the larger, brighter star and only later discovered visually as a faint object, about 10,000 times fainter than Sirius, or 500 times fainter than the Sun. Its mass is slightly less than that of the Sun. Its colour and spectrum correspond roughly to spectral class A, with a surface temperature of 8,000° to 10,000° K. Hence, the energy emission per unit area from the surface must be much greater than that of the Sun. Since it is so faint, the surface area and therefore the volume must be very small, and the average density is of the order of 100,000 times that of water. Under these conditions the electrons are nearly all stripped off the atoms; the resulting material, called a degenerate gas, has remarkable properties. It does not obey the ordinary gas laws and is an almost perfect conductor of heat.

The companion to Sirius

One white dwarf, known as BD+16°516, is paired to a much cooler K0V dwarf star in an eclipsing system. The two stars, whose centres are separated by 1,300,000 miles, revolve about each other in a period of 12.5 hours. The white-dwarf star produces pronounced excitation and heating effects in the K star's atmosphere. The white dwarf's mass is about 0.6 that of the Sun, but its diameter is only 10,000 miles; hence, the density is about 650,000 times that of water.

In the faint object BD+70°8247, a measured value of 1 to 3 percent circular polarization implies the existence of a magnetic field of about 10,000,000 gauss. For comparison, the Earth's field is of the order of one gauss.

A white dwarf represents the terminal stage of the life of an average star; more massive stars may evolve into neutron stars (see PULSARS). Just before the white-dwarf phase, the star may be like the central stars of planetary nebulae. On the other hand, evidence suggests that T Tauri stars are objects either recently formed or in the process of formation from the interstellar medium.

T Tauri stars. T Tauri variables are dwarf stars of spectral classes G and K. They invariably lie in dense clouds of absorbing material. Their changes in brightness are erratic; some, such as R Monocerotis, are associated with small, fan-shaped nebulae. Their spectra show bright lines and washed-out absorption lines, suggesting that they are rapidly spinning objects in clouds of diffuse material. On an H–R diagram they always fall above the normal main sequence.

Flare stars. Flare stars are cool dwarfs (spectral class M) that display flares apparently very much like, but much more intense than, those of the Sun, in fact sometimes so bright as to overwhelm the normal light of the star. Solar flares are associated with copious radio emission, and simultaneous opitical and radio events appear to have been found in UV Ceti, YZ Canis Minoris, and V371 Orionis.

Spectrum and magnetic variables. Spectrum and magnetic variables, mostly of spectral class A, show only small amplitudes of light variation but often pronounced spectroscopic changes. Their spectra typically show strong lines of metals such as manganese, titanium, iron, chromium, and rare earths or lanthanides, which vary periodically in intensity. These stars have strong magnetic fields, typically from a few hundred to a few thousand gauss; one star, HD 215441, has a field that can reach 30,000 gauss. Not all magnetic stars are known to be variable in light; the variables seem also to possess variable magnetic fields. The best interpretation is that these stars are rotating about an inclined axis. As with the Earth, the magnetic and rotation axes do not coincide. Different ions are concentrated in different areas; for example, chromium in one area and the lanthanides in another.

Metallic-line A stars. The so-called metallic line stars in class A, called Am stars, appear to be deficient in calcium, sometimes in other metals. They are all binaries and appear to form a distinct group.

The Sun as a radio emitter

Radio stars. The Sun is an emitter of radio waves, but, with present techniques, even in its most active phases the Sun could not be detected at a distance greater than that of the nearest star, about four light-years. Most discrete radio-frequency sources have turned out to be old supernovae, radio galaxies, quasars, etc.; but in recent years well-recognized radio stars also have been recorded.

These probably include flare stars, possibly red supergiants such as Alpha Orionis, the high-temperature dwarf companion to the red supergiant Antares, and the shells ejected from Nova Serpentis 1970 and Nova Delphini. The radio emission from the last-named objects is consistent with that expected from an expanding shell of ionized gas and fades away as the gas becomes attenuated. The central star of the Crab Nebula has been detected as a (radio) pulsar.

X-ray stars. Measurements from rockets, balloons, and satellites reveal distinct X-ray sources outside the solar system. The strongest galactic source, named Scorpius X-1, appears to be associated with a hot variable star resembling an old nova. Three radio sources have been observed near the X-ray star. One coincides with the X-ray star, which shows rapid irregular time variations, as do the optical and radio sources. The other two radio sources lie on opposite sides of the X-ray source and are constant in brightness. X-ray stars may be massive objects near the end of their evolution and may be closely related to pulsars. Other objects similar to Sco X-1 have been detected.

V. Stellar structure

STELLAR ATMOSPHERES

To interpret a star's spectrum quantitatively, some idea of the progression of temperature and pressure with depth in the star is necessary. Some general theoretical principles are outlined below.

First, the variation of temperature with depth in the atmosphere will depend on whether energy is transported to the surface of the star by convection currents or by radiation. For the Sun, at least, a discriminant is possible.

Darkening at the limb, or edge, of a star's disk

High-speed photographs of the Sun's disk show that the centre of the disk is brighter than the limb. The difference in brightness depends on the wavelength of the light used. It is large in the violet, is small in the red, and almost vanishes in the infrared. This limb darkening arises because the Sun gets hotter toward its core. At the centre of the disk, radiation is received from deeper and hotter layers (on the average) than from the limb, and the dependence of temperature on depth can be shown to correspond to the transport of energy by radiation, not by convection, at least in the outer layers of the Sun's atmosphere.

The degree of limb darkening in any star will depend on the effective temperature of the star and on the variation of temperature with depth. Limb darkening must be taken into account in the analysis of light curves of eclipsing binaries. Here, again, the results suggest transport of energy by radiation.

The layers of a normal star are assumed to be in mechanical or "hydrostatic" equilibrium. At each point the pressure is such as to support the weight of the overlying layers. This gives a relation between pressure and density at any given depth.

The chemical composition of the layers and the absorptivity, or opacity, of the material must be known. In the Sun the principal source of opacity is the negative hydrogen ion; such an ion is a hydrogen atom with one extra electron loosely bound to it. In the atmospheres of many stars the extra electrons break loose and recombine with absorption or reemission of energy in the form of light. At visible wavelengths the main contribution to the opacity comes from the photo-dissociation of this ion. In hotter stars, such as Sirius ($T = 10{,}000°$ K), atomic hydrogen is the main source of opacity, while in cool stars much of the outgoing energy may be absorbed by molecular bands of titanium oxide, water vapour, carbon monoxide, etc. Additional sources of opacity are absorption by helium atoms and electron scattering in hotter stars, absorption by hydrogen molecules and molecule ions, by certain abundant metals such as magnesium, and Rayleigh scattering (a type of scattering of radiation by particles, named for the British physicist Lord Rayleigh) in cool supergiant stars.

Dependence of temperature on depth

For stars other than the Sun, the dependence of temperature on depth cannot be directly found; and calculations must proceed by a process of successive approximations, during which the flux of energy is taken to be constant with depth. At large depths in the Sun and similar stars, convection sets in. In broad outline, this scheme or model seems to be correct, but there are certain limitations. The outer layers are assumed to be plane-parallel stratified layers. Photographs of the solar granulation (see SUN) belie this simple picture, and realistic models must allow for rising columns of heated gases in some areas and descent of cooler gases in others. Particularly when the model is to be used for calculating the anticipated line spectrum of the star, the motions of the radiating gases are important. Typical velocities are of the order of two kilometres per second in the Sun; in other stars they may be much larger.

The temperature, density, and pressure all increase steadily inward in the Sun's atmosphere. The Sun has no distinct solid surface, so the point from which the depth or height is measured is arbitrary. The temperature of the visible layers ranges from 4,700° to 6,200° K, the density from about 0.0000001 to 0.0000004 gram per cubic centimetre, and the gas pressure from 0.002 to 0.14 atmosphere. The visible layers of stars such as the Sun have very low densities and pressures compared with the Earth's atmosphere, although the temperature is much higher. The strata of the solar atmosphere are very opaque compared with the terrestrial atmosphere.

Calculations have been carried out for atmospheres of a variety of stars ranging from dwarfs to supergiants, from cool K stars to hot O stars. Their validity can be evaluated only one way, by examining how well they predict the observed features of a star's continuous and line spectrum, including the detailed shapes of spectral lines and so on. In view of the known complexities of stellar atmospheres, the results fit the observations remarkably well.

Expanding atmospheres

Severe deviations exist for stars with extended and expanding atmospheres. Material flowing outward from a star produces a stellar wind analogous to the solar wind —but often much more extensive and violent. In the spectrum of certain very hot Of stars, such as Zeta Puppis, strong, relatively narrow emission lines can be seen, but in the ultraviolet, observations from rockets show strong emission lines with strong absorption components on the shorter wavelength side. These absorption features are produced by rapidly outflowing atoms that absorb the radiation from the underlying stellar surface. The observed shifts in frequency correspond to ejection velocities of about 100 kilometres per second. Much gentler stellar winds are found in cool M-type supergiants.

Such line profiles are commonly referred to as the P Cygni type and are shown schematically in Figure 5. The strong emission line is produced mostly by material moving across the line of sight. Broad profiles, presumably produced by expanding shells of gas, are also produced in the atmospheres of Wolf–Rayet stars. These objects are of much higher temperature than P Cygni—typically 30,000° to 50,000° K. Their spectra consist of wide emission lines superposed on a strong blue underlying continuous spectrum. Usually they are found in binary systems of which one component is often a conventional hot O-type star.

By courtesy of Perkins Observatory (Ohio State and Ohio Wesleyan Universities), Delaware, Ohio

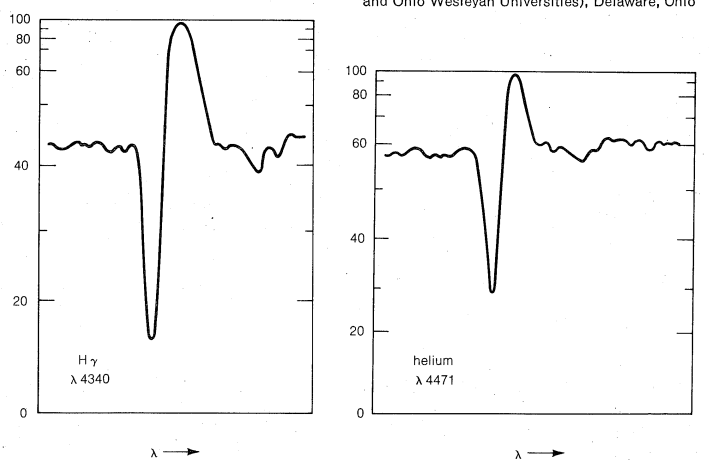


Figure 5: Line profiles in P Cygni. Figures indicate percentage intensities in terms of maximum of emission-line profile (August 24, 1941).

Effects of rapid rotation

Rapid stellar rotation may also modify the structure of a star's atmosphere. Since effective gravity is much reduced near the equator, the appropriate description of the atmosphere will vary with latitude. If the star is spinning at speeds near the breakup point, rings or shells may be shed from the equator.

Some of the most extreme and interesting examples are provided by close binary systems. When attempts are made to interpret the light and velocity curves of these objects, it is found that the spectroscopic observations cannot be reconciled with simple, orderly rotating stars. Instead, emission and absorption lines sometimes appear, jumbled together in such a way as to suggest streams of gas moving between the stars. Beta Lyrae is an eclipsing system with a period of 12.9 days but shows very large orbital-velocity shifts. The brighter component at visible wavelengths is a B9 star. The other component appears to be a hot, abnormal object; its spectral lines have not been observed. The spectrum of the B9 component shows the regular velocity changes expected of a binary star but with an absorption (and associated emission) spectrum corresponding to a higher temperature (near spectral class B5) and a blue continuum corresponding to a very high temperature star. The anomalous B5-type spectrum is evidently excited principally by the hotter source; it envelopes the entire system and shows few changes in velocity with time. Figure 6 shows Beta Lyrae schematically.

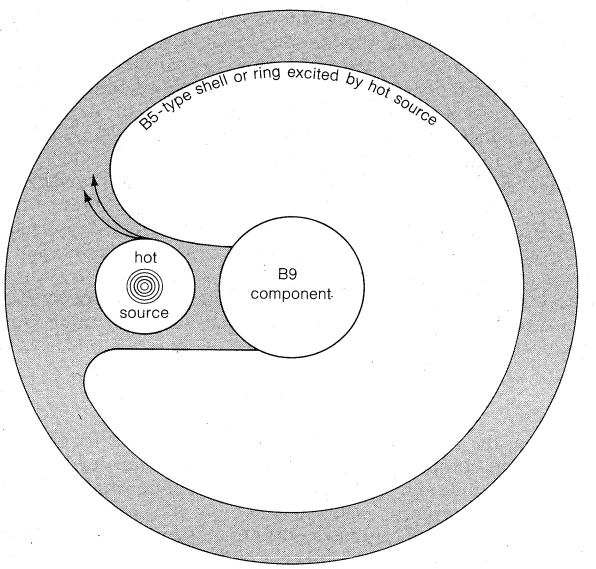


Figure 6: Probable behaviour of gases around "peculiar" double star Beta Lyrae. The nature of the "hot" component is not well understood.

Supergiant stars have very extended atmospheres, which are probably not even approximately in mechanical equilibrium. The atmospheres of M supergiant stars appear to be slowly expanding outward. Observations of the eclipsing binary 31 Cygni show that the K supergiant component has an extremely inhomogeneous, extended atmosphere composed of numerous blobs and filaments. As the B component of this system slowly moves behind the larger star, its light shines through larger masses of the K star's atmosphere. If the atmosphere were in orderly layers, the lines of ionized calcium, for example, produced by absorption of the light of the B star by the K-star atmosphere, would grow stronger uniformly as the eclipse proceeds—but they do not.

Flare-type and nova-type activity are other forms of instability that have been discussed above.

INTERNAL STRUCTURE OF STARS

Calculation of the internal structure of stars—that is, the dependence of pressure, temperature, and density on depth below the surface—depends on basic principles that are explained below. The calculations have to take into account the change of the star's structure with time as its hydrogen supply is gradually converted into helium.

Simplicity of relationships

Because of the basic simplicity of the relationship among temperature, density, and pressure, the internal pressures, temperatures, and densities of main-sequence stars and particularly of white dwarfs can be calculated with greater confidence than can the internal temperature and density of the Earth.

Relations obtained from physical laws. The star is assumed to be in mechanical equilibrium; that is, the pres-

sure at each point exactly balances the weight of the overlying layers. This balance can be expressed as a simple relation between the pressure gradient and the density. A second relation expresses the continuity of mass; that is, if M_r is the mass of material within a sphere of radius r, the mass, ΔM, added when one goes outward a distance Δr through a shell of volume $4\pi r^2\Delta r$, is equal to the volume of the shell multiplied by the density, ρ. In symbols, this is expressed as $\Delta M = 4\pi r^2\rho\Delta r$.

The third relation is an equation of state; it expresses the relation among the pressure, density, and temperature of the material. Throughout the star the material is entirely gaseous, and, except in certain highly evolved objects, it obeys closely the perfect-gas law. The molecular weight is 2 for molecular hydrogen, 4 for helium, and 56 for iron vapour. In the interior of a typical star, however, almost all of the material is completely ionized. Under these circumstances the hydrogen molecules are not only broken down into individual atoms, but the atoms themselves are broken down (ionized) into protons and electrons. Hence, the molecular weight of ionized hydrogen is the average mass of a proton and an electron, ½ on the atom-mass scale. On the other hand, a completely ionized helium atom contributes a mass of 4 with a helium nucleus (an alpha particle) plus two electrons of negligible mass; hence, the average molecular weight is ⅘. A completely ionized nickel atom contributes one nucleus of mass 58.7 plus 28 electrons; hence, its molecular weight is $58.7/29 = 2.02$. Since stars contain a preponderance of hydrogen and helium, which are completely ionized throughout the interior, the average particle mass, μ, is the (unit) mass of a proton, divided by a factor taking account of the concentrations by weight of hydrogen, X, and helium, Y, and α, which is nearly ½. In symbols the equation is as follows:

$$\mu = \frac{1}{2X + (\frac{3}{4})Y + \alpha(1 - X - Y)}$$

Hence, the molecular weight will depend critically on the chemical composition, particularly on the ratio of hydrogen to helium and the total content of heavier material.

If the temperature is sufficiently high, the radiation pressure, P_r, which is numerically equal to one-third of the energy density, must be taken into account as well as the perfect-gas pressure. Hence, the total equation of state is $P = P_g + P_r$. P_g depends on density, temperature, and the molecular weight, μ; P_r on the temperature T and a quantity a equal to 7.56×10^{-15} ergs cm^{-3} deg^{-4}; a is called the radiation density constant. Let $\mu = 2$ (upper limit), and $\rho = 1.4$ grams per cubic centimetre (mean density of the Sun), and the temperature at which the radiation pressure would equal the gas pressure can be calculated. This temperature is 28,000,000°, much hotter than the centre of the Sun, where the density is about 100 times that of water. Hence, radiation pressure in the Sun may be neglected, but it cannot be neglected in hotter, more massive stars. Radiation pressure may then set an upper limit to stellar luminosity.

Radiation density constant

Certain stars, notably the white dwarfs, do not obey the perfect-gas law. Instead, the pressure is almost entirely contributed by the electrons, which obey a law called the degenerate-gas law. If μ' is the average mass per free electron of the totally ionized gas, the pressure, P, and density, ρ, are such that P is proportional to a five-thirds power of the density divided by the average mass per free electron; that is, $P = 9.9 \times 10^{12}(\rho/\mu')^{5/3}$. The temperature does not enter at all. At yet higher densities the equation of state becomes more intricate, but it has been shown that even this complicated equation of state could be employed in the calculation of the internal structure of these white dwarf stars. The white dwarfs are probably better understood than most other objects.

For normal stars such as the Sun, the energy-transport method for the interior must be known. Except in white dwarfs or dense cores of evolved stars, thermal conduction is not significant, as the heat conductivity of gases is very low. One important mode of transport is an actual flow of radiation outward through the star. Starting as X-rays near the centre, the radiation is gradually "softened" as it works its way to the surface to emerge as ordinary light and heat. The rate of flow of radiation is proportional to the temperature gradient, the rate of change of temperature with distance. This relation uses the following important quantities: a, the radiation density constant mentioned above; the velocity of light, c; the density, ρ; and k, a measure of the opacity of the material. The larger the value of k, the smaller the transparency of the material and the steeper the temperature fall required to push the energy outward at the required rate. The opacity, k, can be calculated for any temperature, density, and chemical composition and is found to depend in a complicated way on the density and temperature.

In the Sun's outer layers and particularly in certain giant stars, transport of energy takes place by quite another mechanism—large-scale mass motions of gases, or convection. Large volumes of gas deep within the star become heated, rise to higher layers, and mix with their surroundings, thus releasing great quantities of energy. The extremely complicated flow patterns cannot be followed in detail, but when convection occurs a relatively simple relation exists between pressure and density. Wherever convection can occur, it is much more efficient than radiative energy transport.

Energy generation in stars. The most fundamental property of stars is that they radiate energy and must derive this energy from sources within themselves. For normal main-sequence stars such as the Sun, the source of the energy lies in the conversion of hydrogen into helium. The reaction believed to occur in the Sun is the proton–proton reaction. Two protons collide to form a deuteron (heavy hydrogen nucleus), with the liberation of a beta ray (positive electron, or positron). Also emitted is a neutral particle of negligible mass, called a neutrino; it escapes forthwith from the solar interior. The positive electron meets a negative electron, and the two annihilate one another with the liberation of much energy. The annihilation energy amounts to 1.02 megaelectron volts (MeV). This process is one of low probability, and many proton–proton collisions occur before a positive electron is created.

The proton–proton reaction

Next, a proton collides with the deuteron, ^{2}H, to form the nucleus of a light helium atom of atomic weight 3, ^{3}He. A "hard" X-ray or gamma ray is also emitted. The most likely event to follow in the chain is a collision of this ^{3}He nucleus with a normal ^{4}He nucleus or α-particle to form a nucleus of a beryllium atom of atom weight 7, ^{7}Be, with the emission of another gamma (γ) ray. The ^{7}Be nucleus in its turn captures a proton to form a boron nucleus of atomic weight 8, ^{8}B, with the emission of another gamma ray.

Boron-8, however, is very unstable. It decays to beryllium of atomic weight 8, ^{8}Be, with the emission of another positron, e^+, and a neutrino, ν; and then the nucleus itself decays into two α-particles (the helium nucleus, ^{4}He). These interactions can be represented by the following equations:

$$^3\text{He} + {}^4\text{He} \rightarrow {}^7\text{Be} + \gamma \qquad \text{(rather slow reaction)}$$
$$^7\text{Be} + {}^1\text{H} \rightarrow {}^8\text{B} + \gamma \qquad \text{(rapid reaction)}$$
$$\left.\begin{array}{l} {}^8\text{B} \rightarrow {}^8\text{Be} + (e^+) + \nu \\ {}^8\text{Be} \rightarrow 2\,{}^4\text{He} \end{array}\right\} \qquad \text{(instantaneous reactions)}$$

In the course of these reactions four protons have been consumed to form one helium nucleus, and two negative electrons have perished.

The mass of four hydrogen atoms is $4 \times 1.00797 = 4.03188$; that of a helium atom is 4.0026. Hence, 0.02928 atomic mass units, or 0.7 percent of the original mass, have disappeared. Some of this has disappeared in the form of the elusive neutrinos, but most of it has been converted into radiant energy. To keep shining at its present rate, a typical star like the Sun has to convert 674,000,000 tons of hydrogen into 670,000,000 tons of helium every second. More than 4,000,000 tons of matter literally disappear into radiation every second. The rate of energy generation according to the proton–proton reaction varies as about the fourth power of the temperature.

The theory gives a good picture of solar-energy generation except for one thing. Attempts have been made to observe the neutrino flux from the Sun by means of a very intricate and difficult experimental setup. To date, none have been found. No explanation for this discordance is known. Possibly the neutrino is an unstable particle.

The carbon cycle

The principal source of energy in hotter stars is the carbon cycle, in which hydrogen is converted into helium with carbon serving as a catalyst. The reactions are as follows: first, a carbon nucleus, ^{12}C, captures a hydrogen nucleus or proton, ^{1}H, to form a nucleus of nitrogen, ^{13}N; thus, $^{12}C + {}^{1}H \rightarrow {}^{13}N +$ gamma ray. A gamma ray is emitted in the process. The light ^{13}N nucleus is unstable, however. It emits a positive electron, e^+, which encounters a negative electron e^-, and the two annihilate one another. A neutrino is also emitted; the resulting ^{13}C nucleus is stable. Eventually the ^{13}C nucleus captures another proton, forms ^{14}N, and emits another gamma ray. In symbols the reaction is represented by the equations: $^{13}N \rightarrow {}^{13}C + e^+ + \nu$; $^{13}C + {}^{1}H \rightarrow {}^{14}N + \gamma$. The ^{14}N, ordinary nitrogen, is stable. When it captures a proton to form a nucleus of light oxygen-15, ^{15}O, the resulting nucleus is unstable against beta (β) decay; that is, it emits a positive electron, e^+, and a neutrino, ν. In symbols this is expressed as $^{14}N + {}^{1}H \rightarrow {}^{15}O + \gamma$; $^{15}O \rightarrow {}^{15}N + e^+ + \nu$. Again, the positive electron meets a negative electron and is annihilated while the neutrino escapes from the Sun. Eventually the ^{15}N nucleus encounters a fast-moving proton, ^{1}H, and captures it. It might be supposed that at this point a nucleus of ordinary ^{15}O would be formed. Actually, this process occurs rarely, the most likely reaction being $^{15}N + {}^{1}H \rightarrow {}^{12}C + {}^{4}He + \gamma$. The capture of the proton causes the resultant nucleus to break down into the original ^{12}C nucleus and an alpha particle, or helium nucleus, of mass 4. Thus, the original carbon nucleus reappears, and the four protons that have been added permit the formation of an alpha particle, or helium nucleus. The same amount of mass has disappeared, although a different fraction of it may have been carried off by the neutrinos.

In the interior of a star the particles are moving rapidly in every direction as a consequence of high temperatures. Once in a while a proton is sufficiently close to a nucleus for it to be captured and a nuclear reaction to occur. Only protons of very highest energy (many times the average energy in a star such as the Sun) are capable of producing such nuclear reactions. Since the energies of the protons are proportional to the temperature, the rate of energy production rises steeply as temperature increases. The temperature dependence is much steeper at the higher temperatures at which the carbon–nitrogen cycle takes place than it is for the proton–proton reaction.

The hottest stars of the main sequence shine with energy produced by the carbon cycle. The faint red dwarf stars use the proton–proton reaction exclusively, whereas stars such as the Sun shine mostly by the proton–proton reaction but have some contribution from the carbon cycle.

The aforementioned mathematical relationships, including the law of energy generation, permit the problem of stellar structure to be tackled. An early assumption that stars have a uniform chemical composition throughout their interiors simplified the calculations greatly; but it had to be abandoned when the importance of stellar evolution effects were taken into account, since evolution changes the composition. Calculations have to be carried out by a step-by-step process known as numerical integration and must take into account that at the surface the pressure and density vanish; at the centre of the star the pressure, density, and temperature must remain finite. A frequent procedure is to carry the computations from the surface inward and from the centre outward; then the two solutions must fit together, giving the same pressure, density, temperature, etc., when they meet. One of the most important generalizations that arise from such calculations is the mass–luminosity relation, discovered by Eddington. The mass–luminosity relation, which is, in its theoretical form, really a mass–radius–luminosity relation, expresses the fact that the luminosity, L, is proportional to a power of the mass and inversely proportional to a constant power of the radius, R. The molecular weight, μ, also appears in the formula, but this factor is almost constant. The luminosity depends on the mass but even more on the molecular weight and therefore on chemical composition. Thus: $L = C\frac{M^a\mu^b}{R^c}$, where a is of the order of 5, b about 7 or 8, and c near 1.5.

Theoretical derivation of a mass–luminosity relation

The mass–luminosity law does not explicitly contain the law of energy generation. The equations governing stellar structure require such a relation without invoking the detailed law of energy generation, which was not known in Eddington's time. Given the precise law of energy generation, an additional expression involving the chemical composition, luminosity, and radius may be derived. For a chemically homogeneous star shining on the carbon cycle, taking X and Y as the proportions by weight of hydrogen and helium, respectively, and with the powers d somewhere near 20, e about 22, and f about 23, since the carbon cycle depends on a high power of the temperature, the energy output relation can be written:

$$L = B(1 - X - Y)\mu^d \frac{M^e}{P^f}$$

It and the mass–luminosity relation are actually valid only for young stars.

Energy transport

The star model and the relation among mass, luminosity, and radius are determined by the mode of energy transport. In the Sun and the fainter main-sequence stars, energy is transported throughout the outer layers by convection currents, whereas, in the deep interior, energy is transported by radiation. Among the hotter stars of the main sequence, the reverse appears to be true. The deep interiors of stars that derive their energy primarily from the carbon cycle are in convective equilibrium, whereas in the outer parts the energy is carried by radiation. The observed masses, luminosities, and radii of most main-sequence stars can be reproduced with reasonable and uniform chemical compositions.

Chemically homogeneous possible models of giant and supergiant stars cannot be constructed. If a yellow giant star such as Capella is assumed to be built like main-sequence stars, the central temperature will turn out to be so low that no known nuclear process can possibly supply the observed energy output. Progress has been made only by assuming that these stars were once main-sequence objects that, in the course of their development, exhausted the hydrogen in their deep interiors. As a consequence, inert cores formed, composed mostly of helium, which is the ash of the hydrogen-burning processes. Since no nuclear helium reactions occur at the few tens of millions of degrees temperature and the densities likely to prevail in these interiors, no thermonuclear energy could be obtained from these depleted cores. Energy is assumed to be generated in a thin shell surrounding the inert core where some fuel remains and is presumably produced by the carbon–nitrogen cycle. Such models are called shell-source models. As the star uses up more and more of its hydrogen supply, the core increases in mass; but the outer envelope of the star continues to expand. These shell-source models can explain the observed luminosities, masses, and radii of giant and supergiant stars (see below).

Depletion of hydrogen in older stars

The depletion of hydrogen fuel is appreciable even for a dwarf star as old as the Sun (about 5,000,000,000 years); the Sun seems to have been shining at its present rate for about 1,000,000,000 years. For its observed luminosity to be maintained, the Sun's central temperature must have increased noticeably since the solar system was formed, largely as a consequence of the depletion of the hydrogen in the interior with an accompanying increase in molecular weight and temperature. Over the past 5,000,000,000 years the Sun may have brightened by about half a magnitude; and in the early Precambrian era (about 2,000,000,000 years ago) the solar luminosity was about 20 percent less than now.

Table 5 gives results of one calculation of the internal structure of the Sun. The carbon cycle and the proton–proton reaction and evolutionary effects are all taken into account. Density, temperature, and the fraction of

the total mass within a radius r (shown in Table 5 as a fraction of the Sun's radius R) are given. Nearly all the matter within the Sun is hotter than 1,000,000°. The central temperature attains about 15,000,000° K.

Table 5: Internal Structure of the Sun

r/R	mass within r/R fractional	density (water = 1)	temperature (°K)
0.0000	0.0000	149.8000	14,920,000
0.1026	0.0810	86.3000	12,460,000
0.2027	0.3510	35.0900	8,953,000
0.3022	0.5910	14.6300	6,731,000
0.4000	0.8110	3.9200	4,727,000
0.5048	0.9110	1.2030	3,557,000
0.5980	0.9560	0.4470	2,770,000
0.7020	0.9810	0.1550	2,095,000
0.8039	0.9930	0.0608	1,346,000
0.9000	0.9985	0.0197	630,000
0.9540	0.9997	0.0061	290,000
0.9760	0.9999	0.0025	162,000

Source: From J. and N. Bahcall and R.K. Ulrich.

VI. Stellar evolution

Evolutionary effects have not been negligible even for the Sun. A more massive star must show more spectacular effects, since the rate of conversion of mass into energy is higher. The Sun produces energy at the rate of about two ergs per gram per second; a more luminous main-sequence star may produce energy at the rate of 2,000 ergs per gram per second. Consequently, effects that require billions of years to be easily recognized in the Sun may occur within a few millions of years in highly luminous stars. A supergiant star such as Antares, a bright main-sequence star such as one of the belt stars of Orion, or even a more modest star such as Sirius cannot have endured as long as the Sun.

Rates of energy production

THE LIFE HISTORY OF A TYPICAL STAR

These stars must have been formed relatively recently. Stars might be found within the Galaxy (and even near the Sun itself) that are well evolved or even approaching extinction or both, as well as occasional stars that must be very young or even in the process of formation. Stars are presumably formed from the grains and gas in the interstellar medium. The material between the stars is not uniformly distributed in space but is spread in a patchy fashion. Occasionally, a sufficiently massive cloud will form so that gravitational attractions can draw it still further together. The details of this process are not at all well understood.

Protostars. As the cloud or protostar begins to pull itself together, the internal temperature and density start to rise until it reaches incandescence with a faint red glow. At this stage the star is still shining by the energy released by its gravitational contraction. As the internal temperatures rise to a few million degrees, deuterium (heavy hydrogen) is first destroyed. Then lithium, beryllium, and boron are broken down into helium as their nuclei are bombarded by protons moving at high speeds.

Detailed calculations show that a star originally falls above the main sequence in the Hertzsprung–Russell diagram because it is too bright for its colour. As it contracts, it moves downward and to the left toward the main sequence.

The star reaches the main sequence. As the central temperature and density continue to rise, the proton–proton or carbon-cycle reactions become active, and the development of the star is stabilized. The star then reaches the main sequence, where it remains for most of its active life. The time required for the contraction phase depends on the mass of the star. A star of the Sun's mass may require millions of years to reach the main sequence, whereas one of much greater mass may reach the main sequence in a few hundred thousand years.

By the time the star reaches the main sequence, it is still chemically homogeneous. As time goes on and the hydrogen fuel in the core is converted into helium, the temperature slowly rises. If the star is sufficiently massive to have a convective core, the material in this region has a chance to be thoroughly mixed; but the outer region does not mix with the core. In the Sun, on the other hand, there is no convective core; and the helium–hydrogen ratio is maximum at the centre and decreases outward. During the lifetime of the Sun there has been a steady depletion of hydrogen, so that at the centre the concentration of hydrogen is probably only about one-third of the original amount. The rest has all been converted to helium. The rate of evolution becomes the more rapid the greater the mass of the star. A star of twice the Sun's mass burns its fuel at such a rate that it lasts about 3,000,000,000 years, whereas a star of ten times the Sun's mass has a lifetime measured in tens of millions of years.

Effects of composition changes

As the composition of the interior changes, the star begins to move away from the main sequence, slowly at first and then more rapidly. When about 10 percent of the star's mass has been converted into helium, the structure of the star changes drastically. All the hydrogen of the core has been burned out, and this central region is composed almost entirely of inert helium, with small admixtures of heavier elements. The energy production now takes place in a thin shell where hydrogen is consumed and added to the inert core. The outer portion of the star expands outward, and as the star swells up it may gradually rise in luminosity. The details of the evolution depend on the hydrogen–metal ratio, and the course of evolution will differ for stars of different population types (see above).

The main sequence for a globular cluster (Figure 4) is discussed above. The colour–magnitude diagram exhibits the effects of stellar evolution; all main-sequence stars more massive than 1.5 solar masses have evolved off the main sequence, and most of them have died. The colour–magnitude array then represents an evolutionary picture of the cluster. The giant branch is not the path of evolution but rather the envelope of evolutionary paths. It represents a snapshot of stars of different masses but of the same age. Presumably, the stars leave the main sequence and evolve into stars of steadily increasing luminosity until they reach the peak of the diagram. Later stages of evolution are presumably connected with the horizontal branch at absolute magnitude 0, where the RR Lyrae stars and other blue objects are found. The last stages of evolution are represented by the white dwarfs, which are too faint to be detected in globular clusters, although they must be present there in great numbers.

Evolution away from the main sequence. Theoretical calculations suggest that, as the star evolves from the main sequence, the hydrogen–helium core gradually increases in mass but shrinks in size as more and more "ash" is fed in through the hydrogen-burning shell. Energy is carried outward from the shell by rapid convection currents. The temperature in the shell rises; the star becomes more and more luminous and finally approaches the top of the giant sequence. The core shrinks by gravitational contraction and becomes denser and hotter until it reaches a temperature of about 120,000,000° K. At that temperature the previously inert helium is consumed in the production of heavier elements.

Helium burning

If two helium (He) nuclei each of mass 4 are jammed together, it might be expected that they would form a nucleus of beryllium, Be, of mass 8; in symbols, the reaction would be: $^4He + {}^4He \rightarrow {}^8Be$. Actually, however, ^8Be is unstable and breaks down again into two helium nuclei. If the density and temperature are high enough, however, the short-lived beryllium nucleus may capture another helium nucleus in what is essentially a three-body collision to form a nucleus of carbon-12; thus $^8Be + {}^4He \rightarrow {}^{12}C$. Further reactions producing oxygen, ^{16}O, neon, ^{20}Ne, magnesium, ^{24}Mg, and so on, may occur. In symbols, $^{12}C + {}^4He \rightarrow {}^{16}O$; $^{16}O + {}^4He \rightarrow {}^{20}Ne$; and $^{20}Ne + {}^4He \rightarrow {}^{24}Mg$. Energy is released, and the star again has a central energy source; but the details of its evolution are uncertain. Since the energy output of helium burning (per ton of material) is much less than for hydrogen burning, the star's lifetime in the helium-burning phase cannot be long. The star eventually evolves on the horizontal branch to bluer colours and lower luminosities but certainly ends as a white dwarf. The energy

output of the white dwarfs is so low that they may go on shining mainly by radiating away their store of energy until none is left to emit. The final stage of a star is typically a ball not much larger than the Earth but with a density perhaps 50,000 times that of water, covered possibly with a thin layer of ice and surrounded by an atmosphere a few feet thick.

Estimates of stellar ages. The shapes of the colour–magnitude diagrams permit an estimate of globular-cluster ages. Stars more massive than about 1.3 solar masses have evolved away from the main sequence at a point just above the position occupied by the Sun. The time required for such a star to exhaust the hydrogen in its core is about 5,000,000,000 or 6,000,000,000 years; and the cluster must be at least as old. More ancient clusters have been identified. In the Galaxy, globular clusters are all very ancient objects, with ages measured in billions of years. In the Magellanic Clouds, globular clusters exist, such as NGC 1866, that contain numerous blue stars and therefore must be relatively young. Exact ages, however, cannot yet be assigned to globular clusters. The details of the evolutionary tracks depend on the hydrogen–metal ratios, helium–hydrogen ratios, and the precise theory of stellar evolution.

Ages of cluster stars

Open (galactic) clusters in the spiral arms of the Galaxy, of extreme Population I, tell a somewhat different story. Figure 7 shows colour–magnitude diagrams for several

From *Astrophysical Journal*, reproduced by permission of the University of Chicago Press

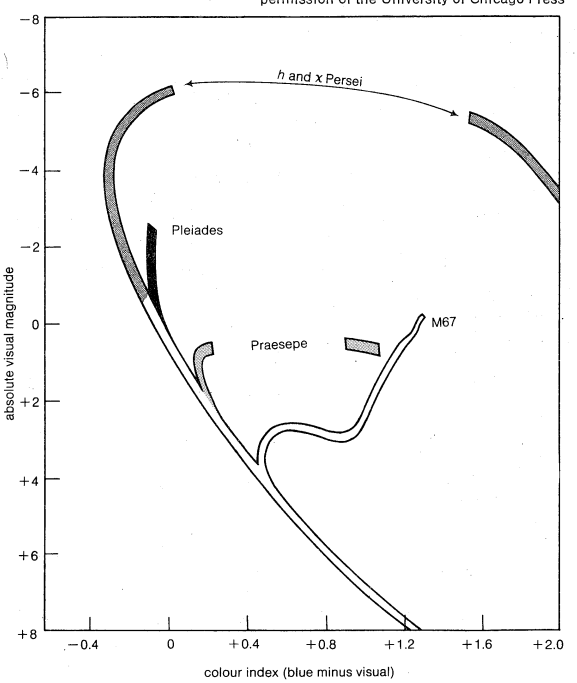


Figure 7: Colour–magnitude arrays for four galactic clusters of different ages.

different open clusters. The feature distinguishing these clusters is the difference in their ages. The youngest cluster, the pair h and χ Persei, a few million years old, contains stars of a great range in luminosity. Some stars have already evolved into the supergiant stage so that the top end of the main sequence is bent over. The stars of luminosity greater than 10,000 times that of the Sun have already largely depleted the hydrogen in their cores and are leaving the main sequence.

The brightest stars of the Pleiades have started to leave the main sequence and are approaching the critical phase when they will have exhausted all the hydrogen in their cores. There are no giants in the Pleiades. Presumably, the cluster contained no stars as massive as some of those found in h and χ Persei.

The Beehive, or Praesepe, cluster is older than the Pleiades. All stars much more luminous than first magnitude have started to leave the main sequence; there are some giants. The Hyades, about 1,000,000,000 years old, display a similar colour–magnitude array. These clusters contain a number of white dwarfs, indicating that the initially most luminous stars have already run the full gamut of evolution. Finally, M67 must be the oldest cluster of all; the best data of stellar structure indicate that it is several thousand million years old. All the bright main-sequence stars have disappeared.

The colour–magnitude diagrams for globular and open clusters differ quantitatively because the latter show a wider range of ages and because of differing chemical compositions. Most globular clusters have smaller metal–hydrogen ratios than have open clusters or the Sun. The gaps between the giants and main-sequence stars of the open clusters (Population I) often contain unstable stars such as variables; the Cepheids fall in these gaps.

The giant stars of the Praesepe cluster are comparable with the brightest stars in M67. The M67 giants have evolved from the main sequence near absolute magnitude +3.5, whereas the Praesepe giants must have evolved from main-sequence stars of absolute magnitude about +0.5. Consequently, the Praesepe giants must have masses about twice as great as those of the M67 giants. Giant stars of the same luminosity may therefore have appreciably different masses. The failure of the mass–luminosity law for giant stars is easily understood.

The components of the giant star Capella have masses three times that of the Sun, whereas each component of the giant binary TW Cancri has a mass 1.3 times that of the Sun. Capella is presumably a relatively young star, TW Cancri an ancient one.

The great spread in the luminosities and colours of giant, supergiant, and subgiant stars is also easily understood. When a star leaves the main sequence, its future evolution is precisely determined by its mass, rate of rotation (or angular momentum), chemical composition, and whether or not it is a member of a close binary system. Giant or supergiant stars of nearly the same radius and surface temperatures may have evolved from main-sequence stars of different ages and masses.

The spread in the main sequence

The spread in the main sequence can also be understood as a consequence of evolution. At the beginning of their lives as hydrogen-burning objects, stars define a nearly unique line in the Hertzsprung–Russell diagram called the zero-age main sequence. Without differences in initial chemical composition or in rotational velocity, all the stars would start exactly from this unique line. As the stars evolve, they burn hydrogen in their cores; the stars adjust to the change in the helium-to-hydrogen ratio in their cores and gradually move away from the zero-age main sequence. When the core is exhausted and a shell source develops, the internal structure of the star changes rapidly; and it quickly leaves the region of the main sequence, as it moves toward the giant or supergiant region. Hence the gap between the main sequence and the domains of these types of stars. Less massive stars leave the main sequence more slowly, as illustrated by the colour–magnitude diagram for M67. Table 3 refers to an average of main-sequence stars, not to the zero-age main sequence.

Possible observations of forming stars

In the clouds of dark obscuring matter in Taurus and other regions of the Milky Way are found variables of the T Tauri type, discussed above. These objects may represent stars in the actual process of formation. Several observations in which stars were actually "turned on" may have been identified. Other evidence is provided by color–magnitude diagrams such as that obtained for the cluster NGC 2264, associated with the star S Monocerotis. In this the brighter main sequence is well defined by stars somewhat more luminous than the Sun. The fainter redder stars, whose colours correspond to spectral classes G, K, and M, all fall above the main sequence defined by normal dwarf stars. Presumably, they are contracting toward the main sequence, shining mostly by liberation of gravitational energy; but they may be burning the light helium isotope helium-3 before they reach the main sequence and the proton–proton reaction is turned on.

Many sites of possible star formation have been identified in the Milky Way, and these are being investigated carefully for possible changes.

Table 6: The Most Abundant Chemical Elements
(by numbers)*

element	symbol	abundance	element	symbol	abundance	element	symbol	abundance
Helium	He	6.3×10^7	Magnesium	Mg	45,000	Potassium	K	76
Carbon	C	500,000	Aluminum	Al	2,140	Calcium	Ca	2,300
Nitrogen	N	100,000	Silicon	Si	32,000	Titanium	Ti	35
Oxygen	O	800,000	Phosphorus	P	320	Chromium	Cr	166
Fluorine	F	80	Sulfur	S	16,200	Manganese	Mn	75
Neon	Ne	93,000	Chlorine	Cl	420	Iron	Fe	22,000
Sodium	Na	1,900	Argon	Ar	1,200	Nickel	Ni	500

*Hydrogen = 10^9.

FORMATION OF CHEMICAL ELEMENTS IN STARS

The relative abundances of the chemical elements provide significant clues to their modes of origin. The Earth's crust has been affected severely by fractionation, so that its present varied composition offers few clues regarding the early stages. The composition of the material from which the solar system was formed is taken from that of stony meteorites, called chondrites, and from the composition of the solar atmosphere supplemented by data from hot stars and gaseous nebulae. Table 6, compiled from the efforts of many workers, lists the most abundant chemical elements.

The most obvious feature is that the abundances of light elements tend to be greater than those of heavier elements. That is, a plot of abundance against atomic mass, A, shows a decline with increasing A to about $A = 100$. Thereafter the abundance is more nearly constant. The abundance plot is not smooth. Among the lighter elements, those of even atomic number tend to be more abundant; and those with atomic numbers divisible by 4 are especially favoured. In particular, lithium, beryllium, and boron are very rare compared with carbon, nitrogen, or oxygen. There is a pronounced abundance peak at iron and less conspicuous peaks at other elements, often connected with the so-called magic-number nuclei. There is also a maximum at lead, the most stable of the heavy elements.

Preponderance of hydrogen

The preponderance of hydrogen suggests that all nuclei were built up from hydrogen, a hypothesis first proposed many years ago. This subject is considered in the article CHEMICAL ELEMENTS, ORIGIN OF. It has been suggested that the elements were created as the consequence of the initial expansion of the universe. At one stage, in this hypothesis, all matter was compressed into a huge neutron ball. As the density decreased, the neutrons decayed into protons and electrons. The protons then captured neutrons, one after another, underwent beta decay (ejection of electrons), and built up heavy elements. The major difficulty with this hypothesis is that atomic masses 5 and 8 are unstable, and there is no way to build many nuclei beyond helium.

A large body of evidence supports the idea that heavy elements are built in stars, although perhaps occasionally in stars of rather unusual types, particularly supernovae, and maybe even in certain supermassive objects not yet identified.

The splitting in the spectral sequence among the cooler stars can be understood in terms of composition differences. The M stars appear to have a "normal" (*i.e.*, solar) composition, with oxygen more abundant than carbon and the zirconium group of elements much less abundant than the titanium group. The R and N, or carbon, stars contain more carbon than oxygen, whereas the S stars appear to have an enhanced content of zirconium as compared with titanium. Other abundance anomalies are found in the high temperature Wolf–Rayet stars, in which objects containing predominantly helium, carbon, and oxygen are distinguished from objects containing helium and nitrogen, some carbon, and little observable oxygen. Significantly, all these abundance anomalies are found in stars believed to be well advanced in their evolutionary development. No main-sequence dwarf stars show such effects.

The most critical observation, however, was the detection of the unstable element technetium, Tc, in the S stars. The element has been produced on the Earth synthetically. The longest-lived isotope, ^{99}Tc, which has a half-life of 200,000 years, is produced from molybdenum, Mo, by reactions in which molybdenum of mass 98 captures a neutron to form molybdenum of mass 99, which decays in about 67 hours to form technetium of mass 99 with the ejection of a negative electron. In symbols this is expressed as $^{98}\text{Mo} + n \rightarrow {}^{99}\text{Mo}$, $^{99}\text{Mo} \rightarrow {}^{99}\text{Tc} + \beta^-$. Furthermore, the technetium in the S stars appears to be about as abundant as the neighbouring elements, ruthenium and molybdenum. The implication is that all these elements were built in the S stars and that these objects have lifetimes of about 200,000 years. How the star gets the heavy elements from the core to the surface without exploding provides an impressive challenge to theoreticians.

Theoretical workers have demonstrated how elements may be built in stars by nuclear processes occurring at very high temperatures and densities. No one mechanism can account for all the elements; hence, several distinct processes, occurring at different epochs during the late evolution of a star, have been proposed.

After hydrogen, helium is the most abundant element. Much of it may have been produced in the initial "big bang," in which the universe was created. Furthermore, it is the normal ash of hydrogen consumption. In the dense cores of highly evolved stars, the helium itself may be consumed to form, successively, carbon-12, oxygen-16, neon-20, and magnesium-24. When the core density and temperature attain values of the order of 20,000 grams per cubic centimetre and 200,000,000° K, respectively, three helium nuclei can become jammed together to form a carbon-12 nucleus. Then, by successive capture of alpha particles, nuclei of oxygen-16, neon-20, and magnesium-24 are built up. Finally, the helium becomes exhausted in the core, which contracts until the temperature reaches about 700,000,000° K, when the carbon nuclei interact with one another. Then fragments such as protons, neutrons, and alpha particles may be produced. These can interact with the oxygen and neon nuclei and others to produce a variety of other elements, such as fluorine, sodium, and others. At 10^9° K, also, oxygen-16 nuclei interact with other oxygen-16 nuclei, producing a wide range of elements from oxygen to silicon. As the core continues to contract and the temperatures and densities are forced even higher, a semi-equilibrium situation is set up wherein nuclides of the "iron peak" are produced.

Special conditions leading to supernova formation

If the temperature and density continue to rise, the iron-group nuclei can break down into alpha particles, but a large amount of energy is suddenly consumed in the process. The star suffers a violent implosion or collapse and then explodes as a supernova. In the supernova explosion, a great variety of nuclear reactions can occur.

When a nucleus captures a neutron, its mass is increased by one unit, and its charge remains the same. Such a nucleus may be too heavy for its charge and may emit an electron or beta ray to reach a more stable state. It then becomes a nucleus of the next element in the periodic table.

Slow and rapid processes

Two extreme situations have been envisaged. In one, called the slow, or *s*, process, the flux of neutrons is low. A nucleus captures a neutron and leisurely emits a beta ray; the nuclear charge increases by one.

Beta decay is often very slow, and, if the flux of neutrons is high, the nucleus may capture another neutron before there is time for it to undergo decay. In this rapid, or *r*, process, the evolution of a nuclide may be very different

from that in a slow process. In supernovae explosions vast quantities of neutrons may be produced, and these could result in the rapid buildup of massive elements. One interesting feature of the building up of heavy elements by neutron capture at a high rate in a supernova explosion is that nuclei much heavier than lead or uranium can be built. These may decay by fission, releasing further amounts of energy.

The superabundant elements in the S stars come from the slow neutron process. A variety of mechanisms of how the neutrons are produced have been proposed. One example, in which the symbols are the same as before, is: $^{13}C + {}^{4}He \rightarrow {}^{16}O + n + (2.6\ MeV)$ energy. Another is between neon-21 and helium-4 to give magnesium-24 and a neutron: $^{21}Ne + {}^{4}He \rightarrow {}^{24}Mg + n$. These neutrons are available for capture by heavier nuclei, particularly those with large target areas, to form yet heavier elements. Thus, the elements of the zirconium group, for example, could be built from lighter elements by successive neutron capture. Other nuclear processes could result in the production of neutrons and the subsequent formation of heavy elements. About 40 percent of the iron in a mixture would be transformed into heavier elements if about 20 neutrons could be produced for every iron atom in the mixture.

Some low-abundance atomic nuclei are proton-rich—*i.e.*, neutron deficient—and cannot be produced by the *s* or *r* process. Presumably, they have been created in relatively rare events—*e.g.*, one in which a quantum of hard radiation, a γ-ray, causes a neutron to be ejected.

No known nuclear process is capable of producing deuterium (heavy hydrogen), lithium, beryllium, and boron in stellar interiors. They are probably produced by the breaking down or spallation of heavier elements, such as iron and magnesium, by high-energy particles in stellar atmospheres or in the early stages of stellar formation. Presumably, these high-energy particles originate as a consequence of electromagnetic disturbances in the neighbourhood of sunspots and are probably connected with phenomena such as flares, which may be more active in early stages of a star's gravitational contraction or early in its life. Some light-element nuclei may be produced by cosmic rays shattering atoms of carbon, nitrogen, oxygen, and other elements in the interstellar medium.

The peculiar A stars comprise a class of objects with strange abundance anomalies, which may arise from mechanical effects—selective radiation pressure, photospheric diffusion and element separation, and so on—rather than from nuclear effects. Some stars show enhanced silicon, others enhanced lanthanides. The so-called manganese stars show great overabundances of manganese and gallium, usually accompanied by an excess of the element mercury. The latter stars show weak helium lines, low rotational velocities, excess amounts of gallium, strontium, yttrium, mercury, and platinum, and absences of elements such as aluminum and nickel. When these stars appear in binaries, the two components often show differing chemical compositions. It is very hard to envisage plausible nuclear processes that would account for the peculiarities of the abundances—particularly the strange isotope ratios of mercury.

Peculiarities of the galactic cosmic-ray abundances are consistent with the suggestion that they originate in supernovae, being particles accelerated in the strong, rapidly varying magnetic fields involved in these objects (see below). On the way to the Earth they suffer shattering encounters with light atoms.

PROBABLE FATE OF A STAR: WHITE DWARFS, PULSARS, AND BLACK HOLES

The final stages of the evolution of a star will depend on its mass, its angular momentum, and whether it is a component of a close binary.

White dwarfs and novae. The evolution of a member of a close double-star system can be markedly affected by the presence of its companion. As the stars age, the more massive one swells up more quickly as it moves away from the main sequence. It becomes so large that its outer envelope falls under the gravitational influence of the smaller star. Material is continuously fed from the more rapidly evolving star to the less massive one, which still remains on the main sequence. U Cephei is a classical example of such a system for which spectroscopic evidence shows streams of gas flowing from the more highly evolved star to the hotter companion, which is now the more massive of the two. Eventually, the latter will also leave the main sequence and become a giant star, only to lose its outer envelope to the companion, which by then may have reached a white-dwarf stage.

Possible causes of novae

Novae appear to be binary stars that have evolved from contact binaries of the W Ursae Majoris type. One component may have reached the white-dwarf stage. Material fed to it from its distended companion appears to produce instabilities that result in violent explosions or nova outbursts. The time interval between outbursts may range from a few score years to hundreds of thousands of years.

Single stars appear to evolve through the red-giant phase to their ultimate state along more straightforward paths. Perhaps in most instances the distended outer envelope of the star simply drifts off into space, while the core settles down as a white dwarf. In other instances the star may explode as a supernova. If the residual mass is less than 1.3 solar masses, the stellar remnant will become a white dwarf (see above). The material becomes a degenerate gas, wherein the electrons are all stripped from individual atoms. In a sense, they resemble conduction electrons in a metal. With sufficient momenta supplied to the electrons, this gas can be compressed to very high densities. The star has no longer any source of energy and simply cools down, although very slowly.

Neutron stars. When the mass of the core lies between 1.3 and 2 solar masses, it can settle down as a neutron star with a density about 10^{14} times that of the Sun and a diameter of the order of 30 miles (50 kilometres). Such a star would have a crystalline solid crust, wherein bare atomic nuclei would be held in a lattice of rigidity and strength some 18 orders of magnitude greater than that of steel. Below the crust the density is similar to that of an atomic nucleus, so the residual atomic cores lose their individuality as their nuclei are jammed together to form a nuclear fluid. The whole star resembles a single atomic nucleus.

Although neutron stars were predicted in the 1930s, their existence was recognized only very recently. In 1967 a radio source (a pulsar) emitting weak pulses, each lasting about 0.3 second with a remarkably constant period of approximately 1.337 seconds, was discovered. Other examples were soon found, including many detected in the Southern Hemisphere, where a pulsar was first identified with a supernova remnant (see PULSARS).

Pulsars as rotating neutron stars

A large body of evidence identifies pulsars as rotating neutron stars. All energy emitted in the pulses comes from a slowing down of this rotation, but only a small fraction appears in radio-frequency pulses. The rest goes into cosmic rays, maybe some into the emission of gravitational energy.

The pulsar at the centre of the Crab Nebula is observed not only in radio frequencies but also in optical and X-ray frequencies, where it emits 100 and 10,000 times, respectively, as much radiation as in the radio-frequency region. The slowing down of the pulsar rotation supplies the energy to account for the synchrotron emission from the Crab Nebula, which ranges from X-ray to radio frequencies.

Pulsar radiation is polarized—both linearly and circularly—and can be understood in terms of a rotating star with a magnetic field of 10^{12} gauss. Various mechanisms have been proposed whereby charged particles can be accelerated to cosmic-ray energies. Possibly most, if not all, galactic cosmic rays originate from pulsars.

Sudden changes have been observed in the rotation rates of pulsars. Early in 1969, the period of the Vela pulsar suddenly decreased and then began to increase again at its normal rate. This period change could be explained by a change of radius of one centimetre, possibly a sudden shrinkage of the crust. The pulsar phenomenon lasts much longer than the observable supernova remnant, but

pulsars cannot be detected at great distances in the Galaxy because the signals become distorted and wiped out by the ionized interstellar medium.

An object similar to a pulsar is probably the X-ray source in Scorpius, whose energy distribution was recently measured by workers who find that it cannot be interpreted by synchrotron radiation nor by interaction of fast electrons with protons; *i.e.*, bremsstrahlung. The emission is that of a hot body with a surface temperature in the millions of degrees and with a diameter of about ten miles (16 kilometres).

Black holes. If the remnant mass exceeds about two solar masses, it will just continue to contract. The gravitational field of the collapsing star will become so powerful that neither matter nor light can escape any more. It is said then that the star has collapsed to a Schwarzschild singularity—or, in popular terms, a black hole. Its powerful gravitational effects on other objects will still be felt. Bodies trapped in its gravitational field, or a laser beam directed at such an object, would just disappear—pulled into a "bottomless pit." To date, such objects remain theoretical, as none have been discovered.

Schwarzschild singularity

Eventually, the Sun will die as a white dwarf, but before that happens it will evolve as a giant star. The Earth will become uninhabitable and probably will be swallowed up in the vastly extended envelope of the red-giant-stage Sun. These events will not come to pass for thousands of millions of years.

BIBLIOGRAPHY. G.O. ABELL, *Exploration of the Universe*, 2nd ed. (1969), a good introductory account of the properties of stars—well written and lucid; L.H. ALLER, *Atoms, Stars, and Nebulae*, rev. ed. (1971), an introductory and semi-popular work emphasizing analyses of starlight, stellar spectroscopy, and evolution; *Astrophysics: The Atmospheres of the Sun and Stars*, 2nd ed. (1963), and *Astrophysics: Nuclear Transformations, Stellar Interiors, and Nebulae* (1954), textbooks intended for senior students, treating ordinary stellar atmospheres, the Sun, variable stars including novae, and stars with extended envelopes; M. SCHWARZSCHILD, *Structure and Evolution of the Stars* (1958), one of the classic books on the subject that it treats, written before widespread applications of modern computers (moderately advanced level); D.D. CLAYTON, *Principles of Stellar Evolution and Nucleosynthesis* (1968), a fully modern, authoritative account of stellar evolution and synthesis—also for the moderately advanced student; B.M. MIDDLEHURST and G.P. KUIPER (eds.), *Stars and Stellar Systems*, vol. 3, 6, and 8 (1961–65), a collection of summarizing review articles written by authorities in each field.

(L.H.A.)

Star Clusters and Stellar Associations

The Sun is a single star. A high proportion of stars in the Earth's Galaxy, however, exists in physically related groups, pairs, multiple stars with a few members, or clusters with many, held together by mutual gravitation. The line of demarcation between such groups is not always strong. Star clusters fall into two main types. Open (formerly called galactic) clusters contain from a dozen to many hundreds of stars, usually in an unsymmetrical arrangement. Globular clusters are old systems containing thousands to hundreds of thousands of stars closely packed in a symmetrical, roughly spherical form. In addition, groups called associations, formed of a few dozen to hundreds of stars of similar type and common origin, whose density in space is less than that of the surrounding field, have now been recognized.

Four open clusters have been known from earliest times: the Pleiades and Hyades in the constellation Taurus, Praesepe (the Beehive) in the constellation Cancer, and Coma Berenices. The Pleiades was so important to some early peoples that its rising at sunset determined the start of their year. The appearance to the naked eye of the Coma Berenices cluster led to the naming of its constellation after the hair of Berenice, wife of Ptolemy Euergetes of Egypt, 3rd century BC, the only constellation named after a historical figure.

Though several globular clusters, such as Omega Centauri and Messier 13 in Hercules, are visible to the unaided eye as hazy patches of light, attention was paid to them only after the invention of the telescope. The first record of a globular cluster, in Sagittarius, was in 1665; it was later named Messier 22; the next, Omega Centauri (see Figure 1), was recorded by the English astronomer and mathematician Edmund Halley in 1677.

By courtesy of Eric Lindsay

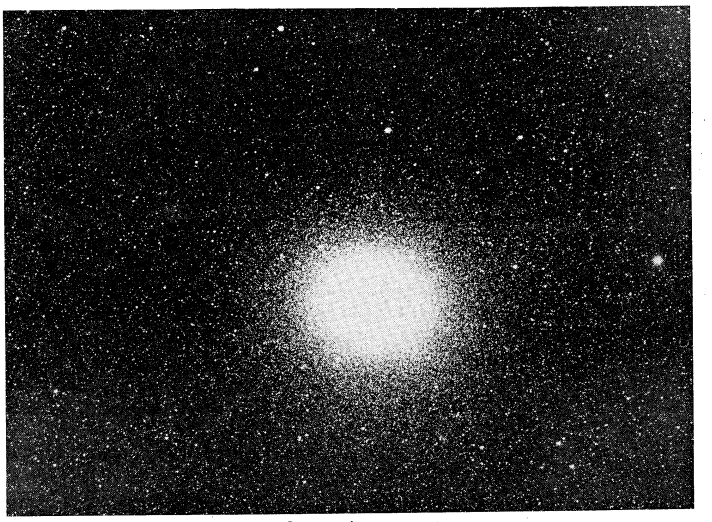


Figure 1: Globular cluster Omega Centauri.

In the late 18th century, a French astronomer, Charles Messier, made a catalogue of 107 "misty objects" so that he would not mistake them for the comets for which he was searching. Objects in this catalogue have a number preceded by M. At the same time, William Herschel began his catalogues of thousands of objects, which were expanded in the first half of the 19th century from observations of the southern sky. In 1888 Johann Ludwig Emil Dreyer, then director of Armagh Observatory, combined all the earlier lists and catalogues into his New General Catalogue (NGC), followed by its supplements, the Index Catalogues (IC), whose numbers are still in use. Of the 125 known globular clusters, 104 have these numbers.

Names and nomenclature

Investigations of globular and open clusters greatly assisted the understanding of the Galaxy. In 1917, from a study of the distances and distributions of globular clusters, Harlow Shapley of Mount Wilson Observatory determined that the galactic centre lies in the Sagittarius region. In 1930, from measures of angular sizes and distribution of open clusters, Robert Trumpler of Lick Observatory showed that light is absorbed as it travels through many parts of space.

The discovery of stellar associations depended on knowledge of the characteristics and motions of individual stars scattered over a substantial area. In the 1920s, it was noticed that young hot blue stars (spectral types O and B) apparently congregated together. In 1949 a Soviet astronomer, Victor Amazaspovich Ambartsumian, suggested that these stars were members of physical groupings of stars with a common origin and named them O associations. (The name OB association is also now sometimes used.) He also applied the term T associations (see below) to groups of dwarf irregular T Tauri type variable stars first noted at Mount Wilson Observatory by Alfred Joy.

The study of clusters in external galaxies began in 1847, when Sir John Herschel at the Cape Observatory published lists of such objects in the nearest external galaxies, the Magellanic Clouds. In this century, the identification of clusters has been extended first to more remote galaxies by the use of large reflectors and later by Schmidt telescopes (see TELESCOPE).

A recent catalogue lists 1,039 open clusters, 5 moving clusters, 11 stellar groups, 70 OB associations, 125 globular clusters, and 28 extragalactic objects. In the present century the number of globular clusters known in the Galaxy has increased by only one third. The number of open clusters known has more than trebled, and the total number is estimated to be 40,000 with 1,000 associations.

GLOBULAR CLUSTERS

Although globular clusters are widely scattered in galactic latitude (see Figure 2), about a third of them are concentrated around the galactic centre, as satellite systems,

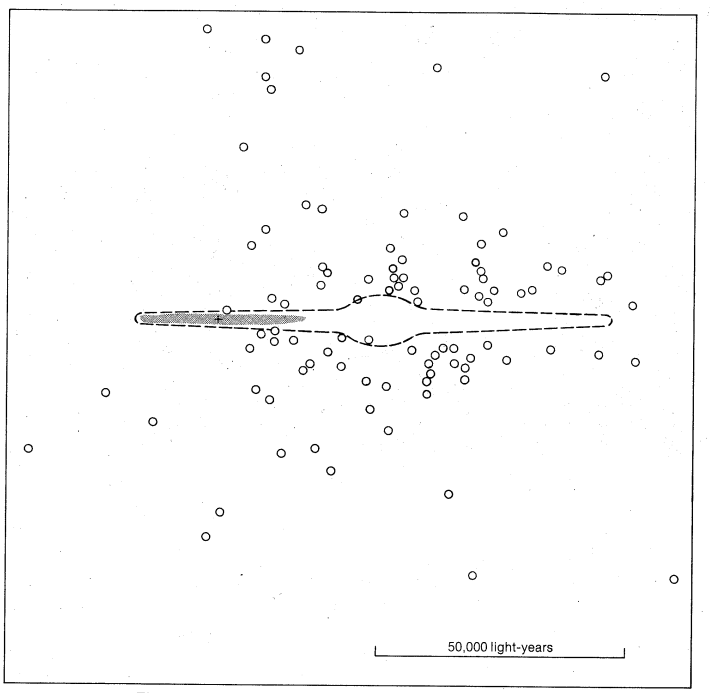


Figure 2: *Distribution of open and globular star clusters in the Galaxy.*
Dotted line indicates probable outline of the Galaxy, a flattened lens-shaped system formed by the stars, as seen edgewise from outside. Eccentric position of the Sun is shown by a cross. Some of the known open star clusters are scattered among the stars in shaded region. Small circles represent globular clusters.

in the rich Sagittarius–Scorpius star fields. Individual cluster masses include up to a million suns, and their linear diameters may be several hundred light-years (one-light-year equals 9.5×10^{12} kilometres); their apparent diameters range from one degree for Omega Centauri down to knots of a minute of arc. In a cluster like M 3, 90 percent of the light is contained within a diameter of 100 light-years, but star counts and the study of RR Lyrae member stars (whose intrinsic brightness varies regularly within well-known limits) indicate a larger one of 325 light-years. The clusters differ markedly in the degree to which stars are concentrated at their centres. Most appear circular and are probably spherical, but a few, such as Omega Centauri, are noticeably elliptic. The most elliptical cluster is M 19, the major axis being about double the minor.

Contents. Globular clusters are composed of what are called Population II objects; that is, old stars. The brightest stars are known as red giants, bright red stars with absolute magnitude (a measure of the true brightness on a logarithmic scale) −2, about 600 times the Sun's brightness or luminosity. In only six clusters have stars as intrinsically faint as the Sun been photographed, and in no globular clusters have the faintest stars yet been recorded. The luminosity function for M 3 shows that 90 percent of the visual light comes from stars at least twice as bright as the Sun, but more than 90 percent of the cluster mass is made up of fainter stars. The density near the centres of globular clusters is two stars per cubic light-year, compared with one star per 300 cubic light-years in the solar neighbourhood. Studies of 50 clusters showed a difference in spectral properties from stars in the solar neighbourhood. (The spectroscopic study of light from a star reveals the structural elements, and compounds, that emitted the light.) This difference proved to be due to a deficiency of metals in the clusters, which have been classified on the basis of increasing metal abundance. Globular cluster stars are 2 to 300 times poorer in metals than stars of the solar neighbourhood. The ratio of metals is higher for clusters near the galactic centre than for those in the halo. The amounts of other elements, such as helium, also differ from cluster to cluster. Radio astronomical studies have set a low upper limit on the amount of neutral hydrogen in globulars. Puzzling features in some globular clusters are dark lanes of nebulous material; though it is difficult to explain the presence of distinct, separate masses of unformed material in old systems, the nebulosity cannot be foreground material between the cluster and observer.

Nearly 2,000 variable stars are known in the 100 globular clusters that have been examined. Of these perhaps 90 percent are members of the class called RR Lyrae variables after the prototype RR Lyrae (formerly called cluster type variables). Other variables in clusters are W Virginis, RV Tauri, and U Geminorum stars, Cepheids, Mira stars, eclipsing binaries (pairs of stars that periodically eclipse each other as seen from Earth), and novae (exploding stars). Most of these variables are named after their prototype star. They are valuable in studies of distance and stellar evolution.

Colour-magnitude and spectrum-magnitude diagrams. The colour of a star has been found generally to correspond to its surface temperature and in a somewhat similar way the type of spectrum shown by a star depends on the degree of excitation of the light-radiating atoms in it, and therefore also on the temperature. All stars in a globular cluster are, within a very small percentage of the total distance, at equal distances from Earth so that the effect of distance on brightness is common to all. Colour-magnitude and spectrum-magnitude diagrams can thus be plotted for the stars of a cluster and the position of the stars in the array, except for a factor that is the same for all stars, will be independent of distance.

All such arrays show a major grouping of stars along a diagonal line called the Main Sequence, and the categories range from hot, bright bluer stars to cool and redder. In many cases a Giant Branch containing very luminous stars is also present (see Figure 3), as well as some other

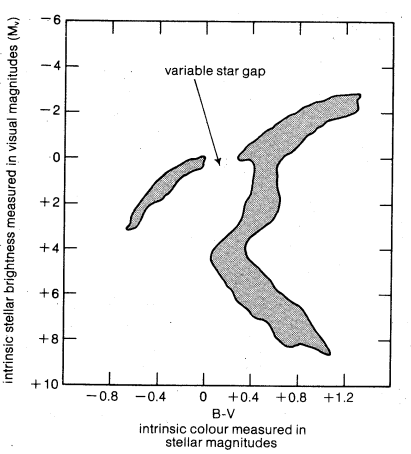


Figure 3: *Colour–magnitude diagram of globular star cluster.*
B and V are magnitudes in blue and visual light. Shaded areas indicate regions occupied by globular stars. Variable-star gap indicates region in which cluster variables occur.

minor groupings, but the extent and concentration of the stars in the various parts of the diagram can be quite different from cluster to cluster.

In 1917 a more distinct difference in colour–magnitude diagrams between open and globular clusters was noticed. This was later explained as a result of differences in the courses of evolutionary change that stars with similar compositions but different masses would follow after long intervals of time. The magnitude at which the brighter stars leave the Main Sequence (called the turnoff point or knee) is a measure of the age of the cluster, assuming that most of the stars formed at the same time. RR Lyrae

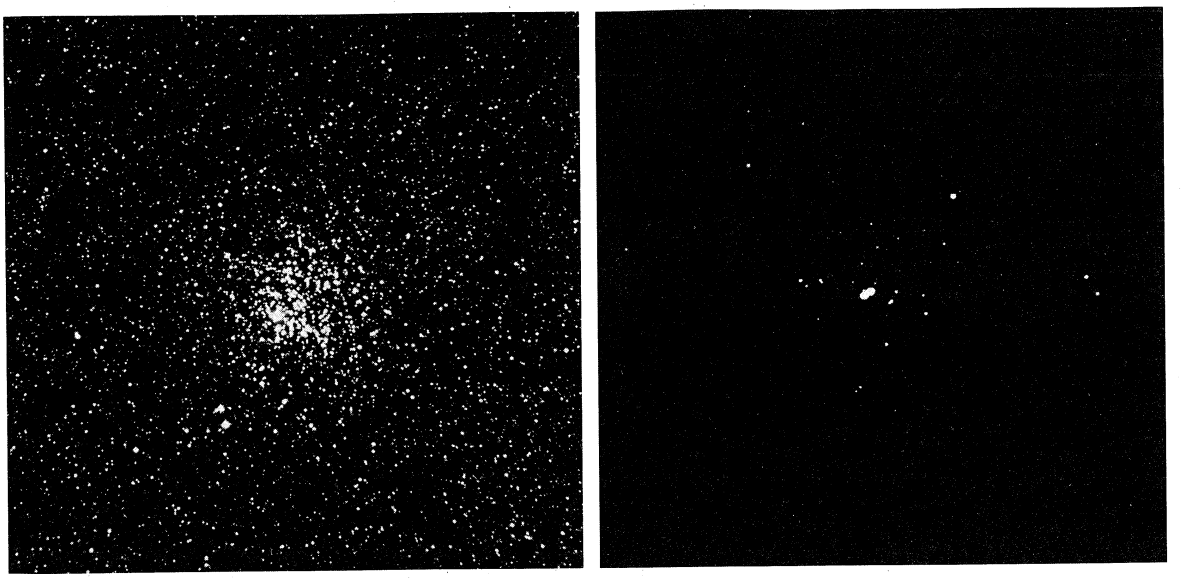


Figure 4: *Typical open clusters.*
(Left) NGC 6705, a rich cluster and (right) NGC 1508, a poor cluster.
By courtesy of Lick Observatory, California

variables, when present, lie in a special region, called the variable gap, near the mean visual magnitude and the blue end of the horizontal (giant) branch of the colour-magnitude diagram. The hydrogen in cluster stars is thought to amount to 70 to 75 percent by mass, helium 30 to 25 percent, and the heavier elements .1 to .01 percent. Globular clusters prove to be as old as the universe, perhaps 12 to 13,000,000,000 years.

Distances and motions. Integrated magnitudes (measures of the total brightness of the clusters), their diameters, and the mean magnitude of the 25 brightest stars made possible the first distance determinations on the basis of the assumption that the apparent differences were due entirely to distance. The colour-magnitude diagram, or the apparent magnitudes of the variables, if their absolute magnitude (the magnitude reduced to a standard distance) is known, however, leads to the best distance estimates. The correction factor for interstellar reddening, due to intervening material that absorbs and reddens their light, is substantial for many but small for those in high galactic latitudes away from the Milky Way. Distances range from about 8,000 light-years for M 71 and NGC 6397 to intergalactic distances of 270,000 light-years for the highly luminous NGC 2419 and 325,000 for the clusters called Palomar 3 and 4.

Velocities and orbits of clusters

The radial velocities (the speed at which objects approach the observer or recede from him, taken as positive when the distance is increasing) measured by Doppler effect have been determined from integrated spectra for a total of 70 clusters. The largest negative velocity is −360 kilometres per second (km/sec) for NGC 6934, the largest positive is +493 km/sec for NGC 3201. These velocities suggest that the globular clusters are moving around the galactic centre in highly elliptical orbits. The globular cluster system as a whole has a rotational velocity of about 140 km/sec relative to the Sun, or 110 km/sec on an absolute basis. For one cluster, Omega Centauri, motions of the individual stars about the massive centre have actually been observed and measured. Though proper motions (the movement across the sky relative to the distant background) of the clusters are very small, those for individual stars provide a useful criterion for cluster membership.

Noteworthy individual clusters. The two clusters of highest absolute luminosity are in the Southern Hemisphere, in the constellations Centaurus and Tucana. Omega Centauri, also called NGC 5139, with (integrated) absolute blue magnitude −9.4, is the second richest cluster in variables, with 171. From this large group, the three types of RR Lyrae stars were first distinguished in 1902. Omega Centauri is relatively near, at a distance of 17,000 light-years, and it lacks a sharp nucleus (Figure 1). The cluster designated 47 Tucanae, or NGC 104, with absolute magnitude −9.5 at a similar distance of 16,000 light-years is of different appearance with strong central concentration. It is near the Small Magellanic Cloud in the sky but is not connected with it. For an observer placed at the centre of this great cluster, the sky would have the brightness of twilight on Earth due to the light of the thousands of stars nearby. In the Northern Hemisphere, M 13 in the constellation Hercules is the easiest to see and is the best known. At a distance of 20,000 light-years, it has been thoroughly investigated, is relatively poor in variables, and has conspicuous lanes of dark nebulosity. M 3 in Canes Venatici, 35,000 light-years distant, is the cluster richest in variables, with 189 known. Their study led to putting the RR Lyrae stars in a special region of the colour–magnitude diagram.

OPEN CLUSTERS

Open clusters are strongly concentrated toward the Milky Way. They form a flattened disklike system 3,000 light-years thick with a diameter about 30,000 light-years; the younger clusters serve to trace the spiral arms of the Galaxy since they are found invariably to lie in them (see GALAXY, THE). Very distant clusters are hard to detect against the rich Milky Way background. A classification based on central concentration and richness (see Figure 4) is in use and has been extended to 852 open clusters. Probably about half the known open clusters contain fewer than a hundred stars, but the richest have a thousand or more. The largest have apparent diameters of several degrees, the diameter of the Taurus cluster being 400′ (nearly 7°), and the Perseus cluster 240′.

The linear diameters run from the largest, 75 light-years, down to five. Increasingly, it is found that a large halo of actual cluster members surrounds the more noticeable core and extends the diameter severalfold; cluster membership is established through common motion, common distances, and so on. Tidal forces and stellar encounters lead to the disintegration of open clusters in long periods of time, as stars "evaporate" from the cluster.

Contents. Stars of all spectral classes from O to M (high to low temperatures; see STAR) are found in open clusters, but frequency of types varies from one cluster to another, as does concentration near the centre. In some (O or OB clusters) the brightest stars are blue and very hot, spectral types O or B, while in others they are type F, whitish yellow and cooler. High luminosity stars are more common than in the solar neighbourhood and dwarfs are much scarcer. The brightest stars in some open clusters are 150,000 times as bright as the Sun. The luminosity of the brightest stars at the upper end of the main se-

Luminosity of stars in open clusters

quence varies in clusters from about −8 to −2 visual absolute magnitude (visual magnitude is that seen by the eye, as opposed to that registered on ordinary photographs; see STAR).

Because of the high luminosity of their brightest stars, some open clusters have a total luminosity as bright as that of some globulars (absolute magnitude −8), which contain thousands of times as many stars. In the centre of rich clusters, the stars may be only 1½ light-years apart. The density may be 100 times that of the solar neighbourhood. In some, such as the Pleiades and the Orion cluster, nebulosity is a prominent feature, while others have none. Masses of neutral hydrogen have been detected with radio telescopes in clusters younger than 25,000,000 years extending over three times the optical diameter of the cluster. Many of the OB clusters mentioned above contain globules, relatively small, apparently spherical regions of absorbing material. The most numerous variables connected with open clusters are the T Tauri type and related stars which occur by the hundreds in some nebulous regions of the sky. Conspicuously absent from open clusters is the type most common in the globular clusters, the RR Lyrae stars. Other variables include: eclipsing binary stars, both Algol type and contact binaries; Cepheids; flare stars and spectrum variables, such as Pleione. The last named, one of the Pleiades, is known to cast off shells of material from time to time, perhaps a result of its high rotational speed (up to 200 miles per second).

Colour–magnitude and spectrum–magnitude diagrams. The colour– or spectrum–magnitude diagram derived from the individual stars holds vital information. Colour–magnitude diagrams are available for 189 clusters on the UBV photometric system; in this system the colour is measured from the amount of light radiated by the stars at ultraviolet, blue, and visual (yellow) wavelength regions. In young clusters, stars are found along the luminous bright blue branch, while in old, beyond a turnoff only a magnitude or two brighter than the Sun, they are red giants and supergiants (see Figure 5).

From A. Sandage and O. Eggen, *The Astrophysical Journal* (1969); The University of Chicago Press

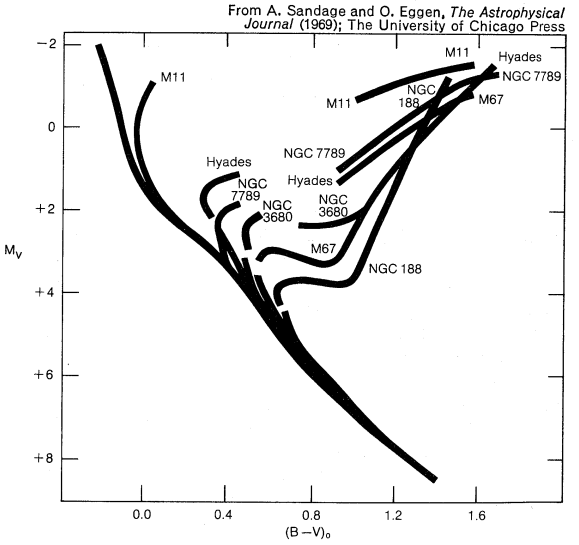


Figure 5: Composite colour–magnitude diagram for six open clusters fitted to the main sequence for very young stars (zero-age). M_V is visual magnitude; $(B-V)_0$ is the colour (blue magnitude minus visual magnitude) for the zero-age stars.

Distances and motions. Distances can be determined by many methods; geometric, photometric, and spectroscopic, with corrections for interstellar absorption. For the very nearest clusters, direct (trigonometric) parallaxes may be obtained and these are inversely proportional to the distance (see PARALLAX, ASTRONOMICAL). Distances may be derived from proper motions, apparent magnitudes of the brightest stars, and spectroscopically from individual bright stars. Colour–magnitude diagrams, fitted to a standard plot of the Main Sequence, give a common and reliable method of determining distance. The nearest open cluster is the nucleus of the Ursa Major group at 75 light-years; the most distant clusters are thousands of light-years away.

Motions, including radial velocities and proper motion, have been measured for thousands of cluster stars. The radial velocities are much smaller than for globular clusters, averaging tens of kilometres per second, but the proper motions are larger. Open clusters share in the galactic rotation. Used with galactic rotation formulas, the radial velocities give another means of distance determination.

Moving clusters

A few clusters are known as moving clusters because the convergence of the proper motions of their individual stars toward a "convergent point" is pronounced. The apparent convergence is caused by perspective; the cluster members are really moving as a swarm in almost parallel directions and with about the same speeds. The Hyades is the most prominent example of a moving cluster. (The Hyades stars are converging with a velocity of 45 km/sec toward the point in the sky with position coordinates, right ascension 94°, declination +7° .6.) The Ursa Major group, another moving cluster, occupies a volume of space containing the Sun; but the Sun is not a member. The cluster consists of a compact nucleus of 14 stars and an extended stream.

Stellar groups are composed of stars presumed to have been formed in a batch together, but the members are now too widely separated to be recognized as a cluster.

Noteworthy individual clusters. The Pleiades is the best known and perhaps the most thoroughly studied. This cluster, with a diameter of 35 light-years at a distance of 400 light-years, has about 500 stars and an age of 150,000,000 years. Near the Pleiades in the sky, but not so conspicuous, the Hyades is the second nearest cluster at 130 light-years. Its stars are similar to those in the solar neighbourhood, and it is an older cluster, with an age of 2,000,000,000 years. Measurements of the Hyades form a basis for astronomical determinations of distance and age because its thoroughly studied Main Sequence is used as a standard. A "poor" cluster with few stars is Coma Berenices at a distance of 260 light-years, containing only about 40 stars. Of the extremely young clusters, that associated with the Orion nebula is the nearest, 1,500 light-years away, with an age of 3,000,000 years. A still younger cluster is NGC 6611, whose age after formation is 800,000 years. At the other end of the scale some open clusters have ages approaching those of the globular clusters and seem to be of intermediate type. Messier 67 in Cancer, at a distance of 3,000 light-years, with 500 solar masses, is 5,500,000,000 years old. NGC 188 in Cepheus is the oldest open cluster at 8×10^9 or 9×10^9 years.

ASSOCIATIONS

The chief distinguishing feature of the members of a stellar association is that the large majority of its stars have similar physical characteristics. The OB association consists of large numbers of hot, blue giant stars, spectral classes O and B, and relatively small numbers of other objects. A T association consists of cooler dwarf stars, many of which exhibit irregular variations in brightness. The stars clearly have to be relatively near to each other in space, though in some cases they may be widely dispersed in the sky and are less closely placed than in the open clusters.

OB associations. The existence of an OB association is usually established through a study of the space distribution of early type O and B stars. It will appear as a concentration of points in a three-dimensional plot of galactic longitude and latitude and distance. Seventy are catalogued, designated by constellation abbreviation and number; *e.g.*, Per OB 1 in the constellation Perseus. In dimensions they are larger than open clusters, from 100 to 700 light-years, and usually contain one or more open clusters as nuclei. They frequently contain a special type of multiple star, the Trapezium, named for its prototype in Orion, as well as supergiants, binaries, H II regions and globules. Associations are relatively homogeneous in age. Best distance determinations are from spectroscopic parallaxes of individual stars; *i.e.*, estimates

of their absolute magnitudes made from studies of their spectra. Most of those known are nearer than 10,000 light-years, with the nearest, named Sco OB 1, at 650 light-years.

Associations appear to be almost spherical, though rapid elongation would be expected from the shearing effect of differential galactic rotation. It has been suggested that expansion, which is of the order of 10 kilometres per second, masks the tendency to elongate, and this is confirmed in some. Tidal forces will break up an association in less than 10,000,000 years through differences in the attraction by an outside body on members in different parts of the association.

Lifetime of an association

A good example is Per OB 1 at a distance of 6,000 light-years, spreading out from the double cluster *h* and Chi Persei. A large group of 20 supergiant stars of spectral type M belongs to it. Associations with red supergiants may be in a relatively advanced evolutionary stage, almost ready to disintegrate.

T associations. The T Tauri associations or T associations are formed by groups of T Tauri stars generically associated with the clouds in which they are found. About three dozen are recognized. A T Tauri star is characterized by irregular variations of light, low luminosity, and hydrogen line (H-alpha) emission. It is a newly formed star of intermediate mass that is still in the process of contraction from diffuse material. The small motions of the T Tauri stars relative to the nebula indicate that they are not field stars passing through the nebula. They are found in greatest numbers in regions with bright O and B stars. T associations are found only in or near regions of galactic nebulosity, either bright or dark, and only in obscured regions, showing the presence of dust. Besides T Tauri stars, they include related variables, nonvariable stars, and Herbig-Haro objects. The last-named are small nebulosities 10,000 astronomical units in diameter; each contains several starlike condensations in configurations similar to the Trapezium, Theta Orionis, in the sword of Orion. These objects are considered to be star groups at the very beginning of life.

The constellation of Cygnus has five T associations and Orion and Taurus four each. The richest is Ori T2 with over 400 members, with a diameter 50 by 90 light-years, at a distance of 1,300 light-years, around the variable T Ori.

RELATIONSHIP AND SIGNIFICANCE

Seen from intergalactic space the Galaxy would appear as a giant luminous pinwheel, with more than a hundred globular clusters dotted around it. The richest parts of the spiral arms of the pinwheel would be marked by dozens of open clusters. If this panorama could be seen as a time-lapse movie, the great globular clusters would wheel around the galactic centre in elliptical orbits with periods of hundreds of millions of years; open clusters and associations of stars would be seen to form out of knots of diffuse matter in the spiral arms, gradually disperse, run their life cycle and fade away, while the Earth's Sun pursued its course around the galactic centre for thousands of millions of years.

Young open clusters and associations, occupying the same region of space as clouds of ionized hydrogen, help to define the spiral arms. A concentration of clusters in the bright inner portion of the Milky Way between galactic longitudes 283° and 28° indicates an inner arm in Sagittarius. Similarly, the two spiral arms of Orion and Perseus are defined between 103° and 213°, with a bifurcation of the Orion arm. Associations show the existence of spiral structure in the Sun's vicinity. Older clusters, whose Main Sequence does not reach to the blue stars, show no correlation with spiral arms because in the intervening years their motions have carried them far from their place of birth.

All the O and B type stars in the Galaxy may have originated in OB associations. The great majority, if not all, of the O stars (5,000) were formed and still exist in clusters and associations; though only 10 percent of the total number of B stars are now in OB associations or clusters, it is likely that all formed in them. At the other (fainter) end of the range of stellar luminosities, the number of dwarf variable stars in the nearby T Tauri associations is estimated at 12,000. These associations are apparently the main source of low luminosity stars in the Sun's neighbourhood.

While large numbers of associations have formed and dispersed and provided a population of stars for the spiral arms, the globular clusters have survived relatively unchanged except for the evolutionary differences that time brings. They are too massive to be disrupted by the tidal forces of the Galaxy, though their limiting dimensions are set by these forces when they most closely approach the Galactic centre. Impressive as they are individually, their total mass of 10,000,000 suns is small in comparison to the Galaxy—only about 1/10,000. Their substance is that of the Galaxy in a very early stage. The Galaxy may have collapsed from a cloud of gas made up almost entirely of hydrogen and helium. Between 12×10^{12} and 13×10^{12} years ago, before the last stages of the collapse, material forming the globular clusters may have separated from the rest. The fact that metal-rich clusters are near the galactic nucleus while metal-poor clusters are in the halo or outer fringes may indicate a nonuniform distribution of elements throughout the primordial mass.

In a study of star clusters a time panorama unfolds—from the oldest things existing in the Galaxy, the globular clusters, through clusters in existence only half as long, down to extremely young open clusters and associations that have come into being since the ancestors of man first trod the Earth.

CLUSTERS IN EXTERNAL GALAXIES

Clusters have been found and studied in at least eight external galaxies, mostly members of the Local Group. At their great distances the classification is difficult, made from studies of the colours of the light from the whole cluster (integrated colours) or, for a few, from colour magnitude diagrams. An electronograph (image tube) on a large telescope in the uniquely good observing conditions of Chile has extended studies to cluster stars of apparent magnitude 23, that is a factor of 10^{10} fainter than the bright star Sirius.

Clusters have been found by the hundreds in some of the nearest galaxies. At the distance of the Magellanic Clouds, a cluster like the Pleiades would appear as a faint 15th magnitude object, subtending 15 seconds of arc instead of several degrees, and would perhaps not yet have been found. In the Small Magellanic Cloud, at a distance of 200,000 light-years, about 120 clusters are known. In the Large Magellanic Cloud, at a distance of 180,000 light-years, nearly 1,200 have been catalogued. Most of them are young open clusters, but about 60 have globular characteristics. In both clouds, the brightest clusters are the young blue giant open clusters such as NGC 330 and NGC 1866. The open clusters contain some Cepheid variables, and in chemical composition are similar to, but not exactly the same as, those of the Galaxy. The globular clusters fall into two distinct groups. Those of the first group, the red, have a large metal deficiency similar to those in the Galaxy; RR Lyrae variables have been found in some. Those of the second group are large, circular in outline, with colours much bluer than normal galactic globular clusters, and ages of about 10^6 to 10^9 years. They are similar to the open clusters of the Clouds but very populous. The observed differences between clusters in the Galaxy and the Magellanic Clouds are due to small differences in helium or heavy-element abundances. There are 122 associations with a mean diameter of 250 light-years, somewhat richer and larger than in the Galaxy. Sixteen of the associations contain coexistent clusters. Also 15 star clouds are recognized.

Cluster characteristics

In the great Andromeda spiral galaxy, M 31, 2,200,000 light-years away, more than 300 clusters are known, of which, perhaps, two-thirds are globular. Colour studies of 46 show them to have a higher metal content than galactic globulars. A total of 188 OB associations is known with distances up to 80,000 light-years from the nucleus. Diameters of their dense cores are comparable to those of galactic associations. NGC 206 (OB 78) is the richest star cloud in M 31, with a total mass of 200,000 suns and

a strong resemblance to the double cluster in Perseus. Around the dwarf elliptical companions to M 31, NGC 185, and NGC 205, some globular clusters have been found.

M 33 in Triangulum, a late-type spiral galaxy, has about 300 known clusters, not many of which have globular characteristics. Of the six dwarf spheroidal galaxies in the Local Group, only the one in the constellation Fornax has clusters. Its five globular clusters are similar to the bluest galactic globular clusters. In the irregular galaxies NGC 6822 and IC 1613 no clusters have been found.

Beyond the Local Group, at a distance of 45,000,000 light-years, the giant elliptical galaxy M 87, in the Virgo cluster, is surrounded by an estimated 2,000 globular clusters. Inspection of other ellipticals in Virgo shows that they too have globular clusters whose apparent magnitudes are similar to those in M 87 although their stellar population is substantially smaller. There is indication that the mean absolute magnitudes of globular clusters are constant and independent of the absolute luminosity of the parent galaxy.

The total number of clusters now known in external galaxies far exceeds the number known in the Galaxy.

BIBLIOGRAPHY

Comprehensive works: H. SHAPLEY, *Star Clusters* (1930); H.S. HOGG, "Star Clusters," in *Handbuch der Astrophysik*, vol. 53 (1959); G. ALTER *et al.*, *Catalogue of Star Clusters and Associations* (1970).

Globular clusters: H.C. ARP, "Globular Clusters in the Galaxy," in *Galactic Structure* (1965); G.E. KRON and N.U. MAYALL, "Photoelectric Photometry of Galactic and Extragalactic Star Clusters," *Astr. J.*, 65:581–620 (1960); R.E. WHITE, "A Bibliography of Color Magnitude Diagrams for Globular Clusters," *Astrophys. J. Suppl.*, 19:343–366 (1970); I.R. KING *et al.*, "Star Counts in 54 Globular Clusters," *Astr. J.*, 73:456–491 (1968); I. IBEN, JR., "Globular-Cluster Stars," *Scient. Am.*, 223:26–39 (1970).

Open clusters: R.J. TRUMPLER, "Preliminary Results on the Distances, Dimensions, and Space Distribution of Open Star Clusters," *Lick Obs. Bull.*, no. 420 (1930); O.J. EGGEN, "Moving Groups of Stars," in *Galactic Structure* (1965); G.L. HAGEN, "An Atlas of Open Cluster Colour-Magnitude Diagrams," *Publs. David Dunlap Obs.*, vol. 4 (1970); M.F. WALKER, "Studies of Extremely Young Clusters," *Astrophys. J.*, 155:447–468 (1969); A. SANDAGE and O.J. EGGEN, "Isochrones, Ages, Curves of Evolutionary Deviation, and the Composite C-M Diagram for Old Galactic Clusters," *Astrophys. J.*, 158:685–698 (1969).

Associations: V.A. AMBARTSUMYAN, "Stellar Associations," *Endeavour*, 18:45–51 (1959); M.S. ROBERTS, "The Number of Early-Type Stars in the Galaxy and their Relation to Galactic Clusters and Associations," *Publs. Astr. Soc. Pacif.* 69:59–64 (1957); P.N. KHOLOPOV, "A Revised List of T-Associations and their Members," *Soviet Astr.*, 3:291–300 (1959); G.H. HERBIG, "The Properties and Problems of T Tauri Stars and Related Objects," in *Advances in Astronomy and Astrophysics*, vol. 1 (1962); S. SHARPLESS, "Distribution of Associations, Emission Regions, Galactic Clusters, and Supergiants," in *Galactic Structure* (1965).

Clusters in external galaxies: (Magellanic clouds): S.C.B. GASCOIGNE and G.E. KRON, "Colors and Magnitudes of some Star Clusters in the Magellanic Clouds," *Publs. Astr. Soc. Pacif.*, 64:196–200 (1952); E.M. LINDSAY, "The Cluster System of the Small Magellanic Cloud," *Mon. Not. R. Astr. Soc.*, 118:172–182 (1958); P.W. HODGE and F.W. WRIGHT, *The Large Magellanic Cloud* (1967), an atlas and booklet; B.J. BOK, "The Magellanic Clouds," *Jl. R. Astr. Soc. Can.*, 63:105–124 (1969); P.B. LUCKE and P.W. HODGE, "A Catalogue of Stellar Associations in the Large Magellanic Cloud," *Astr. J.*, 75:171–175 (1970); M.F. WALKER, "Image-tube Observations at Cerro Tololo," *Sky Telesc.*, 40:132–138 (1970). (Messier 31): S. VAN DEN BERGH, "Stellar Associations in the Andromeda Nebula," *Astrophys. J. Suppl.*, 9:65–80 (1964), "Photometric and Spectroscopic Observations of Globular Clusters in the Andromeda Nebula," *ibid.*, 19:145–174 (1969). (Other galaxies): R. RACINE, "2000 Globular Clusters in M87," *J. R. Astr. Soc. Can.*, 62:367–376 (1968); G. DE VAUCOULEURS, "The Brightest Star Clusters in Galaxies as Distance Indicators," *Astrophys. J.*, 159:435–441 (1970).

(H.S.H.)

State, The

The "state," as the term refers to the political organization of society, is used in two ways, one more general, meaning the whole body of persons who are politically organized, the other more specific, meaning the institutions of government. In the first usage the state comprehends all the citizens; in the second, it may be contrasted with the citizens. Since the second meaning is just as well, if not more clearly, expressed by the term government, it is best, for purposes of exposition and understanding, to concentrate on the idea of the state as being the whole body politic. Modern political theorists generally define the state as a special form of association (an association being a deliberately organized social group). It is distinguished from other forms of association by its purposes, by the methods it employs to carry out those purposes, by having territorial limits, and by its sovereignty. The primary purpose of the state is to maintain order and security. It does so by means of a system of law backed by force. The exercise of its jurisdiction is limited to a definite geographical area, but within that area the state is sovereign; that is to say, it has supreme and exclusive authority.

THE CONCEPT OF THE STATE

The use of the word state for politically organized society is relatively recent. It became prominent in the 16th century, being first popularized by the Italian political theorist Niccoló Machiavelli in *The Prince* (1513). The ancient Greeks spoke of the *polis* ("city-state"). In classical Latin —as, for example, in Cicero—*res publica* ("commonwealth") was the most favoured expression, though *civitas* ("community of citizens") was also used. Medieval writers tended to prefer the terms *civitas* and *regnum* ("realm" or "kingdom"). The Greek *polis* differed from a modern state not only in being smaller but also in embracing more of social life. The Greek conception of the *polis* included religious and cultural activities as well as government, so that the later distinction between church and state could not be expressed in terms of church and *polis*. The Roman *res publica* was a larger entity, for it extended beyond the city of Rome to Roman citizens throughout Italy and the provinces. Another difference, at least in Cicero's usage, was that the Roman idea of *res publica* emphasized the legal structure of political association. In medieval political writings the words *regnum* and *civitas* were used sometimes indiscriminately but often with a distinction, *civitas* referring to the city-state that existed in some parts of Europe and especially in Italy, *regnum* referring to the larger type of monarchy that was growing up in other parts of Europe. The term *res publica* came to be used for a larger entity still, *res publica christiana*, the community of Christendom.

Earlier terms for the body politic

In the political philosophies of Plato and Aristotle, the *polis* was treated as the ideal form of community, because it was self-sufficient in providing for economic and social needs and for moral development. The ideas of value that were included in the Greek concept were continued in the connotations of *res publica* and of its English translation, "commonwealth." Self-sufficiency, which Aristotle called the essence of the *polis*, is quite different from sovereignty, the essential feature of the modern state. Machiavelli thus used a new word for a new idea.

Sovereignty and the state. Sovereignty is the most fundamental of the distinctive attributes of the state as it has been conceived and understood since the 16th century. Some theorists have emphasized sovereignty in its legal aspect, supreme and independent authority for making, interpreting, and applying a system of law over a defined territory. Other theorists have laid more stress on "political" sovereignty, the possession of coercive power sufficient to enforce obedience internally and to maintain independence in external relations. The two forms of sovereignty in fact go together, because the effectiveness of legal authority depends in the last resort on its being backed by coercive power superior to any possessed by individuals or groups within the state.

The word state is used to refer to some politically organized communities that are not sovereign states, notably the constituent members of a federal union, such as the constituent states of the United States of America or of the Commonwealth of Australia. In such instances the feder-

al union is the sovereign state; it not only has a superior authority in internal affairs, but it alone has the right to treat with foreign states and international bodies. Other types of political unit that possess some of the powers of a state but lack sovereignty are protectorates, self-governing colonies, and trust territories.

Other attributes of statehood. Certain other distinctive attributes of the state emerged during the 16th, 17th, and 18th centuries. First, there must be a considerable measure of stability and continuity in respect of defined borders and consequently of population. Since states recognize each other's sovereign jurisdiction over their respective territories, these territories need to be clearly defined. Uncertain or frequently shifting borders are a source of friction and hostility. Men need to know that a particular area and its inhabitants owe political allegiance to one definite authority. Stable borders and populations go along with the growth of a common nationhood. The state in the modern world tends to be a nation-state, although there are many exceptions. A nation is a community that shares a number of attributes such as a common language and culture; it is a community of persons who feel that they belong together because of shared interests and history. To some extent the bonds of nationhood have contributed to the formation of states, but it is equally true that the experience of being a state, of being ruled by the same authority and the same laws, has caused a community gradually to feel a sense of common nationality.

State and nation

The primary function of a state, maintaining order and security by means of law backed by force, naturally entails a number of permanent institutions. There must be governmental bodies for forming policies and making laws; a civil service for carrying out governmental policy; a judicial system with police and penal services to apply the law; and armed forces for defense against external enemies and, if need be, for putting down internal riot. It is all these organs of administering a state that are referred to when the word state is used in the narrower of the two senses described at the beginning of this article.

THE POLITICAL PHILOSOPHY OF THE STATE

Since sovereignty is the feature of the modern state that distinguishes it most markedly from earlier conceptions of political society, it is not surprising that political theory should initially stress the importance of absolute authority. The French and English political philosophers Jean Bodin and Thomas Hobbes are the first who made explicit the idea of state sovereignty. Machiavelli, who preceded them, took it for granted.

Machiavelli. Machiavelli was not a political philosopher in the ordinary sense. He did not try either to define the state or to justify its existence. His views about the state, as about much else, are implied as a matter of course when he describes how a ruler may retain or acquire firm control, how he is liable to lose it, which qualities are necessary for a republic to remain strong, or how precarious a thing is its liberty. Medieval thinkers had taken the political authority of any prince or king in the community of Christendom to be necessarily limited —by the supposedly superior authority of the emperor, by the parallel authority of the church in spiritual matters, and by the authority of natural law (that is, universal moral principles), which determined the bounds of justice. Machiavelli did not argue against this theoretical position. He simply ignored it, writing as if it were an obvious matter of fact that a state has absolute authority. He thought that the value of religion lies purely in its contribution to social order and that the ordinary rules of morality must go by the board if security requires it.

Machiavelli's absolutist assumptions

Machiavelli again differed from medieval thinkers in taking it for granted that the power of the state is a single whole and can be centrally controlled, irrespective of whether a particular state has the form of a monarchy or a republic. He preferred a republic because he preferred liberty. But he believed that the liberty of a republic can be sustained only if the people have the "virtue"—that is, the vitality, courage, and independence—that the ancient Romans possessed, but which, as he thought, existed hardly anywhere in the Europe of his own day except among the Germans. Where the people are "corrupted," especially by the power of nobles, the state needs a single ruler with a strong hand, ready to be ruthless in preserving order.

Bodin. Bodin was the first political theorist to isolate the concept of sovereignty and to see it as the essential feature that distinguishes the state from other organizations. He defined sovereignty as the absolute and perpetual power of making and unmaking laws. By "absolute" he meant unlimited, and by "perpetual" he meant not confined to any fixed period of time. If the right to legislate were of limited duration, Bodin argued, the holder of it would be merely a trustee for the real sovereign, the possessor of permanent authority. Bodin also implied that sovereignty is indivisible, although this is not included in his definition. He followed the traditional classification of forms of government into monarchy, aristocracy, and democracy, and said that in an aristocracy or democracy the sovereign is the ruling group as a whole; sovereignty cannot be shared. In consequence, he believed that there cannot be a state with a mixed constitution, in which the sovereign authority is divided among different offices. The sovereign often has to act through agents, but the agents do not thereby share in the sovereignty. Their authority is strictly subordinate.

Bodin's limitations on state sovereignty

Bodin, nevertheless, did not make as clean a break with medieval thought as did Machiavelli. There is a real sense in which his conception of sovereignty is not absolute but is limited by natural law, by constitutional law, and by property rights. In the first place, Bodin was not prepared to treat arbitrary power as political authority. The chief of a robber band who holds absolute sway over his associates and obeys no superior is not a sovereign, in Bodin's view, because his rule is not law; it is not based on justice; sovereign authority must have a moral foundation in natural law. Secondly, Bodin made sovereignty subject to the constitutional law that determines succession. And thirdly, Bodin held that sovereignty may not interfere with property rights. Legislation pertains to the state, property to the family. The sovereign has an "absolute" power to legislate, but he may not levy taxes without consent.

The fact is that, when Bodin described the sovereign's authority as absolute, he meant unlimited by the authority of any other office. The sovereign's commands are subordinate to natural law but not to the interpretation of natural law by the church; the sovereign is perfectly capable of seeing for himself what natural law requires. Bodin's purpose was to deny that feudal lords, whether temporal or spiritual, had any final right of jurisdiction independent of the king's authority.

Hobbes. Philosophical justification of an absolute state reached its acme in the political theory of Hobbes, who painted a grim picture of what life would be like without political order. Since men are largely selfish, are roughly equal in strength, and are liable to compete with each other in satisfying their wants, the "natural condition" of mankind is a "war of every man against every man," in which there is "continual fear, and danger of violent death; and the life of man, solitary, poor, nasty, brutish, and short." The only way to avoid this, Hobbes argued, is to maintain a state ruled by a sovereign (man or assembly) with absolute authority. Supreme authority must be absolute and undivided, because otherwise there may be conflict between the holders of different offices (such as Hobbes saw before and during the English Civil War between Charles I and his Parliament), and that is to be back in the natural condition of war. Absolute rule may be oppressive, but at least it secures men from the worst of evils, a state of war.

Hobbes's concept of natural right

In the hypothetical condition of nature, according to Hobbes, every man has a "natural right" to do whatever he thinks fit. It must be assumed that, when men form or accept the political order of a state, they have given up the major part of their natural right, they have authorized a sovereign to issue laws, and they have promised to obey those laws. The consequence is that the citizens, in place of their former natural right or freedom to do as they like, are under obligation to the sovereign.

The sovereign retains the whole of his natural right; unlike the citizens, he makes no promise and so does not limit his freedom of action by undertaking any obligation. In addition, since the citizens have obliged themselves to obey his commands, he is entitled to their support when he orders it.

Although the sovereign, in Hobbes's theory has an absolute right to issue whatever commands he thinks fit, the subjects do not have an equally absolute obligation to obey any and every command. Some parts of their original natural right cannot be given up. Hobbes believed that the fundamental motive in all men, as in other living creatures, is the desire for self-preservation. This is what induces men to accept political society in place of the natural condition, in which life is apt to be short. But then one cannot suppose that a man who promises political obedience for the sake of self-preservation intends to promise obedience to *any* command, such as a command that he should kill himself or that he should not resist an attempt to execute him. There are therefore certain things that the sovereign cannot expect of his subjects.

The implicit promise of obedience that the subjects have given, moreover, is for the sake of protection; and, if the sovereign is in fact unable to protect them, their obligation ceases. This is not to say that there is a contract of mutually dependent obligations between sovereign and subjects. The sovereign has not promised to provide protection in return for the subjects' promise to obey. But he has a duty of prudence to provide it. A wise sovereign knows that his power will last only so long as he can maintain order and protection. He, like everybody else, wants to avoid a relapse into the condition of nature, and for this reason he should follow the moral principles called laws of nature, which, in Hobbes's view, are maxims of prudence. In practice, therefore, a sovereign's freedom of action is limited if he wishes to retain his power. But his formal authority, what he is entitled to do, is absolute.

Locke. Absolute sovereignty for the state need not mean absolute monarchy. Machiavelli thought that a monarchy was the only practical regime for most peoples; Bodin positively preferred it to aristocracy and democracy; and Hobbes wrote as if his sovereign were more likely to be an individual than an assembly. All three, nevertheless, regarded popular forms of government as perfectly legitimate. The idea that absolute monarchy is the only proper form was put forward in the 17th century by certain other writers, notably Sir Robert Filmer in England and Jacques-Bénigne Bossuet in France. Both supported their view with the doctrine of the divine right of kings. It naturally provoked reactions and of more than one kind.

The English philosopher John Locke, for example, built his political philosophy upon the natural rights of all men, as did Hobbes, but with a different conception of natural right. By a right Hobbes meant the absence of obligation, a freedom or license to act as one wishes. Locke meant a justified claim on another person, correlative to an obligation or duty that that other person has toward the one who has the right. Locke's natural rights arise from the traditional view of natural law as a moral law imposed upon men by God. If God commands all men to do no murder, the consequence is that every man has a duty not to kill others, and correspondingly every man has a right or justified claim against others that they should not kill him. This is the natural right to life. Locke wrote the three main natural rights: to life, liberty, and property. Unlike Hobbes, Locke believed that a state of nature is not a state of war or one in which there is no sense of community. The state of nature is, however, inconvenient, because it lacks an impartial judge to arbitrate disputes and to punish breaches of natural law. Men are therefore willing to set up a state or political society, the function of which is to protect their natural rights.

Locke differed from Hobbes again in believing that rebellion is sometimes justified, as in the revolution of 1688 against James II. To avoid Hobbes's supposition that the overthrow of a sovereign means the dissolution of the state, Locke distinguishes the state's source of authority from that of the government. The state rests on a social contract, government on a trust. A government is entrusted with power in order to protect life, liberty, and property. If it fails to do so, it forfeits its claim to obedience, and the citizens are entitled to set up a different government instead.

Influence of Locke's ideas

Locke's theory had a considerable influence on the basic political ideas of the American Declaration of Independence, except that the pursuit of happiness replaced property as the third of the cardinal natural rights. Although Locke himself said that property, understood in a wide sense, comprises all natural right, it would be more accurate to regard liberty as the most comprehensive of his natural rights. The state for Locke has the negative function of preventing men from encroaching on each other's proper sphere of liberty in regard to their persons, property, and desires generally. From this perspective Locke's political theory marked the beginning of liberalism. The same tendency (albeit qualified) appears in his arguments for the toleration of religious dissenters. But Locke's distinction between the legislative and executive functions of government is less significant for the development of ideas about liberty. It was the 18th-century French theorist Montesquieu who saw the separation of the legislative, executive, and judicial powers as the best guarantee of liberty in the state. The explicit separation of powers in the American Constitution is due to the influence of Montesquieu.

Rousseau. While the American conception of the state owes much to Locke and Montesquieu, the French one owes more to Jean-Jacques Rousseau. The French Declaration of the Rights of Man and of the Citizen begins, like the American Declaration of Independence, with an assertion that the purpose of political association is to preserve natural rights, but it then goes on to the novel principle that sovereignty resides in the nation and that law expresses the general will. These are the views of Rousseau, who went further than Locke in limiting the powers of the executive but who insisted that the sovereign authority to legislate lay with the whole people and could be neither transferred nor even deputed to representatives. This last idea, rejecting representative democracy and requiring that laws be made by all the citizens, is of course impractical in any modern state and is not followed in the declarations and constitutions of the French Revolutionaries.

In many ways Rousseau's political theory was a reversion to Greek ideas and especially to those of Plato. Whereas Locke restricted the functions of the state, Rousseau went back to the notion of an all-embracing political order that provides the only environment in which men can develop their moral capacities. Having himself been born in the city-state of Geneva, Rousseau thought of the small city-state as the ideal form of political community, and his theory of popular sovereignty reflects his view of the ideal rather than the actual. Like Plato again, Rousseau was a collectivist in his ideas about the moral character of man and society. He disagreed with the individualist notion of Locke that men in a state of nature would have natural rights and moral relationships. All rights are social for Rousseau. It is only in society, and this means political society, that men can realize their moral—indeed, their human—qualities. One can become a man only by being a citizen.

Rousseau's theory of popular sovereignty and the general will is a novel method of dealing with the relation between liberty and authority. Hobbes limited liberty to make way for authority, and Locke limited authority to leave scope for liberty. Rousseau believed that the two can go together. He regarded freedom as essential to morality and so as essential to being a man; and, since citizenship of a state is equally essential to being a man, it must be possible to retain freedom when accepting political authority. Now if the authority is exercised by all citizens, who legislate for themselves, then submission to the laws is following decisions that are as much one's own decisions as they are those of other people; it is doing what one has willed to do oneself, and so it is a voluntary or free act. This would be clear and

simple enough if all decisions were unanimous, but if decisions were unanimous there would be no problems of political order at all. Rousseau had to argue that a man is free even when the law that he is required to obey is one with which he disagrees. Hence, Rousseau distinguished the "general will" from particular wills. The general will is a moral will, a will that aims at the common good. Rousseau assumed that all men are capable of taking the moral standpoint of aiming at the common good and that if they did so they would reach a unanimous decision. But men are also prone to judge things from the particular standpoint of aiming at their own good, and then of course they are liable to disagree. In an ideal state, laws express the general will. An individual who disagrees with a law must be failing to look at things from the moral standpoint. The general will that the law expresses is what that individual, no less than others, would share if he took the moral, the truly human, view. Further, since moral action depends upon freedom of choice, that freedom is true freedom and shows itself in the general will. Freedom to do as one wants for one's own particular interests is not true freedom. Rousseau thus concluded that obedience to the law is following the true freedom of the general will, and even went so far as to say that an individual who is forced by the state to obey the general will is being "forced to be free."

Rousseau's concept of the general will

Hegel. Such a doctrine, which begins with democratic sentiments, can clearly lead to a more authoritarian attitude, an attitude that was developed by the German philosopher Georg Wilhelm Friedrich Hegel, who took up some of Rousseau's ideas and molded them into a political theory of a very different character. He elaborated Rousseau's concept of freedom, speaking of a universal or rational will instead of a general will but without the notion of popular sovereignty. He admired great leaders such as Napoleon, and he advocated monarchy. Rousseau's popular sovereignty was intended for a city-state, and this was bound to seem unrealistic to Hegel, who took history to be the guide to theory and practice alike. The history of Germany in particular led Hegel to idealize the nation and to venerate the state as the embodiment of national aspiration.

Hegel, like Rousseau, rejected the individualistic notion of freedom as the facility to do as one likes. Complete freedom in that sense is impossible. True freedom is the freedom of moral choice, expressing a universal or rational will. Hegel meant by this that, if a man acts as a moral agent, his aims coincide with those of others; he does not think of his own particular interest but of justice and the public good. Since there is no conflict in aims when people act morally, their aims can be fully satisfied, which implies that they can be completely free. Freedom, therefore, like morality, is possible only in a moral community, the highest form of which is the state. Hegel's concept of freedom does not drown the individual in the collective. His view is that genuine human individuality can exist only within the milieu of civilization. Freedom as he understood it involves conscientiousness, and he was ready to admire reformers such as Socrates or Martin Luther who stood up against conventional views for conscience' sake, but the general tenor of his theory of the rational will made him speak as if the requirements of the state must necessarily be right. He had no sympathy with democratic individualism and would build the state on groups such as classes, estates, and corporations. Hegel thought of the state as being like an organism whose parts can have no life or significance except in relation to the whole. The overall effect of Hegel's political theory, therefore, is to subordinate the individual to the state.

Restriction between the state and civil society

Hegel drew a sharp distinction between the state and "civil society." This is partly due to the "dialectical" logic of the whole of his philosophy, the idea that all process contains three elements, thesis, antithesis, and synthesis; in his political theory the state fills the role of synthesis, combining and transcending the values of the family (thesis) and civil society (antithesis). Partly, however, the distinction is intended to show the supposed superiority of Hegel's theory to those of earlier thinkers who had used the expression civil society for the state. By civil society Hegel meant the set of institutions that meet the needs of economic life and regulate people's pursuit of their private affairs. All the mundane business of administering the law and providing public services falls within the sphere of civil society. Civil society is society as conceived by individualists and Utilitarians, a means to the satisfying of private interests. The state, by contrast, is society conceived as a moral order, in which private interest is merged into public. The state is not a means, as civil society is; it is an end—indeed, the ultimate end for social action. It is "the march of God in the world."

Hegel thought of the state as divine, not only for its ethical character but also for its power. In his discussion of international relations and war he glorified power and war as much as justice and was evidently unmoved either morally or intellectually by the prospect of continuing conflict between states. His general philosophy should have led him to envisage a universal political community as the proper embodiment of universal spirit, but he rejected the vision of "perpetual peace" and a league of nations outlined by Immanuel Kant, his predecessor in German philosophy.

The political thought of Hegel stands in striking contrast to the philosophy of individual liberalism initiated by Locke and continued by Utilitarians and radicals, especially in Britain and in France. The Utilitarians rejected the idea of natural rights quite as firmly as did Rousseau and Hegel; the English Utilitarian Jeremy Bentham called it nonsense and insisted that the only rights were those of positive law. The Utilitarians nevertheless regarded society as made up of individuals and the public good as made up of the happiness of individuals. Like Hegel, they posited an identity of private and public interest in the state. The difference is that they regarded the state as a device to produce this identity of interests artificially. The state for liberals is a means to the ends of its members. For Hegel the state is an end in itself. Hegelianism entered into the thought of some liberal thinkers, such as Thomas Hill Green in England and Benedetto Croce in Italy, tempering the somewhat mechanistic and negative concept of the state that the Utilitarians inherited from Hobbes and Locke. Generally speaking, however, the tempering of liberalism has come more from radical and Socialist traditions, which add the aims of equality and fraternity to that of liberty.

Influence of Hegel's ideas on the state

Hegel's theory of the state both reflected and strengthened the rising nationalism of the 19th century. It has had a profound influence on later political thought and on practical political movements. Hegel's dialectical and historical approach was adopted by Karl Marx but with economic Materialism replacing Idealism and with a strongly internationalist instead of a nationalist outlook. Nationalism became conservative and reactionary, culminating in the Fascism and National Socialism that led to World War II (1939–45). Prussian militarism in the 19th century owed the whole of its ideology to Hegel, as may be seen especially in the writings of its spokesman, Heinrich von Treitschke. Modern Fascism and Nazism have a more mixed ancestry, but the influence of Hegel is clear in their exalting of the nation and the nation-state, of the heroic leader, military power, and war.

Marxist analyses. While Hegel saw the state as the matrix of justice and ethical value, Communists and Anarchists see it as a mechanism of exploitation or as an institution for perverting man's natural goodness or capacity for development. This is partly because they take a different view of the history of society but partly also because they assign a different meaning to the word state. Hegel saw the state as the whole society unified by ideas of moral community; it was, of course, organized, but the essence of the state for Hegel lay in its moral spirit, not in its institutions. In the political theories of Marxism–Leninism and of Anarchism the state figures as the institutions of government.

Karl Marx said that in Hegel's philosophy "dialectic stands on its head" and that he himself turned it the right way up by substituting Materialism for Idealism. For Marx the apparatus of the state forms part of the superstructure of society, along with religion, morality, and culture gen-

erally. This superstructure depends on and is less "real" than the substructure constituted by economic facts. Hegel had held that "the real is the rational," meaning that the highest value attaches to the realm of the mind. Marx's Materialism attributed the greatest importance to the work that men do and their economic relations. Marx thus inverted Hegel's order of value as between the state and the mundane life of civil society. But he went further than that. In Hegel's philosophy the state is necessary for true morality and the realization of freedom. For Marx the state is an instrument of class rule and robs men of their freedom. It consists of institutions of oppression in the form of law, police, prisons, an army, all designed to maintain the existing economic system. Marx argued that capitalism, like earlier forms of social organization, develops internal contradictions and so nurtures the seeds of its overthrow by revolution. In this revolution the workers will seize the means of production and make its ownership public. They will thus acquire economic and also political power.

The "withering away" of the state in Marxist theory

The Marxist theory of what happens to the state after the Socialist revolution was explained by Marx's friend and collaborator Friedrich Engels and then elaborated by Lenin. The Socialist stage of society that follows capitalism is not yet a classless society, because the bourgeoisie, although stripped of its possessions and power, still retains its former outlook. Socialist society therefore still has the apparatus of the state in the form of "the dictatorship of the proletariat." In time this Socialist society will be followed by a truly Communist stage of society, in which at last men will have complete freedom, because there will be no need for the enforcement of law. The members of a Communist society are classless and free, because all voluntarily seek the good of the whole society, instead of the interest of a limited section. Engels said that the state will "wither away" and that in a Communist society "the government of persons is replaced by the administration of things and the direction of the process of production." He meant that economic life will still be organized, and there will still be rules as under a system of law, but there will be no need for institutions of enforcement, because everyone will obey the rules willingly and readily, knowing that they are for the common benefit and are not a means of exploiting one group for the private interests of another. But the withering away of the state, Lenin insisted, cannot happen without a Socialist revolution to place the means of production under public ownership. The state cannot just be abolished; there must first be a period of the dictatorship of the proletariat; that is, of the continuance of forceful political rule.

Anarchist analyses. Lenin was here arguing against the Anarchists of his time. Some Russian Anarchists, such as Peter Kropotkin, believed that the state would gradually disappear with the growth of voluntary cooperation, even in a capitalist society. Others, such as Mikhail Bakunin, agreed with the Marxists that revolution was needed but thought that the state could be abolished immediately. Others again, such as Leo Tolstoy, advocated a merely passive resistance to state authority.

Kropotkin's view of the state

Kropotkin is the clearest exponent of Anarchism—the doctrine that government is unnecessary and undesirable. He noted that people engaged in art, science, industry, and commerce can go about their business successfully by voluntary agreement. Indeed, he argued, the freedom that they enjoy through the absence of state control means that they show greater energy.

According to Anarchism, the state is both harmful and unnecessary. It is harmful because power corrupts those who govern and divides those who are governed. It is unnecessary because it is inefficient. Kropotkin argued that the state is less successful than voluntary effort in providing services or security. Even in the conduct of war, Kropotkin asserted, voluntary groups are more effective than standing armies. As for internal security, the state, far from protecting citizens, creates criminals by its unjust economic system and hardens them by its prisons. Kropotkin thought that private associations would naturally cooperate together to do all that the state does and to do it better.

The state as a legal order. Any theory of the state must take account of the relation between government and law. It has been left to the Austrian legal philosopher Hans Kelsen to put forward the view that the state is nothing other than a centralized legal order. Kelsen's "pure theory of law" is a landmark in the history of jurisprudence. His theory of the state, though less influential, is an original and noteworthy development in the history of political philosophy.

Kelsen compared the notion of the state with that of a legal person. A legal person is not the same as a natural person, as may be seen from the fact that a corporation is a legal person. On the other hand, there is no need to postulate a special metaphysical entity over and above natural persons. A legal person, in Kelsen's view, is the set of legal rights and obligations that are commonly said to belong to that person; the person consists of the rights and obligations. So, too, with the state. The state cannot simply be identified with society, the set of natural persons whose behaviour is the subject matter of sociology. Yet, on the other hand, it is idle to postulate an additional metaphysical entity over and above the members of the state. The state, in Kelsen's view, is the centralized legal order regulating the conduct of the inhabitants of the territory; it is the set of legal norms and the acts performed by natural persons (legislators, judges, officials) in their capacity as "organs of the state."

Kelsen rejected the doctrine of state sovereignty on the ground that it is illogical. It implies, he said, that each state or legal order recognizes only itself as authoritative, and he compared this position with solipsism, the philosophical theory that each person has certain knowledge only of his own existence and experience. Kelsen maintained that supreme authority must be accorded to international law.

The welfare state. The concept of the welfare state, applied to a number of liberal democratic states since World War II, involves a widening of the functions assigned to the state by individual liberalism. For Locke and the early Utilitarians the business of the state is confined to order and security; the state should protect each man's freedom to act as he chooses, provided he does not infringe upon the like freedom of others. This notion highlights liberty but does nothing to help those who cannot help themselves. Aid for the needy is left to the voluntary action of religious and philanthropic institutions and is not thought to be the business of the state. In the welfare state, however, it is. The state has the job of ensuring the means of subsistence and a basic minimum of welfare for all its members.

This change means a move toward the doctrine of collective responsibility held by Socialists and Communists. Something of the new attitude is found in the 19th century in the social theory of the English philosopher John Stuart Mill, who saw virtue in parts of Socialist doctrine despite his being the most influential spokesman of the Utilitarian philosophy and the most powerful advocate of individual liberty. The philosophy of the welfare state was, however, first clearly enunciated in Great Britain in the Beveridge report of 1942, a document of proposals for a unified system of social insurance presented to the government by Sir William (later Lord) Beveridge. Beveridge may have been unconsciously influenced by the ideas of Mill, and, like Mill, he was essentially a radical liberal, not a Socialist.

Enlarged role of the welfare state

The philosophy of the welfare state, as enunciated by Beveridge, involves a more general view of state responsibility. The basic necessities of life are defined so as to include a decent measure of education, health, and housing. Schemes of social insurance redistribute wealth in the interests of social justice. A welfare state has the positive function of promoting justice and welfare as well as the old negative function of protecting liberty and security.

The old function still remains, however, and the political philosophy of the welfare state still regards individual liberty and individual responsibility as fundamental. Individuals continue to be themselves responsible for rising above the level of welfare provided by the state, and that level is deliberately kept to a basic minimum of bare sub-

sistence in order to give people an incentive for effort to go beyond it. The welfare state draws on the ideas of socialistic collectivism but retains the essence of individual liberalism.

THE STATE IN CONTEMPORARY POLITICS

A political philosophy usually reflects and reinforces a tendency that is making itself felt in practice. In the 16th and the 17th centuries the notion of sovereignty came to the fore in the writings of Machiavelli, Bodin, and Hobbes because a new form of political structure was emerging and was needed as a result of the breakdown of the medieval order of society. By the middle of the 20th century, however, state sovereignty was beginning to be something of an impediment for the development of international order, and this is reflected in the theory of Kelsen. The liberal and democratic ideas of Locke and Rousseau are indicative of the rising tide of revolt against monarchical rule. Hegel mirrors and strengthens the movement toward national unity in Germany. Marxism brings into sharp focus the social effects of the Industrial Revolution. The philosophy of the welfare state epitomizes changes that have been growing up for some time in the character of liberal democratic society.

The state and the individual. The introduction to this article began by saying that "the state" has both a wider meaning, the politically organized society as a whole, and a narrower one, the institutions of government. The foregoing survey of leading philosophical theories has indicated why it was natural for the narrower meaning to arise. The authority of the state must be exercised through particular offices, and these come to be identified in thought with the state. Thus, Marxism and Anarchism see the state as the institutions for enforcement of rules —that is to say, the institutions of government—and as something separable from the society as a whole.

One result of this tendency is that some kinds of political conflict come to be viewed in terms of a conflict between the state and the individual, for an individual who is against a law or a governmental policy and is prepared to demonstrate or to use violence in support of his beliefs thinks of his conflict with the powers that be as a conflict with the state. Not all political disputes take on this character—political conflict between rival parties or between rival factions within a party, for example, does not take this form. But in a totalitarian state, in which only one party is permitted, views opposed to the policy of that party are necessarily and permanently opposed to the government and are therefore regarded as being opposed to the state. In a multiparty state, an opposition party that has a reasonable opportunity of gaining power in the future will not see its conflict with the government as a conflict with the state. It is different with a minority group that has no chance of gaining political power, such as a racial minority that is socially depressed or a group of young people who oppose governmental policy on matters of war and peace. Such a group is often liable to think of its opposition to the government in terms of a contest with the state.

Ameliorative influence of the separation of powers

The separation of powers in democratic states helps to reduce this polarization of state and individual. Although an individual who takes a minority view is liable to find himself at odds with the legislature, executive, and judiciary alike, the courts may nevertheless protect such an individual from the excessive exercise of power by the executive.

A further situation in which political conflict may polarize into conflict between the state and the individual is when "reason of state" is invoked as a justification of executive action. The phrase reason of state (a translation of the French *raison d'état*) means a purely political reason for justifying the action of a government. It was used especially when normal moral justifications were conspicuous by their absence, when the action involved some patent injustice, dishonesty, or concealment. Nowadays the expression national interest is more commonly used. In time of war or other emergency the national interest is invoked to justify forms of discrimination that would otherwise be unfair, as in the selection of conscripts for the armed forces or to excuse departures from normal legal procedure, such as detention without trial of persons suspected of being potentially dangerous. Even in time of peace, government departments are liable to appeal to the national interest as a justification for actions that would not otherwise be tolerated, such as a refusal to provide information to a court or to a legislature. There is always a temptation for the executive to find an easy excuse for the exercise of power.

Here again the division of powers acts as a check on the executive and as a protection for the rights and freedoms of the individual. The courts may require the head of a government department to show "good cause" why someone should be detained or why some information should be withheld. The executive's use of its power can be restrained by other checks also, such as criticism or questioning in the legislature or examination of a grievance by an ombudsman.

Attempts to counter the growth of governmental control are also seen in the tendency toward decentralization achieved by increasing the scope of local and regional government. Decentralization does not, of course, limit state power, but it diffuses the exercise of that power and so gives more opportunity for the voice of individual citizens and representative groups to be heard and heeded.

The state in international society. Despite the doctrine of sovereignty, states have never been entirely free in law to do whatever they like. When the international society of medieval Christendom was replaced by sovereign states, the modern system of international law emerged at the same time as a restraining factor. The new political entities, which acknowledged no superior authority, needed some agreed principles for regulating their relations with each other both in peace and in war. Much of international law is rather like the customary law of a tribal society. States that break the rules cannot be coerced by the power of an international police force. They nevertheless have obligations under international law, and they conform to those obligations to a certain extent, if only because general conformity is to the advantage of all states.

Some theorists, drawing upon Hegel, consider that the obligations of states under international law arise from the "auto-limitation of sovereignty." They mean by this that the sovereignty of a state can be limited only by the decision of that state itself. This view implies that a state is free at any time to revoke the limitation and to be no longer bound by international law. On the basis of the actual behaviour of states, however, it would appear that only part of international law is treated as subject to consent. A state gives its consent to international obligations when its government signs a treaty or ratifies a convention or agrees to accept the jurisdiction of the International Court of Justice. But there are many rules of international law that do not arise from treaties or conventions; they reflect customary practices of international society or else general principles recognized by all civilized peoples. Nowadays there are also the precedents of important decisions by the International Court of Justice. Newly formed states are not given the option of accepting or rejecting the existing body of rules of international customary law. They are assumed as a matter of course to be bound by these rules, the tacit acceptance of which is a condition of their diplomatic recognition by other states.

International protection of human rights

Public international law and the United Nations are mainly concerned with the relations between states. A state's treatment of its own citizens was until recently held to be entirely its own business. The peace treaties concluded at the end of World War I, however, contained clauses to guarantee protection for certain minorities, and in 1948 the United Nations adopted the Universal Declaration of Human Rights. This document does not lay any legal obligations upon member states, but it has inspired several conventions, which have obligatory force, on particular topics such as the status of refugees, the political rights of women, and the removal of discrimination in employment or education. In addition, a num-

ber of European states have adopted a more general European Convention for the Protection of Human Rights and Fundamental Freedoms (but restricted to civil and political rights, while the Universal Declaration includes economic, social, and cultural rights also). The most striking development in this regard is the setting up of the European Commission of Human Rights and the European Court of Human Rights, which have the power to investigate alleged breaches of the convention. Under this procedure a complaint can be made against a state not only by another state but also by an aggrieved citizen—a remarkable, if as yet undeveloped, modification of the historical supremacy of the state over its members.

BIBLIOGRAPHY. Virtually all books on general political theory deal with the state as one of the main topics. Among recent works, ALESSANDRO PASSERIN D'ENTREVES, *La dottrina dello stato* (1962; Eng. trans., *The Notion of the State*, 1967) is especially useful in combining analysis with the history of ideas. JOHN D. MABBOTT, *The State and the Citizen*, 2nd ed. (1967), is a lucid philosophical introduction. ERNEST BARKER, *Principles of Social and Political Theory* (1951), though a more general treatise, is largely concerned with the state. ALEXANDER DUNLOP LINDSAY, *The Modern Democratic State*, vol. 1 (the only volume published, 1943, reprinted 1959), relates political philosophy to political, social, and economic history. HUGO KRABBE, *Die Moderne Staatsidee* (1915; Eng. trans., *The Modern Idea of the State*, 1922, reprinted 1930), stresses a juristic approach; ROBERT M. MACIVER, *The Modern State* (1926), is more sociological; EDWARD JENKS, *The State and the Nation*, rev. ed. (1935), is anthropological and historical. HAROLD J. LASKI, *The State in Theory and Practice* (1935), gives a Socialist view. Views of the contemporary relation between state and individual are given by WILLIAM ERNEST HOCKING, *Man and the State* (1926, reprinted 1968); and by JACQUES MARITAIN, *Man and the State* (1951). The changing position of the state in relation to international law is described in JAMES LESLIE BRIERLY, *The Law of Nations*, 6th ed. revised by HUMPHREY WALDOCK (1963); and in IAN BROWNLIE, *Principles of Public International Law* (1966); and FREDERICK L. SCHUMAN, *International Politics*, 7th ed. (1968).

(D.D.R.)

Statistics

Statistics is the art and science of gathering, analyzing, and making inferences from data. Originally associated with numbers gathered for governments, the subject now includes large bodies of method and theory.

Applied fields have stimulated extensive theoretical and practical work. For example, trying to improve the accuracy of physical measurements led the 19th-century German mathematician Carl Friedrich Gauss to develop the method of least squares (a method of analyzing experimental data; see below), which today has evolved to a flexible general theory that is applied and adapted to nearly every quantitative area. This work also contained the seeds of what the 20th-century British statistician Sir Ronald Aylmer Fisher later developed into the design of experiments, a branch of statistics widely used in agriculture for the improvement of crop yield and used to improve the precision of the results obtained in medical, biological, and industrial experimentation.

The ideas of effective design for data gathering are also basic to the construction of sample surveys, a branch of statistics most popularly known for its contributions to public opinion polling, to pre-election forecasting, and to market research. Perhaps even more significant is this branch's service to government through, for example, estimates of amount of unemployment and cost of living, which have become indispensable in attempts to regulate the economy. Sample survey methods have found uses in accounting, inventory control, and other areas.

For modern manufacturing, methods of controlling the quality of manufactured product have been essential. The statistical foundations of sequential analysis were provided by Abraham Wald in 1947 in order to handle problems in quality control.

In psychology and education, the statistical theory of tests and measurements has been extensively developed to compare achievement by individuals from diverse places and backgrounds.

THE BASIC PRINCIPLES OF STATISTICAL INFERENCE

Statistical experiments. The basic principles of statistical inference involve the notions of experiment, outcome, sample space, event, and probability. The idea of an experiment includes such examples as tossing a coin, measuring a length, or asking a person which party he or she will vote for at an election. Such experiments have a number of possible outcomes, perhaps infinitely many. When an ordinary die is thrown there are six possible outcomes, corresponding to the numbers 1, 2, 3, 4, 5, and 6. In statistics this collection of all possible outcomes is called a sample space. For the die the sample space can be written as the set {1, 2, 3, 4, 5, 6}. Any single outcome of an experiment, such as the throwing of a 6, is called a point. Certain specified sets of outcomes are called events. The throwing of an even number on a die would be an event, denoted by {2, 4, 6}.

Die throwing and coin tossing

Sets of events are frequently studied. For example, in tossing a coin twice, one possible outcome is "heads first time, tails second time," which is abbreviated to (H, T). Four possible outcomes make up the sample space—(H, T), (H, H), (T, H), and (T, T). If the experimenter asks only whether he gets the same result on both throws, the natural events chosen would be {(H, H), (T, T)} and {(H, T), (T, H)}.

In statistics the results of a large number of experiments form the data used. Sometimes it is assumed that each outcome is independent of any preceding one, as the number thrown on a die is independent of any preceding throws. It is also assumed that there is an underlying probability mechanism determining the nature of the data, as the probability of a coin's coming up heads determines the outcomes of experiments in which a coin is thrown twice. If the underlying probability mechanism is known, then the probability that certain outcomes occur can be calculated. This is the concern of probability theory (see PROBABILITY, THEORY OF).

Statistics is concerned with the application of probability theory to data. If only partial information is available —for example, if the underlying mechanism is one of several—it may be possible to use the data to infer which mechanism generated it and to associate probabilities or degrees of confidence with the various mechanisms. It is possible to perform the experiment a number of times and obtain a sequence of outcomes. Dividing the number of times an outcome of an experiment falls in event A by the number of experiments gives an estimate of the probability of event A. As the number of experiments grows large, the estimate becomes more likely to be close to the true probability $\mathbf{P}(A)$. In mathematical terms, if $x_1, x_2, \cdots$ is a sequence of outcomes and A an event, and $N(x_1, x_2, \cdots, x_n; A)$ denotes the number of times the outcome falls in A during the first n experiments, then as n tends to infinity, the chance that $N(x_1, x_2, \cdots, x_n; A)/n$ is close to $\mathbf{P}(A)$ tends to certainty.

Although this fact, known as the law of large numbers, tell the experimenter that he will arrive at the right answer eventually, it does not tell him how close he will be to the right answer after having performed the experiment a certain number of times. This is where statistical methods come in. Under certain conditions statistical methods give an estimate of the error that might possibly be made by performing the experiment a finite number of times.

An experiment is described completely by its sample space S, which is the set of possible outcomes, by a family B of events, and by a probability assignment $\mathbf{P}$ that governs the outcome of the experiment. The independence of two events means that the outcome of one does not imply anything about the outcome of the other. As a practical matter, separation of events in space and time often ensures independence. The concept of statistical or probabilistic independence is intended to extend the common meaning of independence to chance situations in which the occurrence of one event does not change the probability of occurrence of another event.

When events are independent, the probability that they jointly occur is the product of their probabilities: If the events are $A_1, \cdots, A_n$ and the associated probabilities

are $\mathbf{P}(A_i)$, $i = 1, \cdots, n$, the probability that all occur is $\mathbf{P}(A_1) \cdots \mathbf{P}(A_n)$.

When a die is tossed twice, for example, the sample space can be written as the set $\{(1, 1), (1, 2), \cdots, (1, 6), (2, 1), \cdots, (2, 6), \cdots, (6, 6)\}$; each of the 36 points in this sample space is composed of an outcome on the first toss and an outcome on the second. Thus, (4, 3) denotes a 4 on the first toss and a 3 on the second. The events A_1 and A_2 are independent and the probability of "A_1 and A_2" may be computed as the product of the individual probabilities (see Box, calculation 1), yielding a probability of ⅓.

These ideas of independence apply to statistical investigations in which the data are generated by repetitions of the same elementary operation, each carried out independently of the others.

Random variables

To develop models for numerical data, a real number or a vector with real components is associated with each outcome in the sample space. The function mapping the outcome to the number or vector is called a random variable. Capital letters are often used to designate the random variable itself and lower-case letters for an unspecified value of a random variable. Thus $\mathbf{P}(X = x)$ represents the probability that the random variable X takes the value x. It is often convenient to forget the functional origin of the idea of random variable and regard the set of values of X as the basic sample space.

For example, when a coin is tossed, the outcome may be heads or tails, and heads may be associated with $X = 1$ and tails with $X = 0$. X, the number of heads, is called a random variable and its values are 0 and 1 with equal probabilities of ½ (see 2).

Using random variables to describe outcomes of experiments gives systematic ways to describe the assignment of probability to sets. In practice, the sample space of the random variable is either a discrete set of values (as in tossing a die) or a continuous set (as in measuring a length). In both cases the sample space consists of a set on the real line. In the discrete case, it may be convenient to give a probability for each point x in the sample space S. Thus for the random variable X a probability mass function f is defined such that its value at x is a probability (see 3). Then the probability of an event A may be computed by summing $f(x)$ for all values of x in A (see 4). For all x, $f(x) \geq 0$ and $\mathbf{P}(S) = 1$.

For example, if p is the probability of success on a single trial, if the outcomes of trials are independent, and if the random variable X is the number of successes in n trials, then (as mentioned in PROBABILITY, THEORY OF) the value of f at x is a term in a binomial expansion (see 5). The probability that the experiment produces all successes or all failures is $p^n + (1 - p)^n$. The distribution of probability represented by the values of $f(x)$ is called the binomial distribution. A more complete representation would use $f(x|n, p)$ (read "f of x given n and p") rather than $f(x)$ so as to specify the parameters n and p of the distribution.

When probability is smoothly distributed over continuous sample spaces, a difficulty in representing the probabilities is that each point has probability zero, yet the total probability must be one: a line segment consists of points, each of zero length, and yet it has a nonzero length. Developing probability through measures on Borel sets resolves such difficulties (see ANALYSIS, REAL).

Distribution functions and their properties. For the continuous random variable X with values $x \in S$, a cumulative distribution function F is defined so that its value at x is the probability (see 6) for the random variable to be less than or equal to x. $F(-\infty)$ and $F(+\infty)$ are also defined by: $F(-\infty) = 0$ and $F(+\infty) = 1$ (the symbol ∞ denotes infinity). The probability in an interval, say $a < x \leq b$, is given by the difference in the values of F evaluated at the endpoints of the interval (see 7).

If F has a derivative $dF(x)/dx = f(x)$ for almost all x, then f is called the probability density function of the random variable X, or more briefly the density function. This brings the integral calculus into play because now, from the fundamental theorem of calculus, the definite integral of f is a probability (see 8). For example, for the uniform distribution on the interval $a < x \leq b$, the density and cumulative are given by specific functions (see 9) that are constant and of constant slope, respectively, over the interval.

The approach via the cumulative distribution function F makes it possible to unify mathematically discrete and continuous distributions as well as mixed ones by employing the Lebesque-Stieltjes integral.

The problem of finding the value of x such that $F(x) = p$ often arises. Such a value of x is called the pth quantile, or $100p$th percentile, of the distribution; the notation x_p is used. The median of the distribution is $x_{0.5}$, and $x_{0.25}$ and $x_{0.75}$ are the first and third quartiles.

When the random variable X has a distribution with density f, the mean of the distribution, if it exists, is given by the integral (see 10) over the whole real line of the product $xf(x)$. Its value μ is also called the first moment of the distribution. More generally, if it exists, the integral (see 11) of the product $x^r f(x)$ is the rth raw moment of the distribution. Moments about the mean are also used, and the rth moment about the mean is obtained from a similar integral (see 12) involving the rth power of $(x - \mu)$. In the discrete case the integrations are replaced by summations over $x \in S$.

Variance

The second moment about the mean is much used in both statistics and probability. It is called the variance of the distribution, or the variance of the random variable X, abbreviated var X, or σ_X^2, or when the name of the random variable is understood, σ^2. Its square root, σ, is the standard deviation.

In addition to moments defined as average values of the functions X^r and $(X - \mu)^r$, the average values of other functions of random variables frequently arise. If h is a function of X, then when it exists, the expected value of $h(X)$ is defined to be the integral over the whole real line (see 13) of the product of $h(x)$ and the probability density function. Therefore, the mean is $\mathbf{E}(X)$, and $\sigma^2 =$

(1) $$\begin{cases} A_1 = \{\text{first toss even}\}, \\ A_2 = \{\text{second toss 4 or less}\} \\ \mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1)\mathbf{P}(A_2) = \frac{3}{6} \times \frac{4}{6} = \frac{1}{3} \end{cases}$$

(2) $$\mathbf{P}(X = 1) = \mathbf{P}(X = 0) = \tfrac{1}{2}$$

(3) $$f(x) = \mathbf{P}(X = x), \quad x \in S$$

(4) $$\mathbf{P}(A) = \sum_{x \in A} f(x)$$

(5) $$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \qquad 0 \leq x \leq n$$

(6) $$F(x) = \mathbf{P}(X \leq x), \qquad -\infty < x < +\infty$$

(7) $$F(b) - F(a) = \mathbf{P}(a < X \leq b)$$

(8) $$\int_a^b f(x)\,dx = F(b) - F(a) = \mathbf{P}(a < X \leq b)$$

(9)

	$x \leq a$	$a < x \leq b$	$b < x$
$f(x)$	0	$1/(b-a)$	0
$F(x)$	0	$(x-a)/(b-a)$	1

(10) $$\mu = \int_{-\infty}^{+\infty} xf(x)\,dx$$

(11) $$\nu_r = \int_{-\infty}^{+\infty} x^r f(x)\,dx$$

(12) $$\mu_r = \int_{-\infty}^{+\infty} (x-\mu)^r f(x)\,dx$$

$\mathbf{E}[(X-\mu)^2]$. It should be noted that $\mathbf{E}$ is an operator, not a notation for a function.

For example, the expected value of the number of successes, X, for the binomial distribution is the product of the number of trials n and the constant probability p of success in each trial (see 14), derived by evaluating a binomial summation. The variance of X is $(1-p)$ times the mean (see 15).

For the normal distribution (see 16a), upon integration the mean μ of the random variable X is θ_1 (see 17), and the variance σ^2 is θ_2 (see 18). Consequently, the parameters of the normal distribution are said to be μ and σ^2. These results (see 16b) make clear that the normal density decreases exponentially, the exponent being proportional to the square of $(x-\mu)/\sigma$, the distance from x to the mean measured in standard deviation units. The coefficient $1/\sqrt{2\pi\sigma}$ assumes that the total density integrates to unity so that the total probability is 1. [In (16a), (16b), and (16c), π (pi) equals 3.14159 $\cdots$.]

If $Y = aX + b$, in which a and b are constants, and X any random variable with mean μ and variance σ^2, then the mean of Y, written μ_Y, bears the same linear relation to the mean of X that Y bears to X, and the variance of Y is the product of a^2 and the variance of X (see 19).

From any random variable X with mean μ and variance σ^2 a new random variable $Y = (X-\mu)/\sigma$ with mean 0 and variance 1 can be constructed. Such a Y is called a standardized random variable, and, if X is normally distributed, it is said to have the standard normal distribution (see 16c).

To generalize the concept of distribution function to several random variables $X_1, \cdots, X_k$ in a k-dimensional Euclidean space, the joint cumulative distribution function F is used. By definition the joint cumulative distribution function (see 20) is the probability of the common occurrence of k events of the type $(x_k \leq x_k)$. If any x_i is replaced by $-\infty$, then F takes the value zero. When all x_i are replaced by $+\infty$, $F = 1$. If F is differentiable with respect to all the variables, then there is a joint probability density function f such that F is the indefinite multiple integral (see 21) of f. If A is a set in the sample space and $x = (x_1, \cdots, x_k)$, then the probability that x belongs to the set A is the multiple integral (see 22) of the density extended over A. Formally the same equality is expressible in terms of a Stieltjes integral (see 23). When integration is performed over all variables but x_i, the cumulative gives $F(+\infty, \cdots, +\infty, x_i, +\infty, \cdots, +\infty)$, which represents the marginal cumulative distribution of X_i alone. More generally, when exactly r $(1 \leq r < n)$ coordinates are set at $+\infty$, the $(k-r)$th-order marginal distribution of the remaining $k-r$ random variables is obtained.

When the X_i are independent with $F_i(x_i) = \mathbf{P}(X_i \leq x_i)$, $i = 1, \cdots, k$, and have probability density functions $f_i(x_i)$, then the joint cumulative can be represented as the product of the separate distribution functions (see 24), and the joint density as the product of the separate densities (see 25). To represent the distribution of repeated measurements, X_i is allowed to represent the ith measurement, and the f_i is taken to be equal to f_0, so that the products above take a particularly simple form, expressible as a product extended over k index values (see 26) of $f_0(x_i)$.

If two random variables X and Y have a joint density f and if their means are μ_X and μ_Y, the covariance between X and Y (if it exists) is defined as an integral of the product of the differences $(x-\mu_X)$ and $(y-\mu_Y)$, weighted with the joint density (see 27).

Sums of random variables have special interest for statisticians. If X_i has mean μ_i and variance $\sigma_i^2 = \sigma_{ii}$ and if cov $(X_i, X_j) = \sigma_{ij}$, then if $Y = \Sigma X_i$, the mean of Y is the sum of the means of the X_i (see 28) and the variance of Y is the double sum over σ_{ij} (see 29). (In the summation defining Y the sum is from 1 to n.) In other words, the mean of a sum is the sum of the means, and the variance of a sum is the sum of the variances plus twice the covariances. The correlation coefficient between variables X_i and X_j is defined as the ratio of σ_{ij} to the product $\sigma_i\sigma_j$ (see 30). When all pairs of variables have $\rho_{ij} = 0$, then $\sigma_Y^2 = \sum_{i=1}^{n} \sigma_i^2$. Independence between all the random variables guarantees $\rho_{ij} = 0$, but having all $\rho_{ij} = 0$, $i \neq j$, does not guarantee independence.

The correlation coefficient

An important exception is the joint multivariate nor-

(13) $$\mathbf{E}(h(X)) = \int_{-\infty}^{+\infty} h(x) f(x)\, dx$$

(14) $$\mu = \mathbf{E}(X) = \sum_{x=0}^{n} x f(x|n, p) = np$$

(15) $$\sigma_X^2 = np(1-p)$$

(16)
$$\begin{cases} \text{a. } \dfrac{1}{\sqrt{2\pi\theta_2}} \exp\left[-\dfrac{1}{2}\dfrac{(x-\theta_1)^2}{\theta_2}\right] \\ \text{b. } \dfrac{1}{\sqrt{2\pi\sigma}} \exp\left[-\dfrac{1}{2}\left(\dfrac{x-\mu}{\sigma}\right)^2\right] \\ \text{c. } \dfrac{1}{\sqrt{2\pi}} \exp\left(-\dfrac{1}{2}y^2\right) \end{cases}$$

(17) $$\mu = \mathbf{E}(X) = \int_{-\infty}^{+\infty} x f(x|\theta_1, \theta_2)\, dx = \theta_1$$

(18) $$\sigma^2 = \mathbf{E}[(X-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x|\theta_1, \theta_2)\, dx = \theta_2$$

(19) $$\mu_Y = a\mu_X + b, \qquad \sigma_Y^2 = a^2\sigma_X^2, \qquad \sigma_Y = |a|\sigma_X$$

(20) $$F(x_1, \cdots, x_k) = \mathbf{P}(X_1 \leq x_1, \cdots, X_k \leq x_k)$$

(21) $$F(x_1, \cdots, x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f(u_1, \cdots, u_k)\, du_1 \cdots du_k$$

(22) $$\mathbf{P}(x \in A) = \int_A f(u_1, \cdots, u_k) \prod_{i=1}^{k} du_i$$

(23) $$\mathbf{P}(x \in A) = \int_A dF(u_1, \cdots, u_k)$$

(24) $$F(x_1, \cdots, x_k) = F_1(x_1) \cdots F_k(x_k)$$

(25) $$f(x_1, \cdots, x_k) = f_1(x_1) \cdots f_k(x_k)$$

(26) $$f(x_1, \cdots, x_k) = \prod_{i=1}^{k} f_0(x_i)$$

(27) $$\operatorname{cov}(X, Y) = \int_S (x-\mu_X)(y-\mu_Y) f(x, y)\, dx\, dy$$

(28) $$\mu_Y = \mathbf{E}(Y) = \mathbf{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbf{E}(X_i) = \sum_{i=1}^{n} \mu_i$$

(29)
$$\begin{cases} \operatorname{var} Y = \mathbf{E}[(Y-\mu_Y)^2] = \sigma_1^2 + \cdots + \sigma_n^2 + 2\sigma_{12} + \\ \qquad + 2\sigma_{13} + \cdots + 2\sigma_{n-1,n} \\ \qquad = \sum_{j=1}^{n} \sum_{i=1}^{n} \sigma_{ij} \end{cases}$$

(30) $$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}$$

(31)
$$\begin{cases} f(x_1, \cdots, x_n | \mu_i, \sigma_i^2, i = 1, \cdots, n) \\ \qquad = \prod_{i=1}^{n} (2\pi\sigma_i^2)^{-1/2} \exp[-\frac{1}{2}(x_i-\mu_i)^2/\sigma_i^2] \end{cases}$$

mal distribution, the only parameters of which are the means, variances, and covariances of the variables. When $\rho_{ij} = 0$ for all $i, j, i \neq j$, the joint density can be represented by a product (see 31) of normal densities with typical mean μ_i and variance σ_i^2. For the general joint multivariate normal distribution density is proportional to the exponent of a quadratic form (see 32). If $||\sigma_{ij}||$ is the covariance matrix of the n random variables ($||a_{ij}||$ denotes the matrix with elements a_{ij}) and σ^{ij} the elements of its inverse, then the quadratic form is expressible (see 33) in terms of the collection $\{\sigma^{ij}\}$, and the constant factor that multiplies the exponential is expressible (see 34) in terms of $|\sigma^{ij}|$, in which $|a_{ij}|$ is the determinant of the matrix with elements a_{ij}. If $\Sigma = ||\sigma_{ij}||$ and the vector $(x - \mu)$ is defined (see 35), the quadratic form Q can be written more simply in vector and matrix notation, in which T stands for transpose.

Under a variety of conditions, the distribution of sums of random variables tends to a normal distribution as the number of random variables tends to infinity. The statements of conditions leading to these distributions are called central limit theorems. To illustrate, if $X_1, X_2, \cdots$ is a sequence of random variables and $Y_n = X_1 + \cdots + X_n$, and if $\mathbf{E}(Y_n) = \mu_n$ and var Y_n exists, then as n becomes large the sequence of distributions of the standardized random variables Z_n (see 36; each with mean 0 and variance 1) has a limit that is the standard normal distribution (see 37). One set of conditions guaranteeing this limiting distribution is independence of the X_i and identical distributions with finite variances, but other conditions are also available that do not require identical distributions, equal variances, or even independence.

This limiting distribution is also shared by many functions of random variables and is one reason for the importance of the Gaussian or normal distribution in statistical theory. In applications, sums of even a few random variables of the types ordinarily met have distributions the central portions of which can be well approximated by normal distributions.

If $Z_1, \cdots, Z_k$ are independent standard normal random variables, the sum of their squares (see 38) forms a new random variable, χ_k^2; its distribution is called the chi-square distribution with k degrees of freedom.

If Z, a standard normal random variable, and χ_k^2 a chi-square random variable with k degrees of freedom, are independent, then the new random variable, written t_k, which is proportional to Z divided by the square root of the chi-square (see 39), has the distribution known as Student's t distribution with k degrees of freedom. Student was the scientific pseudonym of William Sealy Gosset, the statistician and chemist who discovered the form of the distribution by a combination of mathematical work and empirical work with random numbers, an early application of what is now called the Monte Carlo method.

The chi-square and t distributions are frequently used in interval estimation and in testing hypotheses.

(32) $$\begin{cases} f(x_1, \cdots, x_n | \mu_i, \sigma_i^2, \sigma_{ij}, i, j = 1, \cdots, n) \\ \quad = K \exp(-\frac{1}{2}Q) \end{cases}$$

(33) $$Q = \sum_i \sum_j \sigma^{ij}(x_i - \mu_i)(x_j - \mu_j)$$

(34) $$K = [\,|\sigma^{ij}|/(2\pi)^n\,]^{1/2}$$

(35) $$\begin{cases} (x_1 - \mu_1, \cdots, x_n - \mu_n) = (x - \mu) \\ Q = (x - \mu)^T \Sigma^{-1} (x - \mu) \end{cases}$$

(36) $$Z_n = \frac{Y_n - \mu_n}{\sqrt{\text{var } Y_n}}$$

(37) $$\lim_{n \to \infty} \mathbf{P}(Z_n < z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}x^2}\, dx$$

(38) $$Z_1^2 + \cdots + Z_k^2$$

(39) $$t_k = \frac{Z}{\sqrt{\chi_k^2/k}}$$

(40) $$f(x|\theta) = \theta e^{-\theta x}, \qquad x \geq 0, \theta > 0$$

(41) $$\mu = \mathbf{E}(X) = \theta \int_0^\infty x e^{-\theta x}\, dx = 1/\theta$$

(42) $$L(x_1, \cdots, x_n | \theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

ESTIMATION

Point estimation. For a random sample with values $x_1 \cdots, x_n$ independently drawn from a distribution F the parameter of which is θ, the problem of estimation is to find a function of the x's, $\hat{\theta}$, that comes close to θ. Here "close" is a vague word that can be given many meanings. If the sample is regarded as composed of random variables $X_1, \cdots, X_n$, and the estimator $\hat{\theta}$ as a function of those random variables, then $\hat{\theta}$ is itself a random variable. Some desirable properties for an estimator $\hat{\theta}$ are (A) that its average error be small: indeed, if the average error vanishes, $\mathbf{E}(\hat{\theta} - \theta) = 0$, and the estimator is said to be unbiassed; (B) that the mean square error $\mathbf{E}(\hat{\theta} - \theta)^2$ be small: this measure combines, in a somewhat arbitrary way, variability of $\hat{\theta}$ with its bias $b = \mathbf{E}(\hat{\theta} - \hat{\theta}$, because $\mathbf{E}(\hat{\theta} - \theta)^2 =$ var $\hat{\theta} + b^2$. (C) Another attractive property is small variance. If, among several unbiassed estimators, one has smallest variance, it is known as the minimum-variance unbiassed estimator. Which of several estimators may be more desirable depends, among other things, upon the costs of errors $(\hat{\theta} - \theta)$ of various sizes.

Samples contain information about the parameters of a family of distributions, but different families can profitably use different functions to extract this information. Several methods of estimation offer general techniques for finding estimators for parameters. Three techniques are considered here: the method of moments, the method of maximum likelihood, and Bayesian methods.

Method of moments. The moments of a family of distributions are functions of the parameters of the family. If there are only a few parameters, it may be possible to use the first few moments of the distributions together with the sample moments to get estimates of the parameters. If the observed values of the sample moments are equated with the corresponding population moments, expressed as functions of the parameters, the resulting equations for estimates of the parameters may be solved. Thus the estimators choose that member of the family whose moments match the sample moments.

The method of moments may be used, for example, to estimate the parameter θ of an exponential distribution having probability density function proportional to the exponential of $-\theta x$, the values of x being non-negative (see 40). The population mean is calculated (see 41) to be $1/\theta$. The value of the sample mean is $\bar{x} = (x_1 + \cdots + x_n)/n$, and so the method of moments would derive the estimate of θ from $\bar{x} = 1/\hat{\theta}$ and give $\hat{\theta} = 1/\bar{x}$.

The method of maximum likelihood. A widely used method that has attractive large-sample properties is the method of maximum likelihood. For families of discrete distributions, the probability of getting the particular sample that has been observed can be computed, and then those values of the parameters that maximize the probability can be chosen as estimates. If the sample $x_1, \cdots, x_n$ has been independently chosen from a distribution with probabilities $\mathbf{P}(X = x \mid \theta) = f(x \mid \theta)$, in which θ may be a vector, then that probability, called the likelihood, is defined as the product of n probability mass functions, one evaluated at each sample value (see 42).

For continuous distributions the corresponding approach is again to maximize the likelihood function, which is given above (see 42) except that the right-hand side is the product of the values, $f(x_i \mid \theta)$, of the density function evaluated at the sample points. Although symbolically the same, the likelihood in the discrete situation is a probability, but in the continuous case it is a measure of the density of probability at the sample point. The one-parameter case is considered. Under broad regularity conditions (dealing with the differentiability of f with respect to θ, and the existence of integrals of functions of f and its derivatives), the estimator $\hat{\theta}$, the value of θ that maximizes L for the given sample, is the maximum likelihood estimator of θ. Asymptotically (*i.e.*, as n tends to infinity) $\hat{\theta}$ tends to be normally distributed with asymptotic mean θ and asymptotic variance expressed in a simple formula (see 43) in terms of the reciprocal expectation of the second derivative (with respect to θ) of the log of the density.

For example, a sample of size n from the normal distribution with mean 0 and variance θ (or σ^2) has a likelihood function and its derivative given by the equations that are exponential and algebraic in form, respectively (see 44). If the derivative is set equal to zero and the equation solved for θ, the maximum likelihood estimate of θ is given by the sum of the squares of the x_i divided by n (see 45). Because the second derivative of the log of the density is a simple expression that is linear in the square of the variable (see 46), and $\mathbf{E}(X^2) = \theta$, the asymptotic variance is found to be proportional to the reciprocal of n (see 47).

Some attractive asymptotic properties of the maximum likelihood estimator when the regularity conditions are met are:

A. That it converges in probability to the true value of θ as $n \to \infty$. (Estimators that do this are said to be consistent.)

B. That it is asymptotically unbiassed.

C. That among asymptotically unbiassed estimators, no other has smaller asymptotic variance. (Unbiassed estimators having finite variance at least as small as that of any other estimator are said to be efficient.)

D. The maximum likelihood estimator of a function of θ is the same function of the maximum likelihood estimator of θ.

In the example given above the maximum likelihood estimate of θ (or σ^2) was $\Sigma x_i^2/n$. The maximum likelihood estimate of σ is the square root of this expression (see 48).

The extension of this method to vector random variables and vector parameters is straightforward providing further regularity conditions hold. If a sample point consists of the vector $x_i = (x_{1i}, \cdots, x_{ri})$, and the joint distribution of the r variables has k parameters $\theta = (\theta_1, \cdots, \theta_k)$, then one forms the likelihood function that is formally the same as before but now involving vectors (see 49). To find the relative maximum (assumed unique), the partial derivatives with respect to θ_i are taken, they are set equal to zero, and the resulting equations for the maximum likelihood estimates $\hat{\theta}_i$ are solved. Asymptotically the estimators are jointly normally distributed with mean $(\theta_1, \cdots, \theta_k)$ and covariance matrix $||nB_{pq}||^{-1}$, in which the $\{B_{pq}\}$ are expectations of second partial derivatives of the log of the density (see 50). The vectorial nature of the observations is not central to maximum likelihood estimation, and so the method is illustrated by a one-variable probability distribution with two parameters.

As an example, consider the vector parameter. If the sample is drawn from the normal distribution with mean θ_1 and variance θ_2, so that the normal density is specified (see 51) as before, then it is found that the maximum likelihood estimator of the population mean is the sample mean, and of the population variance is the sample variance (see 52). Furthermore, the covariance matrix is diagonal (see 53), so that asymptotically $\hat{\theta}_1$ and $\hat{\theta}_2$ are independently normally distributed with means (θ_1, θ_2) and variances θ_2/n and $2\theta_2^2/n$, respectively.

For a sample of n independent normally distributed observations, the sample mean $\bar{X}$ is exactly, not just asymptotically, normally distributed with mean θ_1 and variance θ_2 as before (see 54).

Interval estimation. The discussion of estimation has emphasized point estimates for parameters, but it may also be useful to give an interval within which a parameter is likely to lie, together with the chance that the interval includes the parameter. More generally, if the parameter were a vector, a region would be given.

The confidence interval

An important illustration gives the confidence interval, based on a sample size of n, for the unknown mean μ of a normal distribution with known standard deviation σ. The probability is $1 - \alpha$ that the sample mean X falls (see 55) within a specified distance, $|z_{\alpha/2}|\,\sigma/\sqrt{n}$, of μ, in which $z_{\alpha/2}$ is the $100\alpha/2$ percentage point of the standard normal distribution.

The sample mean X is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$, so that a rearrangement of the inequalities gives upper and lower bounds for the parameter μ that involve the random variable $\bar{X}$ (see 56). This statement (in 56) is true $100(1 - \alpha)$ percent of the time. If the random variable $\bar{X}$ is replaced

(43) $$\operatorname{var}\hat{\theta} = \frac{1}{-n\mathbf{E}\left[\dfrac{\partial^2 \log f(x|\theta)}{\partial\theta^2}\right]}$$

(44) $$\begin{cases} L(x_1, \cdots, x_n|\theta) = [2\pi\theta]^{-\frac{n}{2}}\, e^{-\frac{1}{2}\Sigma x_i^2/\theta} \\ \dfrac{\partial \log L}{\partial\theta} = -\dfrac{n}{2\theta} + \dfrac{\Sigma x_i^2}{2\theta^2} \end{cases}$$

(45) $$\hat{\theta} = \sum_{i=1}^{n} x_i^2/n$$

(46) $$\frac{\partial^2 \log f(x|\theta)}{\partial\theta^2} = \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}$$

(47) $$\operatorname{var}\hat{\theta} = \frac{2\theta^2}{n}$$

(48) $$\hat{\sigma} = \sqrt{\hat{\theta}} = \sqrt{\Sigma x_i^2/n}$$

(49) $$L(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

(50) $$B_{pq} = -\mathbf{E}\frac{\partial^2 \log f(x|\theta)}{\partial\theta_p \partial\theta_q}, \qquad p, q = 1, \cdots, n$$

(51) $$\begin{cases} f(x|\theta_1, \theta_2) = (2\pi\theta_2)^{-1/2} \exp[-(x-\theta_1)^2/2\theta_2] \\ -\infty < x < \infty, \qquad -\infty < \theta_1 < \infty, \qquad 0 < \theta_2 < \infty \end{cases}$$

(52) $$\hat{\theta}_1 = \frac{\Sigma x_i}{n} = \bar{x}, \qquad \hat{\theta}_2 = \frac{\Sigma (x_i - \bar{x})^2}{n}$$

(53) $$\begin{Vmatrix} \dfrac{\theta_2}{n} & 0 \\ 0 & \dfrac{2\theta_2^2}{n} \end{Vmatrix}$$

(54) $$\theta_2/n = \sigma_X^2/n$$

(55) $$\mu - |z_{\frac{1}{2}\alpha}|\sigma/\sqrt{n} \leq \bar{X} \leq \mu + |z_{\frac{1}{2}\alpha}|\sigma/\sqrt{n}$$

(56) $$\bar{X} - |z_{\frac{1}{2}\alpha}|\sigma/\sqrt{n} \leq \mu \leq \bar{X} + |z_{\frac{1}{2}\alpha}|\sigma/\sqrt{n}$$

(57) $(\bar{x} - |z_{\frac{1}{2}\alpha}|\sigma/\sqrt{n}, \bar{x} + |z_{\frac{1}{2}\alpha}|\sigma/\sqrt{n})$

(58) $\mathbf{P}(\hat{\theta}_l \leq \theta < \hat{\theta}_u|\theta) = 1 - \alpha$

(59) $\text{var } X' = \dfrac{1}{4n[f(x_{0.5})]^2}$

(60) $c(x|\mu, \sigma^2, \gamma) = (1-\gamma)n(x|\mu, 1) + \gamma n(x|\mu, \sigma^2)$
$\sigma > 1$

(61) $L(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$

(62) $h(x_1, \cdots, x_n, \theta) = g(\theta) \prod_{i=1}^{n} f(x_i|\theta)$

(63) $p(\theta|x_1, \cdots, x_n) = kg(\theta) \prod_{i=1}^{n} f(x_i|\theta)$

by the observed sample mean $\bar{x}$, then the statements (in 56) are true or false. They are true in $100(1 - \alpha)$ percent of the times this method is used. The limits $\bar{x} \pm |z_{\alpha/2}|\sigma/\sqrt{n}$ are called upper and lower $100(1 - \alpha)$ percent confidence limits, and $100(1 - \alpha)$ is called the confidence level or confidence coefficient and is expressed either as a percent or a proportion, $1 - \alpha$. The interval composed of the points (see 57) between the upper and lower confidence limits is called a confidence interval. A commonly used confidence level is 0.95, and then $|z_{\alpha/2}| = 1.96$, or about 2.

If, in general, an upper function $\hat{\theta}_u$ and a lower function $\hat{\theta}_l$ of the sample values can be found—so that the probability for these functions to bound θ is $1 - \alpha$ (see 58), then the values of $\hat{\theta}_u$ and $\hat{\theta}_l$ obtained from the sample provide upper and lower $100(1 - \alpha)$ confidence limits for θ.

In a sample of size n drawn from a normal distribution with unknown mean μ and unknown variance σ^2, having sample mean $\bar{x}$ and variance $s_{\bar{x}}^2$ of the mean given by $\Sigma(x_i - \bar{x})^2/n(n - 1)$, symmetrical $1 - \alpha$ confidence limits on μ are given by $\bar{x} \pm |t_{\alpha/2}|\, s_{\bar{x}}$, in which $t_{\alpha/2}$ is the $\alpha/2$ percentage point of the t distribution with $n - 1$ degrees of freedom.

In some investigations, the researcher continues sampling until his confidence limits have been reduced to a preset size. Finding appropriate stopping rules and setting the limits are problems in sequential estimation.

Interval estimation can be used not only for reporting uncertainty in the value of the parameter, but also for hypothesis testing. If a specific value of the parameter θ, say θ_0, has a reasonable chance of being the true value, and if $1 - \alpha$ confidence limits can be set on θ, then if θ_0 is in the confidence region, it can be accepted as possibly being the true value. Otherwise this hypothesis can be rejected.

For example, in a sample of size $n = 3$ from a normal distribution, $\bar{x} = 2$, $s_{\bar{x}} = 0.5$, symmetrical 95 percent confidence limits are set on μ, and a test applied as to whether $\mu = 0$ is a reasonable value. This problem has two degrees of freedom, and $t_{0.025} = -4.30$. From the above formulas, the confidence limits are $2 \pm 4.30(0.5) = -0.15, +4.15$, and the hypothetical value 0 is in the interval (though barely). Consequently, 0 is not rejected as the value for μ at the 5 percent level of significance.

Robust estimation. In real data, the appearance of observations that seem not to belong with the rest of the measurements (such as 22.19 in the set 10.01, 10.04, 10.03, 22.19) is so commonplace that systematic methods are required to defend the analyses against them. Setting observations aside has always been a touchy step, especially to those who believe that good work never leads to bad data. One reasonable attitude is that wild observations should be set aside and studied separately so as not to destroy the information in the main body of the data. A few standard methods are applied to afford protection against temptation to make a special analysis to reach a foregone conclusion—and against the criticism that this has been done.

One estimator of location is the sample median, the middle-sized observation if the sample size is odd, or the average of the middle two if it is even. The large-sample distribution of the sample median X' is approximately normal with asymptotic mean $x_{0.5}$, the 50th percentile point of the distribution, and asymptotic variance expressed in terms of the density function evaluated at $x_{0.5}$ (see 59). When f is a normal density function, the population mean and median are both μ. The sample mean $\bar{X}$ is exactly, and the sample median X' approximately, normally distributed. As n grows large, $\text{var } \bar{X} \approx \frac{2}{\pi} \text{var } X'$, and X' is said to have asymptotic efficiency $2/\pi \approx 0.63$ compared with $\bar{X}$. Thus it takes a sample of 100 using the median to get the accuracy of a sample of 63 with the mean. If wild observations contaminate a distribution, however, the efficiency soon favours the median.

For example, a density function c formed as the weighted sum of two normal densities is expressed in terms of a function n and a weighting parameter γ (see 60), in which $n(x|\mu,\sigma^2)$ is the ordinate at x of a normal probability density function with mean μ and variance σ^2. For small γ, c is said to be a contaminated distribution, and γ is the fraction of contamination by data having variance σ^2. When $\gamma = 0.09$ and $\sigma = 3$, X' has a smaller variance than $\bar{X}$, and thus X' would be preferred to $\bar{X}$ as an estimator of μ.

Trimmed means

Trimmed means are also resistant to contamination; they are formed by setting aside a fraction α of the largest and smallest measurements, 2α in all, and averaging the middle $1 - 2\alpha$. Empirical studies suggest that, for a large class of distributions, trimming with $\alpha \approx 0.25$ seems to be satisfactory for estimating location. For the normal this trimmed mean has asymptotic efficiency of about 0.83 compared with $\bar{X}$ before contamination. For a contaminated distribution c with $\gamma = 0.05$ and $\sigma = 3$, the mean $\bar{X}$ has efficiency only 0.92 of that of the trimmed mean. With $\alpha = 0.25$ and $\gamma = 0.10$ this decreases to 0.775.

Bayesian methods. The methods of estimation described above are often referred to as classical in the sense of standard. The Bayesian approach offers an alternative that gives an especially convenient way of incorporating prior information about the parameters and also of adding further information when it arrives. One disadvantage is that the method apparently cannot proceed without a prior distribution for the parameters. The main device for inference is the posterior distribution of θ, given the data.

For a distribution of the random variable X given parameter θ with probability density function f, the likelihood of a sample of size n is, as before, a product of densities (see 61) each evaluated at sample points. If the prior distribution of θ has density $g(\theta)$, then the joint density of the sample and θ can be represented as the product (see 62) of this density and the likelihood of a sample. The posterior distribution of θ is required, given the sample. It has the density constructed as k times the joint density of the sample and θ (see 63), in which k is a normalizing constant that makes the integral of the function $p(\theta \mid x_1, \cdots, x_n)$ equal to 1, the integral extending over the possible values of θ.

In this present discussion θ is a random variable, whereas in methods of estimation discussed earlier it was regarded as a constant. The interpretation of $p(\theta \mid x)$ is of interest. It tells what the thoughtful observer who provided the prior distribution would believe about the state of the world after he received the data x. If his prior was largely based on personal subjective judgment, then the degree-of-belief interpretation is paramount, and others may prefer different priors and therefore different $p(\theta \mid x)$. When the prior comes from substantial frequency-based observations closely related to the real world, the poste-

rior distribution may well approximate an objective relative frequency distribution. Many practical situations fall between these extremes.

Much of the controversy over Bayesian methods has concentrated on the philosophy of the choice of the prior distribution and the difficulty of attaining agreement or unanimity. After completing a large Bayesian study, the U.S. statisticians Frederick Mosteller and David L. Wallace (1964) concluded that a neglected difficulty in Bayesian inference, as in methods of inference based on likelihood, is the sensitivity of the conclusions to the chosen family of data distributions; that is, to the functional form of f.

For a given $p(\theta \mid x)$, point or interval estimates can be obtained, or $p(\theta \mid x)$ can be used to make decisions. For example, the mean, mode, or median of the posterior distribution might be used as a point estimate for the true value. Alternatively, an interval (u, v) could be set up, like a confidence interval, that includes most of the probability, so that the integral with respect to θ from u to v of $p(\theta \mid x)$ equals $1 - \alpha$.

(64) $$p(\pi|x) = k\pi^{a+x-1}(1-\pi)^{b+n-x-1}$$

(65) $$\mu_{post} = \frac{a+x}{a+b+n}, \qquad \sigma^2_{post} = \frac{\mu_{post}(1-\mu_{post})}{a+b+n+1}$$

(66)

State of Nature	Action a_0 (accept H_0)	Action a_1 (accept H_1)
θ_0 (H_0 true)	$l(\theta_0, a_0)$	$l(\theta_0, a_1)$
θ_1 (H_1 true)	$l(\theta_1, a_0)$	$l(\theta_1, a_1)$

(67) $$\begin{aligned} \text{Risk}(\theta_0) &= l(\theta_0, a_0)\,\mathbf{P}(a_0|\theta_0) + l(\theta_0, a_1)\,\mathbf{P}(a_1|\theta_0) \\ &= \mathbf{P}(a_1|\theta_0) = \alpha \end{aligned}$$

(68) $$\begin{aligned} \text{Risk}(\theta_1) &= l(\theta_1, a_0)\,\mathbf{P}(a_0|\theta_1) + l(\theta_1, a_1)\,\mathbf{P}(a_1|\theta_1) \\ &= \mathbf{P}(a_0|\theta_1) = 1 - \gamma \end{aligned}$$

(69) $$\frac{L(x|H_0)}{L(x|H_1)} \leq c$$

(70) $$\int_W L(x|H_0)\,dx = \alpha$$

(71) $$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}, \qquad -\infty < x < +\infty$$

An example is a machine setup in which the proportion of defective parts produced in the initial setting is π and prior data suggest that π is distributed according to a beta distribution $g(\pi) = c\pi^{a-1}(1-\pi)^{b-1}$, in which c is the normalizing constant and $a > 0$, $b > 0$, $0 \leq \pi \leq 1$. After a setup, a sample of size n with x defectives has likelihood $f(x \mid \pi) = \pi^x(1-\pi)^{n-x}$. The posterior distribution of π given x defectives in the sample is proportional to the product of g and f (see 64). The posterior mean and variance of π are then calculable (see 65).

If both the prior distribution and the data distribution are normal with prior mean and variance μ_0 and σ_0^2, and the data observation and variance are x and σ^2, then the posterior distribution is again normal with mean μ_1 and variance σ_1^2. Their values can be computed from the relations $\mu_1 = \dfrac{h_0\mu_0 + hx}{h_0 + h}$, and $h_1 = h_0 + h$, in which $h_i = 1/\sigma_i^2$, $h = 1/\sigma^2$, so that $\sigma_1^2 = 1/h_1$.

When little information about θ is available, prior distributions for θ might be diffuse so that in the neighbourhood of any particular value of θ, the distribution is nearly flat. This would give the data greater weight in determining the posterior. Such a distribution is called a gentle prior distribution.

One's prior distribution of the length of a dimly seen and unfamiliar distant object would have a very large standard deviation compared to that of one direct measurement taken with a steel tape. If measurement errors were normally distributed with $\sigma = 0.1$ in inches and a gentle prior were used on the true value, the posterior distribution would be approximately normal with mean x and standard deviation 0.1.

HYPOTHESIS TESTING

When the probability model for the mechanism generating the observed data is known, hypotheses about the model can be tested. This involves the question: Could the present set of data reasonably have come from the model if the hypothesis is correct? Usually a decision must be made on the basis of the available data, and some degree of uncertainty is tolerated about the correctness of that decision. These four components (data, model, hypothesis, and decision) are basic to the statistical problems of hypothesis testing.

Use of the null hypothesis

In such problems a null hypothesis is formed (usually denoted H_0) that the assumption to be tested is correct. Then a significance level α $(0 < \alpha < 1)$ is chosen to indicate how improbable a set of data must be if H_0 is to be rejected. The particular samples to be considered improbable make up the critical region W, a subset of the sample space that has $\mathbf{P}(W \mid H_0) \leq \alpha$. In developing a test of a hypothesis, the central task is to construct the critical region.

When the null hypothesis is rejected, an alternative hypothesis, usually denoted H_1, is accepted in its place. The probability that a sample will lead to this action when the true value of the parameter is in H_1 is called the power, γ, of the test. If testing a hypothesis is viewed as a decision problem, there are two possible actions—"accept H_0" and "accept H_1"—and each has a corresponding loss depending on the true state of nature. (The terms "accept" and "reject" may give a rather unrealistic impression of finality, but they are the ones conventionally used in statistics as standard labels for the two actions; they owe part of their currency to one development of these ideas in the field of quality control, in which manufactured lots of material are accepted or rejected on the basis of samples.)

Many hypotheses concern values of a parameter (θ) in the probability model. In the most basic case both H_0 and H_1 involve a single value of θ (θ_0 and θ_1, respectively) and are said to be simple hypotheses. In this case the loss function may easily be displayed in a table: the two states of nature and the two possible actions determine four distinct values of the loss function (see 66). Here the concepts of decision theory provide a generalization of the classical formulation of hypothesis testing in which the losses are $l(\theta_0, a_0) = l(\theta_1, a_1) = 0$ and $l(\theta_0, a_1) = l(\theta_1, a_0) = 1$.

Risk is defined as the expected loss, given the state of nature. In that formulation the values of the risk are α at θ_0 (see 67) and $1 - \gamma$ at θ_1 (see 68). Thus among all tests with significance level α in a given problem, the one with the greatest power γ has the least risk.

To construct a test of H_0 with maximum power against H_1, a general technique, given in 1933 by the Romanian-born U.S. statistican Jerzy Neyman and the British statistician Egon Sharpe Pearson, may be used. The basis for this technique is the likelihood function $L(x \mid H)$, in which x denotes the sample and H a simple hypothesis. The critical region W for the test is that subset of the sample space for each element of which the ratio of the likelihood function for H_0 to the likelihood function for H_1 is less than or equal to c (see 69), with c chosen to give the desired significance level: the probability of W, assuming H_0 to be true (see 70), is required to be α. The quotient $L(x \mid H_0)/L(x \mid H_1)$ is referred to as the likelihood ratio.

For example, a hypothesis about the value of the mean μ of a Gaussian (normal) distribution with variance 1 might be tested (see 71). On the basis of a random sample of n observations $X_1, \cdots, X_n$, the null hypothesis H_0 that $\mu = \mu_0$ is tested against the alternative hypothe-

sis H_1 that $\mu = \mu_1 < \mu_0$. If a prescribed significance level α is used, $\alpha = 0.1$, say, then the appropriate critical region may be constructed. The likelihood function for any value of μ is formed from a product of normal densities each evaluated at sample points (see 72), and the likelihood ratio for the test is a ratio of two such products that differ in the value selected for the parameter μ (see 73). It is more convenient to work with the logarithm of the likelihood ratio, so that $(x_1 \cdots, x_n)$ lies in the critical region when the difference of two sums of squares (namely the quadratics that appear in the exponentials of the normals) yields a value that is less than or equal to $\log_e c$ (see 74). An algebraically equivalent condition is an inequality that sets an upper bound (dependent on c, μ_0, μ_1, and n) on the sum of sample values (see 75). Here c remains unspecified, but that causes no difficulty. In this problem the critical region does not require any distinction between samples that have the same sum. The critical region is described in terms of the sample mean $\bar{x}$: it is composed of the set of n-dimensional points for which the mean is less than or equal to a simple function of z_α (see 76).

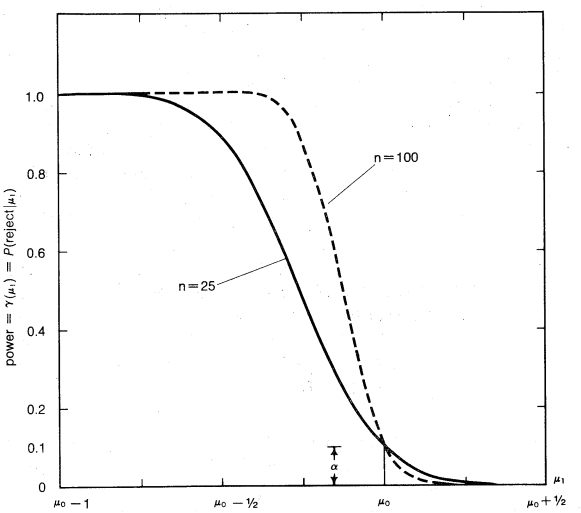
As might be expected from the assumption that $\mu_1 < \mu_0$, the critical region is composed of samples with means to the left of μ_0. For this test, the power is calculated (see 77) to be the function Φ evaluated at the sum of z_α and $\sqrt{n}(\mu_0 - \mu_1)$, in which Φ is the standard Gaussian cumulative distribution function. In the Figure power curves are plotted showing how the power depends on the separation between μ_1 and μ_0 and on the sample size n.

In the example, the critical region W depends only on having $\mu_1 < \mu_0$. Thus, in this problem the alternative hypothesis H_1 that $\mu < \mu_0$ could be used. Such hypotheses, involving more than one value of the parameter, are called composite. The power γ of the test is now a function of the true value of the parameter within the set of alternatives.

The curves in the Figure represent the power function $\gamma(\mu)$ for the two values of n. In general any attempt to construct a test with maximum power against a composite alternative hypothesis involves the formation of a power function $\gamma(\theta)$ that lies above other possible power functions for all values of θ within H_1. In some problems this is not feasible, and a considerable body of theory has been created to deal with such problems in great generality.

The descriptive level of significance

In practical significance testing, instead of presetting the significance level, the investigator may report the level of significance that would have just rejected the null hypothesis for his data. This level, called the descriptive level of significance or the nominal level of significance, has the advantage of giving a more complete report on the test and allows the reader to set his own significance level.



Plot of power curves for the example for $\alpha = 0.10$ and $n = 25$ (solid curve) and 100 (dashed curve; see text).

$$L(x_1, \cdots, x_n|\mu) = \prod_{i=1}^{n} f(x_i) = (2\pi)^{-n/2} \exp\left[-\tfrac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \tag{72}$$

$$\frac{L(x_1, \cdots, x_n|\mu_0)}{L(x_1, \cdots, x_n|\mu_1)} \tag{73}$$

$$-\tfrac{1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2 + \tfrac{1}{2}\sum_{i=1}^{n}(x_i - \mu_1)^2 \leq \log_e c \tag{74}$$

$$\sum_{i=1}^{n} x_i \leq \frac{\log_e c}{\mu_0 - \mu_1} + \frac{n}{2}(\mu_0 + \mu_1) \tag{75}$$

$$W = \left\{(x_1, \cdots, x_n) : \bar{x} \leq \frac{z_\alpha}{\sqrt{n}} + \mu_0\right\} \tag{76}$$

$$\begin{cases} \mathbf{P}(W|H_1) = \mathbf{P}\left(\bar{X} \leq \dfrac{z_\alpha}{\sqrt{n}} + \mu_0 | \mu = \mu_1\right) \\ \qquad = \mathbf{P}(\sqrt{n}(\bar{X} - \mu_1) \leq z_\alpha + \\ \qquad\quad + \sqrt{n}(\mu_0 - \mu_1)|\mu = \mu_1) \\ \qquad = \Phi(z_\alpha + \sqrt{n}(\mu_0 - \mu_1)) \end{cases} \tag{77}$$

$$\lambda(x_1, \cdots, x_n; \theta_{01}, \cdots, \theta_{0r}) = L_0(x_1, \cdots, x_n|\theta)/L_1(x_1, \cdots, x_n|\theta) \tag{78}$$

$$W = \{\lambda : \lambda \leq c\} \tag{79}$$

In more complicated situations a critical region in the original sample space may often be difficult to use because the region is defined by complicated functions of the data. When this happens, a technique that is based directly on a likelihood ratio may be used. The following illustration suggests some of the generality of the technique.

Suppose the parameter is $\theta = (\theta_1, \cdots, \theta_k)$, the null hypothesis is H_0: $(\theta_1, \cdots, \theta_r) = (\theta_{01}, \cdots, \theta_{0r})$ with $r \leq k$, and the alternative is H_1: $(\theta_1, \cdots, \theta_r) \neq (\theta_{01}, \cdots, \theta_{0r})$. These hypotheses are not concerned with $\theta_{r+1}, \cdots, \theta_k$. The approach is to determine the maximum value of the likelihood subject to the constraint imposed by H_0 and then use its ratio to the unconstrained maximum as the test statistic. This procedure, known as the likelihood ratio test, plays a role in hypothesis testing analogous to the role of maximum likelihood in estimation. For a random sample of n the test statistic is a function λ that is a ratio of likelihood functions L_0 and L_1 (see 78), in which L_0 is maximized over θ with the constraints that $\theta_i = \theta_{0i}$, $i = 1, \cdots, r$, and L_1 over all θ. From this definition it is clear that $0 \leq \lambda \leq 1$, and the cirtical region contains values of λ that are too small for H_0 to be reasonable: it is composed of values of λ that are less than or equal to c (see 79), with c chosen so that $\mathbf{P}(W \mid H_0) \leq \alpha$. When the exact distribution of λ is a function of the random sample $X_1, \cdots, X_n$, it is easy to choose c for the desired significance level. Otherwise the distribution of λ must be approximated. Fortunately, a remarkable asymptotic result may be employed: as n becomes large, the distribution of $-2 \log_e \lambda(X_1, \cdots, X_n; \theta_{01}, \cdots, \theta_{0r})$ when H_0 is true approaches a chi-square distribution with r degrees of freedom.

When data can be collected on a continuing basis, the techniques for testing a hypothesis should attempt to reach a decision as soon as the evidence is strong enough. Considered as a decision problem, this situation has three actions: "accept H_0," "accept H_1," and "continue sampling." Such methods are part of the field of sequential analysis.

When hypothesis testing was first mentioned, it was assumed that the probability model underlying the data was known. There is no reason to expect that this will always be the case; the reasonableness of a particular model may be exactly what is to be tested. Alternatively, it may be necessary to test whether the mean of one population is equal to the mean of another, without assuming a particular probability model for either. These are typical of the problems that form the subject of nonparametric hypothesis testing. Ordinarily, nonparametric methods depend less on the detailed shape of the sampling distributions than do methods called parametric. Nonparametric methods when used with measurements typically depend on signs of differences, or on counts, ranks, and permutations, rather than on the specific values.

STRUCTURE IN DATA

Many sets of data arising in actual practice are more complex than the situations and examples that have been described. The data may be collected according to a plan, and its structure must be considered in any analysis. Several properties may be measured for each observation in the sample, and their relationship may be of central importance. A process may be observed over a period of time, and it may be desired to discover or correct for cyclical aspects of its behaviour. These three types of situation are examples, respectively, of experimental design, of regression and correlation analysis, and of time-series and spectrum analysis. There are almost as many techniques of analysis as kinds of structure in data, but the most widely used are those of linear regression.

Regression analysis

The general setting for regression analysis is as follows. Each observation consists of a value of Y, which is regarded as a response, and a value of each of $X_1, \cdots, X_k$, which are regarded as factors affecting the response. A model that restricts Z to equal a linear combination of $\{U_i\}$ plus ε (see 80) is built, for example, in which Z is some function of Y, the U's are functions of the X's, and ε is random error, attributable to measurement error or natural variation. Z may be regarded as a response; it is often called the dependent variable. The U's, which are called carriers, may depend on the X's in ways prescribed by theory or suggested by preliminary examination of the data; they are often called independent variables, even when they may not be probabilistically or functionally independent. The parameters $\beta_1, \cdots, \beta_r$ in the model are usually called regression coefficients. Such a model is said to be linear because Z depends linearly on the U's.

At an early stage in analyzing data according to this linear or multiple regression model the β's are estimated, whether or not their numerical values are the primary goal of the analysis. One standard method for fitting the model to the data is least squares: given n observations $(y_i, x_{1i}, \cdots, x_{ki})$, estimates ($b$'s) for the β's are chosen that result in the smallest value of the sum of squared deviations from the fitted model (see 81). This sum is differentiated with respect to each of the b's, and the minimization problem is reduced to one of solving a set of r simultaneous linear equations in r unknowns. This can be done without consideration of the probability model for the random error ε, but when that error follows a normal distribution, the least-squares estimates of the b's are the same as the maximum-likelihood estimates.

Ordinarily, it is assumed that the random errors $\varepsilon_1, \cdots, \varepsilon_n$ are independent and have mean 0 and common variance σ^2; it is also assumed that they are the only source of random variation in the data. As a result, the estimated coefficients $b_1, \cdots, b_r$ vary randomly with mean values $\beta_1, \cdots, \beta_r$. When the ε's are normally distributed, the vector $b = (b_1, \cdots, b_r)$ will have a multivariate normal distribution. This fact can be used in setting a confidence region for β or a confidence interval for any single β_j.

Certain hypotheses about the model may be tested: is the carrier U_j necessary (usually posed as H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$)? Does the model adequately describe the data? The basis for many such tests is the component of the data that remains when the fitted model is subtracted. For each observation this component is called the residual: it is a typical term (see 82) that appears squared in the sum of squared deviations. The relationship between the residuals and other elements of the model is often examined graphically. Systematic patterns in such plots can show that assuming a common variance σ^2 for the ε's is suspect, that the data have additional structure (to be fitted by adding a carrier to the model), that some U's should have a different form, or that another function of Y is more nearly linear in the U's than the present Z.

$$(80) \quad Z = \beta_1 U_1 + \cdots + \beta_r U_r + \varepsilon$$

$$(81) \quad \sum_{i=1}^{n} [z_i - (b_1 U_{1i} + \cdots + b_r U_{ri})]^2$$

$$(82) \quad z_i - (b_1 U_{1i} + \cdots + b_r U_{ri})$$

$$(83) \quad \hat{y} = A(b_1 x_1 + \cdots + b_k x_k) + B$$

While fitting by ordinary least squares, assuming the common variance σ^2 for the ε's, is usually the first step in regression, it is quite sensitive to the presence of contamination or wild observations in the data. Consequently, a different analysis can be undertaken, based on a pattern of unequal variances or fitting the model by minimizing the sum of absolute-value deviations instead of the sum of squares.

For these calculations, as well as in graphical examination of the data and the model, the computer has become an indispensable tool. The discriminant function illustrates one use of linear regression in biology, medicine, the humanities, and the social sciences.

When no one property correctly classifies objects into their populations, measures on several properties may improve the process. The linear discriminant function applies multiple regression for this purpose. If $X_1, \cdots, X_k$ are the variables used to aid the classification, $Y = 0$ is defined for objects in population Π_0 and $Y = 1$ for those in population Π_1. Using data for objects that can be correctly assigned to Π_0 and Π_1, least squares is applied to predicting y from $(x_1, \cdots, x_k)$. The discriminant function is a linear function, with parameters A and B, of a linear combination of the $\{x_i\}$ in which the coefficients are $\{b_i\}$ (see 83). The parameters A and B are arbitrary scaling factors chosen for convenience, and the b_i are the least squares regression coefficients. These b_i, it can be shown, separate the y scores of the populations as widely as possible. To classify new objects, one chooses a cutoff score y_0, and values above it are classified as belonging to Π_1, below it to Π_0.

For example, the rates of use of the words *and, in, of, the,* and *to* by Alexander Hamilton and James Madison in political writing led to the discriminant function

$$\hat{y} = -3.014 - 0.0195x_1 + 0.0301x_2 + 0.0547x_3 - 0.0167x_4 + 0.0420x_5.$$

Here a piece of writing by Hamilton is scored 1 and a piece by Madison 0. The rate of use per 1,000 words of the word numbered i above is x_i. A cut-point, here 0.5, is chosen, and a piece of writing scoring higher is assigned to Hamilton, lower to Madison. Thus a new paper known to be written by one of these two authors having rates $x_1 = 30$, $x_2 = 20$, $x_3 = 60$, $x_4 = 100$, $x_5 = 40$ would score $\hat{y} = 0.295$ and be classified as by Madison.

BIBLIOGRAPHY. J.L. HODGES, JR., and E.L. LEHMANN, *Elements of Finite Probability*, 2nd ed. (1970); and FREDERICK MOSTELLER, ROBERT E.K. ROURKE, and GEORGE B. THOMAS, JR., *Probability with Statistical Applications*, 2nd ed. (1970), two elementary general texts; PAUL G. HOEL, *Introduction to Mathematical Statistics*, 4th ed. (1971), an elementary calculus-based text; B.W. LINDGREN, *Statistical Theory*, 2nd ed. (1968), an intermediate text; HARALD CRAMER, *Mathematical Methods of*

Statistics (1945, reprinted 1958), the classic work that organized the field of mathematical statistics; MAURICE G. KENDALL and ALAN STUART, *The Advanced Theory of Statistics:* vol. 1, *Distribution Theory*, 3rd ed. (1969); vol. 2, *Inference and Relationship*, 2nd ed. (1967); vol. 3, *Design and Analysis, and Time Series*, 2nd ed. (1968), an encyclopaedic standard text with extensive bibliography; SHELEMYAHU ZACKS, *The Theory of Statistical Inference* (1971), a standard work; FREDERICK MOSTELLER and JOHN W. TUKEY, "Data Analysis, Including Statistics," in GARDNER LINDZEY and ELLIOT ARONSON (eds.), *Handbook of Social Psychology*, 2nd ed., vol. 2, ch. 10 (1968), somewhat philosophical treatment of topics in statistics and data analysis, oriented to applications; DAVID L. SILLS (ed.), *International Encyclopedia of the Social Sciences*, 16 vol. and index (1968), contains 47 articles on special statistical topics and 39 biographies of statisticians; JUDITH TANUR *et al.* (eds.), *Statistics: A Guide to the Unknown* (1972), expositions of applications of statistics in a variety of fields, written for the layman; JOHN W. TUKEY, *Exploratory Data Analysis* (1970), innovative treatment of statistical data analysis; WALTER A. SHEWHART, *The Economic Control of Quality of Manufactured Product* (1931), a classic work; RONALD A. FISHER, *The Design of Experiments*, 8th ed. (1966), the classic work that established this branch of statistics; NATHAN KEYFITZ, *Introduction to the Mathematics of Population* (1968), the standard work that organized this branch of statistical methods; WILLIAM G. COCHRAN, *Sampling Techniques*, 2nd ed. (1963); and FREDERIC M. LORD and MELVIN R. NOVICK, *Statistical Theories of Mental Test Scores* (1968), two additional standard works.

Robust estimation: H.A. DAVID, *Order Statistics* (1970), a standard work with extensive bibliography.

Bayesian methods: D.V. LINDLEY, *Introduction to Probability and Statistics from a Bayesian Viewpoint*, 2 pt. (1965), a standard work; FREDERICK MOSTELLER and DAVID L. WALLACE, *Inference and Disputed Authorship: The Federalist* (1964), large-scale application of Bayesian and other statistical methods to a problem in discrimination; SAMUEL A. SCHMITT, *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics* (1969), an elementary text; GEORGE E.P. BOX and GEORGE C. TIAO, *Bayesian Inference in Statistical Analysis* (1973), a standard work.

Hypothesis testing: HERMAN CHERNOFF and LINCOLN E. MOSES, *Elementary Decision Theory* (1959), an elementary text; I.R. SAVAGE, *Bibliography of Nonparametric Statistics* (1962); ABRAHAM WALD, *Statistical Decision Functions*, 2nd ed. (1971), the classic work; E.L. LEHMANN, *Testing Statistical Hypotheses* (1959), a standard work; HOWARD RAIFFA, *Decision Analysis: Introductory Lectures on Choices Under Uncertainty* (1968), an elementary text; HOWARD RAIFFA and ROBERT SCHLAIFER, *Applied Statistical Decision Theory* (1961), a standard work that develops the mathematical tools for Bayesian decision theory.

Structure in data: THEODORE W. ANDERSON, *An Introduction to Multivariate Statistical Analysis* (1958) and *The Statistical Analysis of Time Series* (1971); GEORGE E.P. BOX and GWILYM M. JENKINS, *Time Series Analysis: Forecasting and Control* (1970); GWILYM M. JENKINS and DONALD G. WATTS, *Spectral Analysis and Its Applications* (1968); and OSCAR KEMPTHORNE, *The Design and Analysis of Experiments* (1952), all standard texts; ARTHUR P. DEMPSTER, *Elements of Continuous Multivariate Analysis* (1969), a standard text with a geometric and computational approach; ARTHUR S. GOLDBERGER, *Econometric Theory* (1964), a standard text for multivariate regression methods.

(Fr.M./D.C.H.)

Steam Power

Steam, or vaporized water, is a basic power source for industrial civilization. Steam power is indispensable in the manufacture of steel and other metals, in oil refining, and in the production of chemicals and many other products. It is used to drive the turbogenerators that furnish the major portion of electric power, to convert nuclear energy into mechanical energy, and to operate heavy transportation machinery.

HISTORY

Steam power before the Watt engine. The first recorded use of steam to produce motion was a simple reaction turbine built by Hero of Alexandria about AD 100.

Though Hero's demonstration of the reactive power of steam had tremendous potential, it was centuries before any attempt was made to give it practical value. In 1784 Baron von Kempelen of Pressburg (modern Bratislava, Czechoslovakia), attempted to develop a reaction turbine similar to the reaction water, or undershot, wheel. When James Watt in Scotland heard of von Kempelen's work, he correctly concluded that a wheel capable of moving fast enough to absorb the energy from steam could not be made with then-existing technology.

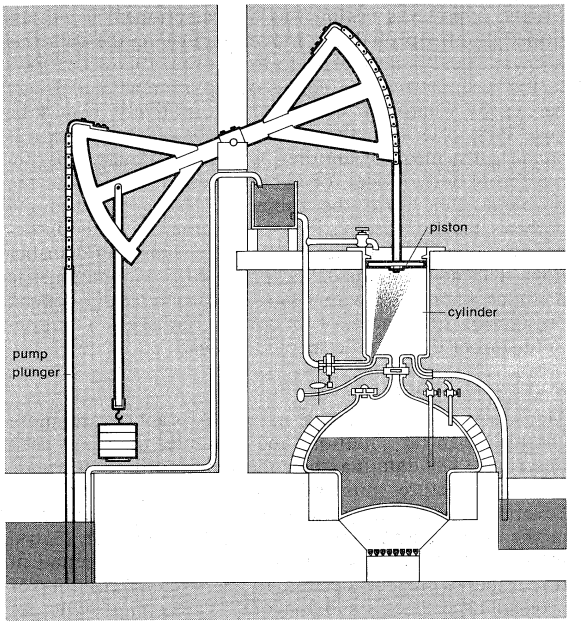


Figure 1: Newcomen engine (see text).

Savery pump and Newcomen engine

Meanwhile, an entirely different line of development led to Watt's steam engine. The problem of raising water from mines had long attracted engineering attention when, in 1698, Thomas Savery, an English military engineer, patented a pump for "raising water by the impellent force of fire." Savery's pump consisted of a boiler, a closed vessel called a receiver, one-way valves in the water inlet and the outlet of the receiver, a steam supply line from boiler to receiver with a shutoff valve, and a waterline with a cock for spraying water on the exterior of the receiver. The pump was operated by hand manipulation of the steam-supply and water-spray valves. When the steam valve was opened to supply steam to the water-filled reservoir, steam pressure forced the water out through the one-way-outlet valve. When the water was all expelled from the reservoir, as indicated by the clatter of the outlet valve, the steam valve was closed. The water cock was then opened and water sprayed over the surface of the reservoir, condensing the steam and creating a vacuum. Atmospheric pressure in the mine caused water to flow into the reservoir. The vacuum created in the reservoir was not perfect, so atmospheric pressure was capable of lifting water only about 20 feet (6 metres). The steam-powered discharge was also limited, owing to the inability of the boilers of that day to withstand high pressure. Direct contact of the steam and water within the reservoir not only caused a heat loss but the steam pressure had to be at least equal to the pressure represented by the height to which the water was pumped.

Thomas Newcomen of Dartmouth, Devon, England, and his assistant, John Calley, developed the first piston-operated steam engine in 1712.

The Newcomen engine (Figure 1) consisted of a cylinder fitted with a piston. When the cylinder was filled with steam the counterweighted pump plunger moved the piston to the extreme upper end of the stroke. With the admission of cooling water, the steam condensed, creating a vacuum. The atmospheric pressure acting on the piston caused it to move down in the cylinder, and the resulting force acting upon the beam lifted the pump plunger. The steam was condensed, at first, by applying the cooling water to the exterior of the cylinder. When this was found to be too slow, the water was sprayed inside the cylinder. Another problem was that air and other noncondensable gases accumulated within the cylinder causing

the engine to stop after a few strokes. This difficulty was overcome by allowing some steam to blow out each time the cylinder was filled. The valves for admitting steam and condensing water to the cylinder were first operated by hand. Later they were operated by the motion of the engine so that after start-up the engines became self-actuating.

Savery and Newcomen probably worked independently in the development of their quite different types of engines. Since, however, Savery's patent included a reference to "creation of a vacuum by the condensation of steam," it was necessary for Newcomen to enter into a partnership with him. Before Newcomen's death in 1729, numerous pumping engines were in use in England and in several other European countries.

The practical success and usefulness of the Newcomen engines is proven by their long service record. A few with separate condensers and other features introduced by Watt were built as late as the 1800s. As mining operations changed, the pumps were moved to new locations and sometimes modified. Some are known to have remained in operation until the early part of the 20th century when they were replaced with electric pumps.

Watt's engine. In 1765 James Watt, an instrument maker in Glasgow, modified a Newcomen engine with a separate condenser to make it unnecessary to heat and cool the cylinder with each stroke. Because the cylinder and piston remained at steam temperature while the engine was operating, fuel costs dropped 75 percent. In 1769 Watt obtained a patent for his "new method of lessening the consumption of steam and fuel in fire engines." He entered into a partnership with Matthew Boulton who owned the Soho Manufactory at Birmingham, England. In 1775 the House of Commons extended the patent for 25 years.

Boulton and Watt sold engines on a basis that was to give them one-third of the saving in fuel from that required by a Newcomen engine. Watt made extensive tests to establish the comparative steam requirements of his engine with those of the common engine.

At Boulton's insistence, Watt applied his great inventive ingenuity to developing a new kind of engine which rotated a shaft instead of providing simple up-and-down motion. Watt found a way to obtain an inflexible connection between piston and piston rod (beam), and invented special gear arrangements to convert the up-and-down movement of the beam into circular motion. A heavy flywheel was added to smooth out the variations in the force delivered to the engine shaft by the action of the piston in the cylinder. An ingenious governor, connected to the flywheel, regulated the flow of steam to the engine. Watt, furthermore, applied steam to both sides of the piston to produce greater uniformity of effort and increased power, and invented packing glands which sealed against steam leakage where the piston rod passed through the cylinder head. Though far more difficult to build, Watt's rotative engine created an entirely new field of application. The rotative engine was immediately accepted, and it is estimated that by 1800 the partnership had built 500 engines, of which 38 percent were pumps and 62 percent were of the rotative type.

Owing to the inherent weakness of the boilers and Watt's emphasis on safety, steam pressures remained at five pounds per square inch. Though Watt understood the advantages of the use of the expansive powers of steam within a cylinder and the use of compound engines, neither was economically attractive at the low steam pressures Watt employed. In a compound engine, high pressure steam is supplied to one cylinder, the exhaust from it is then admitted to a second, the exhaust from the second to a third, and in some cases from the third to a fourth. Each cylinder is larger than the preceding one because of the increase in steam volume as the pressure is reduced.

Watt's pressure indicator

Watt also invented a device which graphically indicated the pressure inside the cylinder of the engine. This invention is still used to test the performance of engines and to adjust them for maximum efficiency.

Development of high-pressure boilers. The first instance of the successful generation of steam for a practical application was the pressure cooker invented in 1680. From this beginning developed the steam boiler, soon widely used in conjunction with the Newcomen and Watt engines. Early boilermakers observing that a hollow vessel usually assumes a cylindrical or spherical shape in response to internal pressure, designed their boilers accordingly. Flat plates, required where water and steam tubes were connected, were supported to prevent bulging.

Shells were constructed of steel plates joined together by rivets. The technique of joining boiler plates by rivetting was highly developed before welding replaced it in the 20th century.

Since Watt's engines used low-pressure steam, simple boilers were adequate and there was but little development in design. By the latter part of the 18th century, two types of boilers had developed. In one, the fire-tube boiler, hot gases in tubes heated the surrounding water. In the other, the watertube boiler, the fire is external to the tube.

Though these first watertube boilers were not very successful, they showed great promise, and development of them continued throughout the 19th century. By the beginning of the 20th century, quite sizable watertube boilers were in common use.

GENERATION OF STEAM

Steam boilers. The modern boiler, like its earliest forerunners, is basically a closed vessel supplied with water which is transformed into steam by heat. But because the pressure at which the steam is generated determines the temperature of the water inside the boiler, the ability of the boiler to withstand pressure is critically important. The application of more heat to the boiler produces additional steam at a constant temperature, provided the pressure does not change.

Pressure requirements

Required operating pressure is a major consideration in the selection of a boiler. If the steam is to be used for heating purposes, the required temperature determines the operating pressure. When steam is used to generate power the pressure must be sufficient to meet the requirements of the steam engine or turbine. Generally, higher pressure results in greater efficiency of conversion from heat to mechanical energy. Pressure is an important consideration when steam from the same source is used both for power generation and heating.

Boilers must be selected to meet specific pressure requirements, and pressure rating determines design. All components must withstand not only the pressure but also the accompanying high temperatures. Though the steam piping and valves on a high-pressure boiler are smaller than those on a low-pressure boiler of comparable capacity, they must withstand higher temperatures. Other considerations include provisions for circulation of the water in the boiler to carry away the heat transmitted through the surfaces and separation of steam from the water to assure that the steam discharged from the boiler is as free of water particles as possible.

Superheated steam

Steam is designated as superheated when it is heated beyond the temperature at which it is produced by boiling water. Superheated steam is produced by passing the steam discharged from the boiler outlet, usually the steam drum, through tubes that are exposed to hot gases. Superheaters, usually installed within the boiler casing, absorb radiant and convective heat from the boiler furnace. In some instances, however, separately fired superheaters are used. The amount of superheat must be controlled since piping systems, engines, and turbines are designed for a given temperature.

Superheaters add to boiler cost since they must be constructed of nickel or molybdenum steel and because additional controls are required, but this added cost can be justified by improved economy.

The heat, absorbed by the boiler in generating and in superheating, is produced by burning fuels such as coal or oil or from hot gases discharged from industrial processes. In order for it to operate at maximum efficiency, as much heat as possible must be transferred from these gases to the water inside the boiler. This is accomplished by providing as much internal heating surface in the boiler as possible, consistent with initial costs.

At high pressures the water in the boiler is at a correspondingly high temperature, so the heating gases leave the boiler at elevated temperatures and much heat may be wasted. To prevent this, heat-recovery equipment is employed. Such equipment may be designed to heat the water before it is fed into the boiler drum, or to heat the air for combustion before it is introduced into the furnace. The equipment used to heat the feedwater is referred to as an economizer, while that which heats the combustion air is an air preheater. In some plants both economizers and air preheaters are installed. The type and size of a heat-recovery unit used depends on fuel costs, fuel analysis, the extent to which the plant will be operated at or near capacity, and other considerations.

Fire-tube boilers. Fire-tube boilers, so-named because the hot gases circulate through tubes to heat the water surrounding the tubes, are extensively used. Tube arrangement may be vertical, horizontal, or inclined. For applications in which weight and space limitations exist, the boiler and combustion equipment (firebox) are constructed in a single unit. The firebox is made of a double wall of steel plates arranged to leave a space between the plates for water. These plates must be internally reinforced to resist steam pressure. The horizontal-return tubular boiler (Figure 2), a fire-tube boiler formerly widely

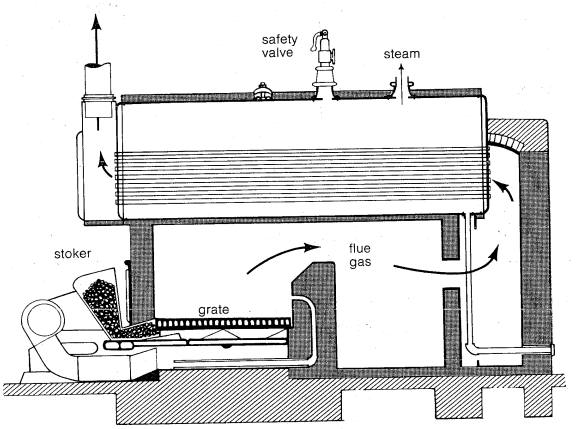


Figure 2: Early horizontal-return fire-tube boiler with single-retort-underfeed stoker.

used in plants for heating and power generation, has a cylindrical shell of steel plates. Tubes that extend through the shell from one head (flat end) to the other provide the major part of the heating surface and serve as stays for a section of the heads.

The shell is placed in a brick-walled enclosure, lined with refractory materials capable of withstanding high temperatures. Boilers may be fired with solid fuels by hand, or mechanically with stokers, or oil or gas can also be used. Combustion of the fuel begins in the front of the unit and continues as the gases pass over a low separating wall made of firebrick, called the bridge wall, and under the boiler shell. Thus, the shell is exposed to the radiant heat from the solid fuel bed on the grate, or flame from the oil or gas burner. The hot gases (combustion products) pass through the tubes and out the stack. These boilers, though limited in maximum pressure and capacity, are very reliable.

Package boiler

The fire-tube principle has been adapted to present-day requirements in the low-cost, prefabricated, and easily installed package-type boiler, which consists of a large shell mounted on a steel foundation, with flat heads (tube sheets) welded to each end. A flue (large tube) and a number of smaller tubes extend through the shell from one tube sheet to the other. The ends are secured to prevent the tube sheet from bulging from the pressure and to resist leakage. This assembly provides the complete unit for the section of the boiler subjected to steam pressure. Since the flue is exposed to the intense heat of combustion, it is usually corrugated around its circumference to allow for expansion.

The oil- or gas-burner flame is directed into the flue that serves as the combustion chamber. Some of the heat is absorbed by the water by transfer through the walls of the combustion chamber; the remaining heat is discharged in the products of combustion (gases) that enter a compartment at the rear of the boiler and are directed back through a section of tubes toward the front, where they are again reversed and pass through another section of tubes toward the rear, completing three passes through the boiler. Some boilers are designed for four gas passes. Three-pass boilers have flue connections in the rear; four-pass boilers in the front. In some designs the compartments in the rear of the boiler, where the gases reverse direction, are lined with heat-resistant materials while in others they are constructed of water-cooled steel plates.

The package boiler has burners, air-supply fan, and controls mounted directly on its front. The fan supplies pressurized air for the burner, where the fuel and air are mixed in correct proportions for efficient combustion. The pressure developed by the fan not only delivers the air to the burner but also forces the products of combustion through the combustion chamber, tubes, and out the smoke stack, which need be only high enough to prevent a nuisance and meet antipollution requirements.

These boilers may be made essentially automatic in operation with controls that vary the fuel supply to meet the steam demand and regulate the air supply to maintain efficient combustion. Safeguards are provided to prevent the boiler from starting up if safe conditions do not exist and to shut the unit down if unsafe conditions develop in the course of its operation. There is a wide variation in legal requirements for boilers and plants, most of which are based on steam capacity and pressure.

Packaged fire-tube boilers are extensively used in supplying steam for space heating and for industrial process requirements. Completely factory-assembled, they may be transported as a single unit, lowering costs. Installation time is reduced since no special foundations are required, and it is necessary only to connect the water, fuel and electric power supplies, drain, and steam and flue gas outlets. The compactness of these units is also an advantage.

Watertube boilers. The watertube boiler, in which the hot gases surround the tubes and the water flows through the tubes, permits the designer to provide almost any amount of heating surface required. The steam drum containing the reserve supply of water is relatively large in diameter and must, therefore, be made of thick plates to withstand the pressure. Advantages of watertube boilers include adequate circulation so that steam formed on the tube surfaces is rapidly carried to the steam drum, reduced hazard in the event of tube failure, and less costly repairs.

Optimum heat transfer

The drum or drums and tubes are the essential components which must be made to withstand steam pressure. Many arrangements are used to provide optimum transfer of heat from the hot gases, circulation of water within the boiler, separation of water from the steam before it leaves the drum, and expansion of the drums and tubes owing to variations in temperature.

Early watertube boilers were encased in brick settings that enveloped the pressure section, provided space for the furnace where the fuel was burned, and directed the hot products of combustion around the exterior of the tubes. To maximize heat transfer, restrictions (baffles) were installed. These forced the products of combustion to flow through the bank of tubes several times before being discharged to the stack.

Boiler tubes originally were installed close to the furnace to maximize heat transfer from the high-temperature bed of burning fuel. It was found, however, that this arrangement did not allow enough space and time for complete combustion; gases leaving the fuel bed entered the tube bank and were cooled before all combustible material had been burned. Part of the heating value of the fuel was lost, smoke was discharged from the stack, and soot formed on the tube surfaces retarding the rate of heat transfer.

Design was directed toward boilers with larger capacity and improved efficiency. These larger boilers were set

higher above the fuel bed, permitting the gases to travel further before coming in contact with the tubes. The increased distance, however, meant that less heat was radiated to the tubes. This resulted in a higher temperature in the fuel bed and furnace. When temperatures exceeded the softening point of the ash in the fuel, clinkers formed in the fuel bed, restricting the flow of air. High temperature combined with the action of the ash resulted in the rapid deterioration of the furnace refractory (firebrick) walls. Particles of ash adhered to the boiler tubes, decreasing heat transfer and, in extreme cases, restricting the flow of gases.

Large steam-generating units are now made by combining adequate space for fuel combustion with a furnace able to withstand the high temperatures developed (Figure 3). Experience with different fuels dictates the required furnace configuration and volume. The furnace enclosure is made of tubes placed side by side, sometimes welded to form a solid steel wall. These tubes (waterwalls) contain water and are actually part of the boiler. The waterwall tubes absorb heat and thereby reduce the furnace temperature, while adequate furnace volume is provided to obtain efficient combustion of the fuel.

By courtesy of the Babcock & Wilcox Co.

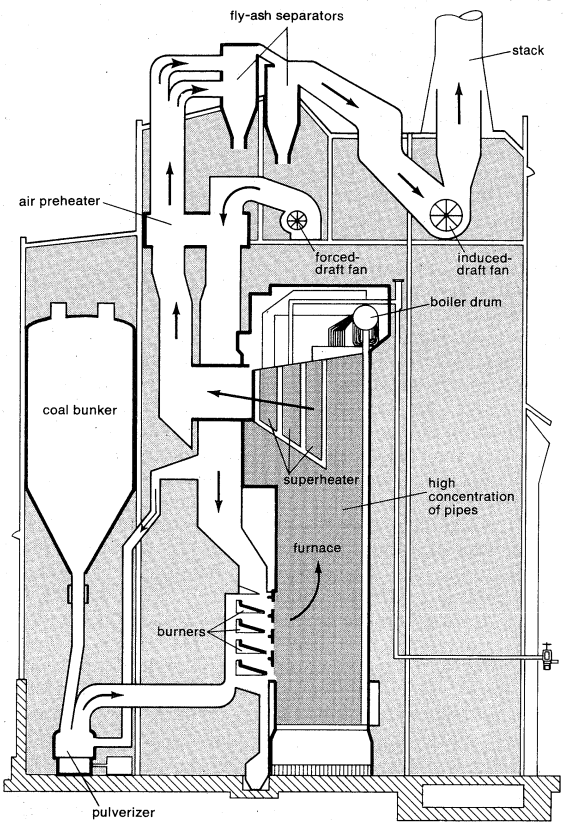


Figure 3: Cross section of a typical central-station boiler.

These large watertube boilers with waterwall furnaces provide adequate furnace volume, reduce furnace maintenance, permit the use of preheated combustion air and decrease the required building size. Such a design allows highly preheated air to be employed for combustion, thus permitting the use of air preheaters and increasing overall efficiency. The large quantities of fuel burned in these boilers justify the use of such heat-recovery equipment; both air preheaters and economizers are frequently installed on the same boiler. Superheaters are also an integral part of such units. After leaving the drum the steam passes through the superheater tubes and is heated to the specified temperature before being discharged into the station piping system.

Water supplies and treatment. Water must be supplied to the boiler in quantities equal to the rate at which steam is being generated. The water normally available contains impurities in the form of suspended solids, dissolved solids, and gases. Oxygen contained in the feedwater and carbon dioxide produced within the boiler may leave with the steam. These gases cause corrosion and other problems. The boiler may corrode, or become embrittled, losing strength. Impurities may produce scale, which by impeding the flow of heat causes the tubes to become overheated and to fail. Foam may form on the surface of the water in the drum and cause moisture and impurities to leave with the steam, creating deposits in the steam lines. Impurities cannot be tolerated in cases in which the steam comes in contact with the product.

To prevent these conditions from arising, the boiler water must be conditioned. Such conditioning includes treatment before it enters the boiler, the addition of chemicals to the boiler, and draining some water from the boiler to prevent excessive buildup of impurities. Pretreatment of water includes settling to remove suspended solids and the addition of chemicals to remove dissolved impurities or to convert them to a less objectionable form. Dissolved gases are removed, and chemicals are added to the water in the boiler to prevent its impurities from forming scale and to prevent it from corroding or foaming.

Recycling water

Recycling the water in a power plant is accomplished by returning the condensed steam (condensate) from the heating coils and condenser to the boiler. Except for possible leakage of cooling water into the condenser and leakage into the heating coils, the condensate is free from impurities. The use of condensate vastly reduces the amount of water that must be treated in a steam-generating plant.

Fuels. Fossil fuels such as coal, oil, and gas contain carbon and hydrogen as essential heat-producing elements. One pound of carbon combines with oxygen to develop 14,540 British thermal units of heat, while one pound of hydrogen develops 62,000 British thermal units. A British thermal unit (BTU) is the quantity of heat required to raise the temperature of one pound (453 grams) of water 1° Fahrenheit at or near 39.2° F (4° C). Air (about 23 percent oxygen and 77 percent nitrogen by weight) is used to supply the oxygen.

Boiler furnaces bring the fuel and air into contact at elevated temperatures in order to effect a complete chemical reaction (combustion) of the carbon and hydrogen. To obtain complete combustion more air must be added than that theoretically required. This excess air results in the presence of oxygen and additional quantities of nitrogen in the products of combustion (flue gases).

Coal is the principal fossil fuel used in steam generation. Combustion of coal leaves ash, an inert material that must be disposed of. Moreover, ash frequently forms into large chunks called clinkers and the fine particles (fly ash) are carried along with the flue gases, thereby leading to possible pollution of the atmosphere. Coal burning plants therefore include ash-handling equipment and collectors for removing a large portion of the ash from the gases before they are discharged from the stack. Sulfur, another troublesome impurity in coal, provides only a small amount of heat, but the sulfur dioxide and sulfur trioxide in the products of combustion are highly corrosive and objectionable atmospheric pollutants.

In early plants, coal, burned on stationary grates, was introduced into the furnace by hand. The air required for the combustion came up through the grates, from which the ashes were removed by hand. As boilers increased in size, hand firing became impractical and mechanical means of conveying fuel to the boiler and feeding it into the furnace were devised. Three basic types of mechanical feeders, or stokers, are in common use.

Basic types of stokers

1. The chain- or travelling-grate stoker consists of a front and a rear sprocket (circular-toothed gear) and a moving grate on which the fuel bed slowly moves through the furnace. Fuel is fed on the grate from a hopper located at the front of the furnace. Upon entering, the fuel is ignited and continues to burn as it moves through the furnace. The ash is discharged as the grates pass over the rear sprocket. Feed rate is varied by adjusting the speed of travel and the depth of coal on the grate.

2. The underfeed stoker employs a ram pusher or screw to force the coal into the furnace, injecting it beneath the

burning fuel bed. Fuel is distributed inside the furnace by moving or inclining the grates. Ash is removed from smaller units by periodic dumping of a section of the grates. In larger units, ash is removed continuously by a power-driven rotary ash discharge (clinker grinder).

3. The spreader stoker simulates continuous hand firing. Coal is supplied at a controlled rate to a power-driven rotary distribution mechanism that throws it into the furnace. Some combustion takes place while the fuel is being thrown through the high-temperature furnace, and the remaining combustion takes place after the fuel falls on the grates. In smaller installations, ash is removed by hand cleaning of stationary or dumping grates. Multiple feed and distribution units must be used since one must be out of service during cleaning. Large spreader-stoker units employ moving grates; ash removal is continuous.

Large watertube boilers with waterwall furnaces utilize pulverized fuel to generate heat. Particles of various degrees of fineness are employed, depending on the specific fuel and the combustion equipment. The fine particles are blown into the furnace by a stream of primary air. Here the fuel is ignited and the required secondary air is added. In this way the fine fuel particles are thoroughly mixed with the air and are burned in suspension as they pass through the furnace.

Fuel oil used to generate steam is divided into several classes according to its specific gravity (that is, its density compared to water). Lighter grades are used in domestic furnaces and for standby service to supplement other fuels. Heavier grades are burned in power boilers. In order to obtain efficient combustion, the oil is broken up into fine particles (atomized), so that it will mix with the combustion air. Heavy oils are heated in order to reduce their viscosity and improve their pumping and atomizing characteristics.

Gas, either natural or man-made, is another fuel widely used for steam generation. Natural gas is widely distributed by pipeline systems. In some instances gas is manufactured as a by-product from coke ovens, blast furnaces, and oil refineries. The gas is introduced into the combustion chamber and burned to provide heat.

In addition to the conventional fuels, certain waste products such as scraps from sawmills and woodworking plants, paper-mill sludge, sugarcane fibres (bagasse), and combustible material in rubbish are employed.

Heat for steam generation is also produced by nuclear reactors. In this case, the boiler and furnace which utilize fossil fuel are replaced by a reactor (see also NUCLEAR REACTOR).

STEAM ENGINES

Development and operating principles. Though relatively inefficient, the steam engine is, nevertheless, a most important device for the conversion of heat into mechanical energy. A piston moving back and forth within a cylinder uses steam to perform work. When steam pressure is exerted on one side of the piston, the resulting force causes it to move the length of the cylinder. Then the steam is applied to the other side of the piston and it moves in the opposite direction. This reciprocating motion is changed into rotary motion of the engine shaft by means of a connecting rod and crank. During the operation the cylinder is alternately filled with steam from the high-pressure supply line and exhausted at a lower pressure. In this way a portion of the heat in the steam is converted into work.

The amount of work delivered by one stroke of the engine (one traverse of the piston within the cylinder) is the product of the average force on the piston times the distance of travel. The force can be increased by the application of higher-pressure steam or by discharging the exhaust steam at a lower pressure.

Simple and compound reciprocating engines. A conventional double-acting steam engine admits steam alternately to each end of the cylinder, resulting in two power-producing strokes for each rotation of the crankshaft. Steam is alternately admitted to and discharged from each end by means of a D-slide valve, so-named because it has a cross section shaped like the letter D. This valve is moved back and forth over the steam passages (ports) to the cylinder by a connection to the engine shaft. The arrangement is adjusted to open and close the ports when the piston is in the correct position to receive and exhaust steam.

Engine speed is controlled by a flyball governor (so-named because it employs two small spinning balls to actuate the valve-controlling mechanism) driven by a belt from the engine shaft. When engine speed increases, the governor slowly closes the steam supply valve, reducing the pressure of the steam admitted to the cylinder and, when speed decreases, the governor slowly opens the steam supply valve.

Engine efficiency

Engine efficiency is improved by having the steam supply to the cylinder shut off, before the piston reaches the end of the stroke. After the supply is cut off, the steam in the cylinder expands and continues to push the piston, but with a continuous decrease in pressure. The power output from a cylinder can be effectively controlled by varying the point of steam cutoff.

Control of the point of steam cutoff cannot be obtained with a simple D-slide valve without also changing other valve functions. This and other disadvantages prompted modification of the D-slide valve and development of other types of valves. The piston valve, for example, permits the use of higher pressures and more superheat. By adding an extra component called a riding cutoff to a D-slide and piston valve, the cutoff can be controlled without varying the other functions of the engine cycle; an engine with a riding cutoff controlled by a flywheel governor is called an automatic engine.

A type called a Corliss-valve engine employs separate intake and exhaust valves at each end of the cylinder (four valves per cylinder) to permit separate control of steam admission and exhaust. The admission valve is not cooled by the exhaust steam. The design also provides for quick closing of the admission valves. Poppet- (perpendicular-) valve engines, which also have separate valves for intake and exhaust, can use highly superheated steam.

The flow of steam in and out of an engine cylinder results in the loss of heat from the heating and cooling of the cylinder (cylinder condensation). The greater the temperature difference between the steam supply and exhaust, the greater the heat loss. This loss is reduced, therefore, by the use of compound or multiple-expansion engines in which high-pressure steam passes through three or four successively larger cylinders. When each cylinder has a connecting rod and crank to deliver its power to a common shaft it is called a cross-compound engine. Tandem-compound engines have the cylinders located in line and a common piston rod, connecting rod, and crank.

The uniflow engine provides another means of reducing cylinder condensation. Steam enters at the end of the cylinder and leaves through exhaust ports located near the middle of the cylinder. Perpendicular or poppet valves admit the steam; the piston uncovers the exhaust ports. Only steam flows through the supply ports and exhaust through the discharge, thus eliminating the repeated heating and cooling which occurs in other designs.

Turbines. The steam turbine changes the heat energy in steam into mechanical energy by discharging the steam through nozzles, where its heat energy is transformed into energy due to velocity. The stream of high-velocity steam strikes vanes on the moving element of the turbine (rotor) causing the rotor to turn. For a given steam supply and exhaust system, the identical amount of heat is available for conversion into mechanical energy whether a moving-piston engine or a steam turbine is used. Thus the useful output of energy depends upon the efficiency of the engine or turbine.

Impulse and reaction turbines

The steam turbine employs both the impulse (forward) and reactionary (backward) forces resulting from the discharge of steam through nozzles and both principles are frequently employed in different sections of the same turbine. In impulse turbines, the nozzles are stationary and the high-velocity steam expends its energy moving

the blading. In a reaction turbine the steam is discharged from moving nozzles against stationary blades or vanes.

Steam supply pressure, superheat, and exhaust pressure all determine the amount of heat energy available to a turbine. The turbine's ability to convert the energy into useful work depends upon many factors involved in turbine design. The best transfer of energy occurs when the turbine blade is traveling about half as fast as the steam discharged from the nozzle. To achieve this velocity ratio, the designer regulates pressure drop in the first set of nozzles and employs several sets of nozzles arranged in sequence to utilize the full amount of the steam pressure. Optimum blade shape, and the use of finely ground and polished surfaces contribute to improved efficiency. When the exhaust steam is used for heating, the conversion efficiency is less important and the added cost for multiple sets of nozzles cannot always be justified.

The energy output of a turbine, and hence its speed, is controlled by a governor. In its simplest form a single valve governor regulates the pressure of steam supplied to the first set of nozzles. When the turbine is operated at less than rated capacity, efficiency may be improved by the use of multivalve control, in which the governor opens a series of valves in sequence as load requirement increases. Each valve supplies steam to a nozzle or set of nozzles. Full pressure is applied to one set before steam is admitted to the next. Turbine governors are very responsive and capable of maintaining the accurate speed control required to drive alternating current generators.

Steam turbines are ideally suited to use high pressure, highly superheated steam, since the steam does not come in contact with lubricated surfaces. Steam temperatures are limited by the material used in construction of the turbine. Provision must be made to reduce leakage around the shaft at the inlet, outlet, and between stages. Small- and medium-sized turbines for moderate steam conditions use carbon-seal rings held against the shaft by springs. Large turbines using high-pressure and high-temperature steam have packing consisting of a grooved shaft running with close clearance inside a grooved sleeve. Because of the close clearances, careful consideration must be given to movements resulting from expansion and to vibration. The rotating elements must be accurately balanced. Minimum vibration is the goal of every turbine manufacturer. Pressure lubrication in which oil is supplied to the bearing under pressure and is then cooled contributes to low maintenance costs.

The versatility of the steam turbine makes it suitable for many industrial applications. Turbines are used to drive fans, pumps, air compressors, presses, and similar equipment; their speed may be controlled to meet almost any requirements. Efficiency decreases, however, when turbines are operated at partial load and reduced speed. Mechanical steam-drive turbines are especially attractive in installations in which the exhaust steam can be used for process and/or space heating, because the heat energy changed to power is very small. Mechanical drive turbines are also used for emergency electric generators and fire pumps. These may be arranged to start automatically when an emergency arises.

The most extensive use of steam turbines is in central electric power stations in which they are used to drive electric generators (see also ELECTRIC POWER). High initial steam pressures and temperatures and low exhaust pressures are used to obtain optimum efficiencies. Further economy is obtained by reheating the steam. When steam has expanded part way through the cycle, the temperature and pressure having been reduced, it passes through a superheater (reheater) in which heat is added and the temperature increased before expanding through the final stages of the turbine.

Regenerative cycle

Since pressure is reduced in the consecutive stages of a turbine, steam at any desired pressure can be removed if connections are provided in the casing. This feature provides a versatile means of balancing power and heating requirements since the steam can be expanded in the turbine until it reaches the pressure required for heating. This principle is applied in electric generating plants to what is called a regenerative cycle. Steam from several stages of the turbine is used progressively to heat the condensate (water discharged from the condenser) on its way back to the boilers. This improves efficiency by reducing the steam flow to the condenser, and hence the heat rejected to the cooling water. The required extraction points to meet steam heating requirements must be incorporated in turbine design.

Utilization of high-pressure and more efficient condensers. Sadi Carnot (1796–1832), a French engineer, established the relation between the temperature of the working fluid and its mechanical effect. In 1824 he published a paper stating that the optimum conversion of heat energy to mechanical energy is the absolute temperature leaving the boiler less the absolute temperature of the vapour in the condenser, divided by the absolute temperature leaving the boiler. Stated in a formula: Efficiency $= (T_1 - T_2)/T_1$. Applied to steam power, T_1 is the absolute temperature (degrees Celsius +273, or degrees Fahrenheit +460) leaving the boiler, and T_2 is the absolute temperature of the vapour in the condenser. This represents the maximum possible theoretical efficiency and serves as a useful means of evaluating results. It cannot be attained in practice owing to losses which are inherent in physical equipment.

The efficiences of power generation have been continually increased by the use of higher and higher steam temperatures. This, in turn, has been made possible by the use of heat-resistant steel alloys in superheater tubes, piping, and turbine components. Temperatures in excess of 1,000° F (538° C) have been successfully employed. Because of the increased cost of material and more rapid deterioration of metals requiring costly maintenance and interruption to service, however, 1,000° F is the nominal upper limit of temperature.

As Watt discovered, a separate condenser is a satisfactory method of reducing the final (exhaust) temperature. The exhaust steam is condensed, the volume reduced, and a vacuum created. The pressure depends upon the temperature of the cooling water, efficiency of the noncondensible gas-removal equipment, and the heat transfer rate in the condenser.

There are two principal types of condensers, a jet condenser in which the cooling water and steam come into direct contact; and a surface condenser in which the cooling water flows through tubes that are surrounded by steam.

Jet condensers are seldom used in generating plants because the condensate is mixed with, and therefore contaminated by, the cooling water. This cannot be tolerated in high-pressure plants because of complications and the cost involved in conditioning the water supplying the boilers.

Widely used in steam electric generating plants, surface condensers contain thousands of small tubes, surrounded by vapour and through which water flows. The amount of cooling water required is from 50 to 100 units per unit of steam, depending upon the temperature rise which takes place in the condenser.

Since the pressure inside a condenser is below that of the atmosphere, there is a tendency for air to leak into it. This air, along with some gases carried along with the steam, collects in the condensers and is noncondensible, since it is not removed by the cooling action of water and must be exhausted by vacuum pumps.

Auxiliary equipment (valves, governors, etc.). Auxiliaries to steam engines and turbines include pumps, lubricating systems, condensers, and heaters. Accessories include piping, shutoff valves, governors, and safety devices.

Function of the throttle valve

The throttle valve admits steam for starting, and stops the engine or turbine by shutting off steam flow. In addition to hand operation for starting, the throttle valve is equipped with a tripping device, which causes it to close in the event an undesirable or hazardous condition develops. Some of the conditions which cause the throttle valve to close are excessive speed, loss of lubricating oil pressure, excessively high temperatures, reduced vacuum, and excessive vibration. During start-up, the throttle valve is opened by hand until the unit attains design speed, at which time the governor starts to function,

controlling the flow of steam and thereby the speed. Once in operation the engine or turbine is controlled automatically to run at nearly constant speed. The simplest type of governor throttles the steam supply and so regulates the steam flow to meet the demand for power. To improve the efficiency at partial load, more elaborate control is used. Engine governors vary the length of time during which steam is admitted to the cylinder and in this way make use of the expansive properties of the steam. Turbine governors admit full line-pressure steam to an increasing number of nozzles as the demand for power increases. In hydraulic-control systems, a centrifugal governor senses the shaft speed and regulates the oil flow to or from a hydraulic cylinder, which in turn positions the steam-supply mechanism. Failure to maintain the oil pressure in this system will cause the throttle valve to close and shut down the unit.

STEAM TRANSPORTATION

Locomotives. Early in the 19th century, English experimenters attempted to use steam engines to run vehicles on tramways. In 1802 Richard Trevithick, an English engineer, obtained a patent for his high-pressure engine and its application for driving carriages but progress was retarded by a boiler explosion alleged to have been caused by an operator having shut off the safety valve.

In 1829 the "Rocket" locomotive, invented by George Stephenson, an English engineer, proved superior in a trial involving three locomotives. At this time development was rapid, and in 1833 Robert Stephenson, the son of George, and himself an engineer, built a locomotive incorporating many features that were retained in future developments. The Stephenson link was a single lever by means of which the locomotive could be made to run either forward or backward, and intermediate positions of the lever controlled the point of cutoff and gave high-steam economy at light loads. These features were incorporated in larger locomotives operating at high pressures and with superheated steam. The steam locomotive retained its importance in transportation until the mid-20th century when it was replaced by diesel-powered locomotives (see also RAILROADS AND LOCOMOTIVES).

Ship propulsion systems. The steam engine has made many contributions to industrial progress, but none has been more important than that made to navigation. Of the many experimenters, the American, John Fitch (1743–98), was one of the first to obtain some degree of success and public recognition. He built several steamboats between 1786 and 1798.

In 1803 Robert Fulton, an American mechanical engineer, successfully operated a steamboat on the Seine River in Paris, but it was rejected in both France and England. In 1807 Fulton's side-wheeler, "Clermont," made a successful trip up the Hudson River from New York to Albany, a distance of 150 miles (241 kilometres) in 32 hours. In the same year Robert L. Stevens, an American engineer and naval architect, and his father, John Stevens, an engineer, constructed a successful paddle-wheel steamer, which was put in service between New York and New Brunswick, New Jersey (see also SHIP).

Reciprocating engines. A reciprocating engine built by Boulton and Watt was used to power Fulton's "Clermont." Improved reciprocating engines were subsequently used, and with the introduction of watertube boilers, referred to as "safe boilers," higher pressures were used. Higher pressures and the use of condensers prompted the use of multiple-expansion engines as a means for reducing fuel consumption. Evaporators were used to condition the seawater for use in the boilers.

Geared turbines. When turbines became practical they were quickly applied to ship propulsion, because they offered a means of developing the increased power required, weighed less, and occupied less space. Gear reducers are required to adapt the economical high-speed turbines to the low optimum speed of the ship propellers.

Road vehicles. The use of steam to propel road carriages dates from 1770 when a French artillery officer built a massive machine. The boiler was inadequate and it stopped every few hundred feet because of a lack of steam. In 1801 Richard Trevithick, a rival of James Watt, built a steam-driven carriage that operated on the streets of London.

Between 1824 and 1828, several coal-burning steam-powered coaches capable of carrying 20 passengers and attaining a speed of 20 miles (32 kilometres) per hour were operated in and around London. Progress on the development of steam-driven carriages was essentially stopped when Parliament passed a law requiring that a self-propelled wagon be proceeded by a man carrying a red flag in the daytime and a red lantern at night. Steam-powered coaches

The invention of the pneumatic tire and the discovery of liquid fuel resulted in great interest in motor vehicles in France, Germany, and the United States. In 1897 F.E. and F.O. Stanley, American inventors in Newton, Massachusetts, completed their first steam-driven car. Their success in competition in Boston, Massachusetts, in 1898, prompted them to continue to develop and build the steam cars. Their car won a speed trial by travelling one mile (1,609 metres) in 2 min 11 sec, not only exceeding the speed of the gasoline powered De Dion car but also the world's record of one mile in 2 min 32 sec. Furthermore, the Stanley was the only car capable of negotiating a 30 percent grade. In 1906 a Stanley steamer attained a speed of 127 miles per hour in the Dewar Cup race on the beach at Ormond Beach, Florida.

By the early 20th century there were several makes of steam-powered automobiles including Eclipse, Grout, Lane, Locomobile, Mobile, Stanley, Stearns, and White. In 1900, 34 cars were exhibited in Madison Square Garden, New York and of this number seven were steam-powered. In 1906, by carrying some supplies and tools, and with even limited mechanical ability, steam cars could be successfully driven on extensive tours.

Much ingenuity and technical knowledge was applied to the early steam-powered automobiles. Steam pressures in excess of 1,000 pounds per square inch were developed in the boilers. Some designs employed flash-type once-through boilers consisting of a single coil in which water supply to the coil and fuel feed to the burner were proportioned to evaporate all of the water in the first part of the coil with the remainder used to superheat the steam. Early engines exhausted the steam to the atmosphere, and additional water had to be supplied to the boiler. Condensers were introduced about 1915 and the resulting water recycled to the boiler. Some manufacturers connected the engine directly to the rear drive axle, while others located the entire power plant under the hood and transmitted the power to the rear wheels.

The steam-driven automobile is characterized by speed, smooth operation, and hill-climbing ability. The time required to develop steam pressure in the boiler, however, is a disadvantage. Extensive application of automatic controls and other advances have decreased the amount of attention required by the operator. The internal-combustion engine, however, requires even less attention.

Interest in steam-driven automobiles has continued. Consideration has been given to the use of steam drive to reduce the air pollution caused by the exhaust from the internal-combustion engine. Working substances other than steam have been tested to eliminate the problem of freezing in cold climates (see also AUTOMOBILE).

STEAM POWER PLANTS

Early individual plants. A steam-power installation consists of boilers, engines, and auxiliary equipment that uses fuel as raw material and produces mechanical power. The general use of steam power in mills began in 1782 when Watt developed the double-acting rotative engine. One of the first of these engines was installed in the Albion Mill at Blackfriars, London. Its success revolutionized the flour-milling industry and rotative engines were soon in use in many industries.

The replacement of waterpower by steam power was a great advantage to mills and industries. It meant that the mill or factory could be located close to raw materials, transportation, and markets rather than where there was waterpower. Whereas the availability of water to turn the waterwheel frequently had limited the size and output of

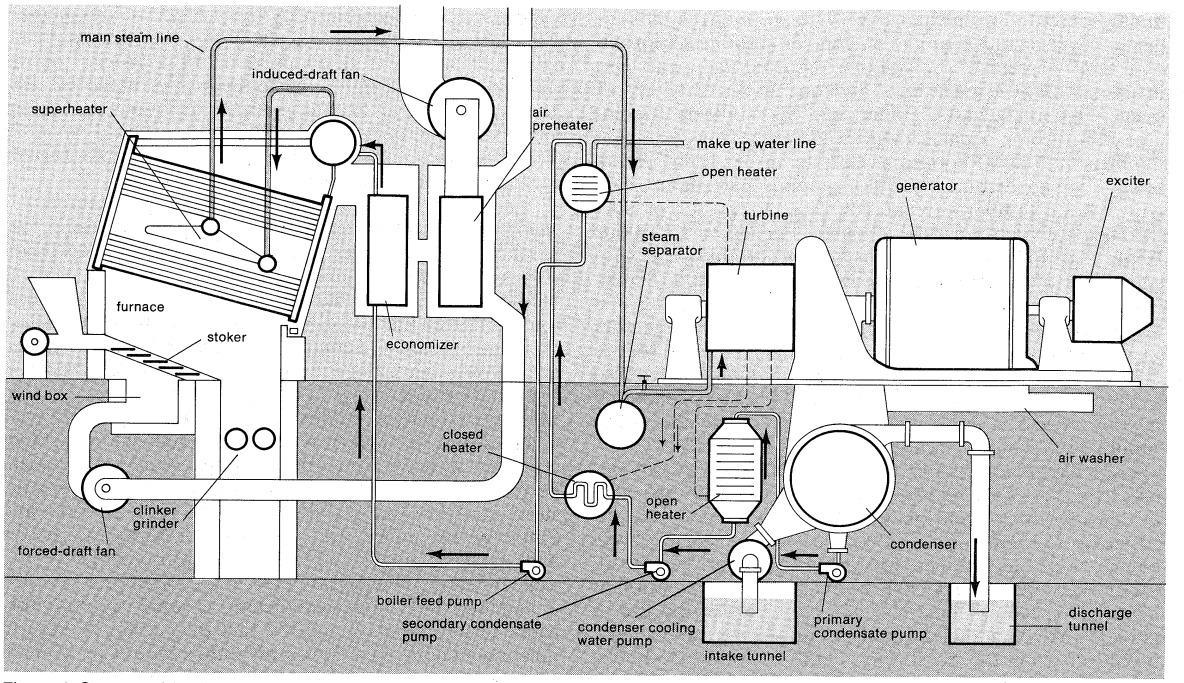


Figure 4: Steam turbine generating station with heat recovery equipment.

a mill, a steam plant could be built to meet the power requirements.

The use of steam to provide power for mills and factories was further enhanced by rapid development of the engine. By mid-19th century engine design was essentially standardized. Steam plants had become more compact, easier to operate, and more reliable and efficient.

Development of standardized design

The early use of steam engines in flour mills was extended to cotton gins, sugarcane mills, lumber mills, and grain-threshing machines. During the harvest season, portable threshing machines were moved from one farm to another; at first horse-drawn, they were later self-propelled. Steam-power tractors were used on large farms to cultivate the land and at harvest to pull combines which cut, threshed, cleaned and delivered the grain ready for market in one operation. Steam tractors were also used to replace horses and mules on construction work involving the moving of soil from excavations and for building dams. These steam-driven tractors were widely used from mid-19th century to mid-20th century, when they were largely replaced by internal-combustion engine drives.

Modern fossil-fuel plants. The power developed by steam engines was transmitted to the driven mill or machine by belts, shafting, or gears. The application of electricity created a new use for steam power. Steam engines were used to drive the electric dynamos. At first used only for lighting, electricity was soon found to be an excellent means of transmitting power. Power plants were increasingly used for generating electricity that could be readily transmitted to the mills and factories where it supplied motors that powered the machinery.

In 1880 Thomas A. Edison constructed a power station with a capacity of 7,200 lamps at 110 volts. An increasing demand for electricity resulted in the rapid development of a central steam plant. The service provided mill and factory operators with adequate power without requiring them to invest their capital in power plant equipment.

In a noncondensing power plant, typical of an early plant used for supplying steam for space heating and for industrial processing, in addition to electricity for power and lighting, the steam generated by use of fossil-fuel supplies the engine which generates electricity. The engine runs at constant speed and automatically admits sufficient steam to develop the required electric power. Exhaust steam from the engine flows to the heating system, process equipment and feedwater heater. When the exhaust steam is insufficient, steam is admitted from the boiler outlet through a pressure-reducing valve. Should the exhaust steam exceed the demand, the excess is wasted to the atmosphere through the back pressure regulating valve. A feedwater heater raises the temperature of the water supply to the boiler by the application of heat from the exhaust steam which would otherwise be wasted. The plant operates most efficiently when the electrical power demand on the engine results in an amount of exhaust steam equal to the combined heating requirements.

Central stations

The conventional central station, having only a limited use of steam for heating, must employ techniques for obtaining maximum amounts of electrical power from the fuel. These include (Figure 4) economizers and air preheaters to recover heat from the flue gases, superheaters to obtain a high-temperature steam supply to the turbines, condensers to reduce the turbine back pressure (pressure at the turbine exit), and heaters which use steam extracted from the turbines to raise the temperature of the condensate being returned to the economizers. The practice of using steam extracted from the various stages of turbines to heat the condensate is generally practiced in central stations.

Central station plants must be located accessible to the fuel and condensing-water supply. The electricity can be transmitted relatively long distances over high voltage lines. In a modern central power plant, coal received in railroad cars is dumped into hoppers and conveyed to overhead bunkers. The coal flows from bunkers to pulverizing mills and the resulting fine coal particles are then blown into the furnace. Radiant heat from the flame is absorbed by the water in the tubes which form the furnace walls and by the steam in the superheater. Additional heat is removed from the gases as they pass through the superheater-convection section of the boiler and the air preheater. Following the air preheater in the system, the precipitator removes dust particles from the gases before they pass through the fan and are discharged out the stack. Steam from the superheater drives the turbine and flows to the condenser where the latent heat is transferred to the condenser cooling water. A portion of the steam that flows through the turbine is, however, diverted from the condenser for use in the regenerative feedwater heating cycle. The returning condensate passes through the deaerating heater for removal of oxygen before being pumped through the economizer into the boiler. The electric power generated at 13,800 volts is stepped up to 114,000 volts for long-distance transmission.

An international survey of plants being designed and

under construction in the early 1970s dramatizes the tendency toward larger units for central stations. Approximately 40 percent of the turbines reported had a capacity in excess of 500 megawatts and one had a capacity of over 1,000 megawatts. Boiler size also reflects the increase in turbine capacity, with some exceeding 5,000,000 pounds (2,270,000 kilograms) of steam per hour. Steam is supplied to turbines at 3,500 pounds per square inch with a total temperature of 1,000° F and, after passing through the high-pressure turbine, is reheated to 1,000° F. Some of these plants are designed to produce one kilowatt-hour of electric energy on less than 8,000 British thermal units of fuel.

Nuclear power plants. Nuclear fuel is used in reactors to develop heat, producing steam that is then used to drive turbines. The reactors replace the fossil-fuel-fired boilers used in the conventional plant. The heat generated by the reactor is controlled so as to regulate the steam supplied to meet the requirements of the turbines.

In nuclear plants it has been quite difficult, if not impractical, to obtain highly superheated steam comparable to that used in conventional plants. In most cases saturated steam is used. This results in a decrease in the cycle efficiency of the turbines. In some cases separately fossil-fuel-fired superheaters have been used. Consideration is being given to methods of superheating steam by use of nuclear fuel. The nuclear plant produces less atmospheric pollution than the fossil-fuel plant but they both discharge large quantities of heat to the condenser water. The heat discharged is low level and not capable of generating power but adds thermal pollution to lakes and rivers. To eliminate this contamination, large cooling towers are being used to receive this rejected heat that is then dissipated to the atmosphere.

Almost all nuclear power central stations utilize the nuclear power to generate steam, which is then used to drive steam turbines. In some, steam is generated within the reactor (boiling-water reactor); in others, the water used to cool the reactor is pressurized to prevent boiling, and the steam generated in a separate heat exchanger (pressurized-water reactor). In western Europe some of the reactors are cooled by gases (high-temperature gas reactor), and gas turbines are used to generate electricity.

The nuclear-powered central station has proven to be a safe and dependable means of generating electricity. When compared with a conventional plant, however, costs are marginal, except in areas where fossil-fuel costs are high. Installations of nuclear-powered stations have been delayed by the reluctance of the public to accept the proposed sites, uncertainty regarding future fuel costs, and long-term delivery of equipment. Surveys show 11,000 megawatts of nuclear-powered central stations on order or being designed in the United States and 20,000 megawatts in western Europe. It has been estimated that of the 12,000 megawatts which will be installed in the United States by the year 2000, one-half will be nuclear powered, and the ratio will be even higher in western Europe.

OTHER APPLICATIONS OF STEAM POWER

Injector. The injector is a device that uses the energy derived by expanding steam through a nozzle to lift water and force it into the boiler from which the steam was obtained. The essential parts of an injector are the steam and suction jets, combining and delivery tube and ring, overflow and discharge tube. The velocity of steam discharged from the jet causes the water to be drawn into the combining chamber where the steam is condensed. The kinetic energy in the steam is transferred into the water which flows through the delivery tube into the boiler. Injectors, because of their compactness and simplicity, were widely used to supply water to small portable boilers, tractors, and locomotives.

Ejector. The ejector makes use of the kinetic energy developed by discharging steam through a nozzle to produce a vacuum. In modern power plants ejectors are used to remove noncondensible gases from surface condensers.

Ejectors are also used in processing operations where a product must be exposed to a vacuum.

Water can be refrigerated by the use of a steam ejector. The water is sprayed into a vacuum created by a steam ejector. This results in evaporation of a portion of the water and a lowering of the temperature of the remainder. For a chamber maintained at 0.12 pounds per square inch the temperature of the chilled water would be 40° F (4.5° C). This chilled water is then available for air conditioning or other cooling applications.

Pumps. The first practical application of steam power was for pumping water and this remains an important function. Steam pumps may be broadly classified as direct-acting and power-driven or rotary.

Direct-acting pumps have a steam piston and water piston or plunger that are directly connected by a rod. When the steam piston is actuated by steam, the water piston also moves, and water is admitted to and discharged from the water cylinder.

Piston or plunger pumps can be driven by engines or turbines, but in modern practice the high-speed turbines are used to drive centrifugal pumps. This arrangement permits a wide range of selection of turbine throttle, steam pressure, and temperature and exhaust pressure. Similarly, the centrifugal pump may be chosen to meet a wide range of capacity and pressure requirements.

Pulsometer. The pulsometer is an adaptation of Savery's pump and, though inefficient, is sometimes used because of its simplicity and the fact that it can be used in locations where other types of pumps would be difficult to install. The pulsometer consists of a casting containing two chambers, a steam-admission valve, two water-inlet and two water-outlet valves. Steam flows into one of these chambers, pushing out the water. When nearly empty the continued flow of steam through the outlet valve disturbs the remaining water, causing the steam to condense and produce a vacuum within the chamber. This vacuum shifts the steam supply to the second chamber where it forces the water out while the vacuum is causing the first chamber to refill. The cycle is repeated and an almost constant stream of water is discharged. The maximum suction lift is 25 feet (7.6 metres), but the practical range is 7 to 15 feet (2.1 to 4.5 metres). The required steam pressure is 50 percent higher than the total discharge pressure of the water.

Fire engines. Steam-driven pumps replaced the bucket brigades and hand-operated fire pumps commonly used in the 1830s and before. In 1829 the first heavy and cumbersome steam-operated fire pumps were developed in London. Similar units were used in New York in 1841. Some of these early fire engines were self-propelled, but it was found advantageous to use horses to pull the engines to the location of the fire and utilize the time enroute to develop steam pressure. Inventiveness and ingenuity were exercised in developing compact and light-weight equipment, and boilers that would steam rapidly.

Horse-drawn fire engines

By the end of the 19th century, these steam driven pumping engines were universally adopted in large cities. By 1925 internal-combustion engines were being used to propel the fire-fighting equipment to the fire and to operate the pumps. Thereafter, horse-driven steam engine pumpers were soon replaced.

Pile drivers. It is an age-old practice to support building structures on wooden poles (piles) driven into the ground. Steel and reinforced-concrete piling is now used. Before the use of steam, a heavy weight was lifted by various hand-powered devices and then dropped on the pile.

At first a steam engine was used to turn a winch that raised a hammer ranging in weight from 1,000 to 8,000 pounds (450–3,600 kilograms). This hammer was then dropped on the piling. Much faster but less forceful blows were obtained by a single-acting steam hammer. Steam pressure is applied directly to lift the weight and, upon release of the steam pressure, the weight falls by gravity and strikes the pile. Additional striking force is obtained by the double-acting steam hammer in which steam is applied not only to lift the weight but also to add to the force of gravity to intensify the striking force. Additional speed and striking force is obtained by use of the differential steam hammer which employs two pistons.

Since World War II diesel-powered hammers have been

developed. Although lacking in some of the versatility of the steam hammer, they are more easily transported because the boiler is eliminated.

Catapults. Steam is used as a means of accelerating airplanes to attain airborne speed when being launched from the deck of carriers. A tube or cylinder extending the length of the carrier deck is provided with a piston. A protrusion from the piston engages the airplane. When steam is introduced into the cylinder the piston moves forward accelerating the plane to 70 miles (112 kilometres) per hour. By an ingenious arrangement the slot in the cylinder, through which the protrusion from the piston moves, is sealed against steam leakage as the piston moves forward.

BIBLIOGRAPHY. JOHN BOURNE, *The Steam Engine* (1868), an outline of the early development of the steam engine, *A Treatise on the Steam Engine* (1869), a detailed treatise of early steam engine design; W.N. BARNARD, F.O. ELLENWOOD, and C.F. HIRSHFELD, *Elements of Heat-Power Engineering*, 3rd ed., pt. 1, *Thermodynamics and Prime Movers* (1926), on the theory and practical application of steam plant design; A. BRINTON CARSON, *Foundation Construction* (1965), on the theory and practice of foundation construction including the application of steam power to driving piling; T.S. DERR, *The Modern Steam Car and Its Background* (1932), a study of the application of the steam engine to power automobiles; H.W. DICKINSON, *Short History of the Steam Engine*, 2nd ed. (1963), a description of the use of steam power from earliest attempts to the present; A. FLEMING, *Wheels* (1960), a popular description of the use of the wheel including steam driven vehicles; G.R. FRYLING, *Combustion Engineering*, rev. ed. (1966), a reference book on fuel burning and steam generation; G.A. GAFFERT, *Steam Power Stations* (1946), a detailed treatise on the design of the modern steam plant with emphasis on central stations; C.F. HIRSHFELD and T.C. ULBRICHT, *Steam Power* (1916), on thermodynamic principles and their application to boilers, engines, and plant auxiliaries; R.K. EVANS, "Nuclear Power Reactors," *Power*, 109:1–24 (1965); G.H. PREBLE, *A Chronological History of the Origin and Development of Steam Navigation* (1883), a general text on the use of steam to power ships; C.W. PURSELL, JR., *Early Stationary Steam Engines in America* (1969), a history of the development and application of the steam engine in America and its influence upon industrial advancement; L.T.C. ROLT, *Thomas Newcomen* (1963), a biography of Newcomen and a description of his contribution to the development of the steam engine; BABCOCK AND WILCOX COMPANY, *Steam: Its Generation and Use*, 37th ed. (1963), a history of the development of the water tube boiler including technical data on modern central station boilers for various fuels; AMERICAN SOCIETY OF MECHANICAL ENGINEERS, *Steam Tables* (1967), complete data on steam as a working substance including pressure, temperature, enthalphy, and entropy; R.H. THURSTON, *A History of the Growth of the Steam Engine*, with a supplementary chapter by W.N. BARNARD (1939), a study of the early development of the steam engine to the early part of the 20th century; J.N. WALTON, *Doble Steam Cars, Buses, Lorries and Rail Cars* (1966), a description of the activities of the Doble Company in developing the steam engine for transportation; R.M. WIK, *Steam Power on the Farm* (1953), on the application of the steam engine to the farm, including its influence on economics and social life; E.B. WOODRUFF and H.B. LAMMERS, *Steam Plant Operation*, 3rd ed. (1967), a description of power plant boilers, engines, turbines, auxiliaries, and accessories, with an explanation of operating procedures to obtain optimum results.

(E.B.Wo.)

Steel Construction

Steel construction refers to a broad range of building construction in which steel plays the leading role. Most steel construction consists of large-scale buildings or engineering works, with the steel generally in the form of beams, girders, bars, plates, and other members shaped through the hot-rolled process. Despite the increased use of other materials, steel construction remained a major outlet for the steel industries of the U.S., U.K., U.S.S.R., Japan, West Germany, France, and other steel producers in the 1970s.

Early history. The history of steel construction begins paradoxically several decades before the introduction of the Bessemer and Siemens-Martin (open-hearth) processes made it possible to produce steel in quantities sufficient for structural use. Many of the problems of steel construction were studied earlier in connection with iron construction, which began with the Coalbrookdale Bridge, built in cast iron over the Severn River in England in 1777. This and subsequent iron bridge work, in addition to the construction of steam boilers and iron ship hulls, spurred the development of techniques for fabricating, designing, and joining. The advantage of iron over masonry lay in the much smaller amounts of material required. The truss form, based on the resistance of the triangle to deformation, long used in timber, was translated effectively into iron, with cast iron being used for compression members —*i.e.*, those bearing the weight of direct loading—and wrought iron being used for tension members—*i.e.*, those bearing the pull of suspended loading.

The technique for passing iron, heated to the plastic state, between rolls to form flat and rounded bars, was developed as early as 1800; by 1819 angle irons were rolled; and in 1849 the first I beams, 17.7 feet (5.4 metres) long, were fabricated as roof girders for a Paris railroad station.

Two years later Joseph Paxton of England built the Crystal Palace for the London Exposition of 1851. He is said to have conceived the idea of cage construction—using relatively slender iron beams as a skeleton for the glass walls of a large, open structure. Resistance to wind forces in the Crystal Palace was provided by diagonal iron rods. Two features are particularly important in the history of metal construction: first, the use of latticed girders, which are small trusses, a form first developed in timber bridges and other structures and translated into metal by Paxton; and second, the joining of the wrought-iron tension members and cast-iron compression members by means of rivets inserted while hot.

In 1853 the first metal floor beams were rolled for the Cooper Union Building in New York. In the light of the principal market demand for iron beams at that time, it is not surprising that the Cooper Union beams closely resembled railroad rails.

Introduction of structural steel

The development of the Bessemer and Siemens-Martin processes in the 1850s and 1860s suddenly opened the way to the use of steel for structural purposes. Stronger than iron in both tension and compression, the newly available metal was seized on by imaginative engineers, notably by those involved in building the great number of heavy railroad bridges then in demand in Britain, Europe, and the U.S.

A notable example was the Eads Bridge, also known as the St. Louis Bridge, in St. Louis (1867-74), in which tubular steel ribs were used to form arches with a span of more than 500 ft (152.5 m). In Britain, the Firth of Forth cantilever bridge (1883–90) employed tubular struts, some 12 ft (3.66 m) in diameter and 350 ft (107 m) long. Such bridges and other structures were important in leading to the development and enforcement of standards and codification of permissible design stresses. The lack of adequate theoretical knowledge, and even of an adequate basis for theoretical studies, limited the value of stress analysis during the early years of the 20th century, as occasional failures, such as that of a cantilever bridge in Quebec in 1907, revealed. But failures were rare in the metal-skeleton office buildings; the simplicity of their design proved highly practical even in the absence of sophisticated analysis techniques. Throughout the first third of the century, ordinary carbon steel, without any special alloy strengthening or hardening, was universally used.

The possibilities inherent in metal construction for high-rise buildings was demonstrated to the world by the Paris Exposition of 1889, for which Alexandre-Gustave Eiffel, a leading French bridge engineer, erected an openwork metal tower 300 m (984 ft) high. Not only was the height —more than double that of the Great Pyramid—remarkable, but the speed of erection and low cost were even more so, a small crew completing the work in a few months. At the same exhibition the huge, single-story Galerie des Machines, housing some of the exhibits, also contributed significantly to the advance of metal construction. Its roof was supported by latticed steel arch ribs spanning 114.3 m (375 ft).

The first skyscrapers. Meantime, in the United States another important development was taking place. In 1884–85 Maj. William Le Baron Jenney, a Chicago engineer, had designed the Home Insurance Building, ten stories high, with a metal skeleton. Jenney's beams were of Bessemer steel, though his columns were cast iron. Cast iron lintels supporting masonry over window openings were, in turn, supported on the cast iron columns. Solid masonry court and party walls provided lateral support against wind loading. Within a decade the same type of construction had been used in more than 30 office buildings in Chicago and New York. Steel played a larger and larger role in these, with rivetted connections for beams and columns, sometimes strengthened for wind bracing by overlaying gusset plates at the junction of vertical and horizontal members. Light masonry curtain walls, supported at each floor level, replaced the old heavy masonry.

Though the new construction form was to remain centred almost entirely in America for several decades, its impact on the steel industry was worldwide. By the last years of the 19th century, the basic structural shapes—I beams up to 20 in. (0.508 m) in depth and Z and T shapes of lesser proportions—were readily available, to combine with plates of several widths and thicknesses to make efficient members of any required size and strength. In 1885 the heaviest structural shape produced through hot-rolling weighed less than 100 pounds (45 kilograms) per foot; decade by decade this figure rose until in the 1960s it exceeded 700 pounds (320 kilograms) per foot.

Coincident with the introduction of structural steel came the introduction of the Otis electric elevator in 1889. The demonstration of a safe passenger elevator, together with that of a safe and economical steel construction method, sent building heights soaring. In New York the 286-ft (87.2-m) Flatiron Building of 1902 was surpassed in 1904 by the 375-ft (115-m) Times Building (renamed the Allied Chemical Building), the 468-ft (143-m) City Investing Company Building in Wall Street, the 612-ft (187-m) Singer Building (1908), the 700-ft (214-m) Metropolitan Tower (1909) and, in 1913, the 780-ft (232-m) Woolworth Building.

The rapid increase in height and the height-to-width ratio brought problems. To limit street congestion, building setback design was prescribed. On the technical side, the problem of lateral support was studied. A diagonal bracing system, such as that used in the Eiffel Tower, was not architecturally desirable in offices relying on sunlight for illumination. The answer was found in greater reliance on the bending resistance of certain individual beams and columns strategically designed into the skeleton frame, together with a high degree of rigidity sought at the junction of the beams and columns. With today's modern interior lighting systems, however, diagonal bracing against wind loads has returned; one notable example is the John Hancock Center in Chicago, where the external X-braces form a dramatic part of the structure's facade.

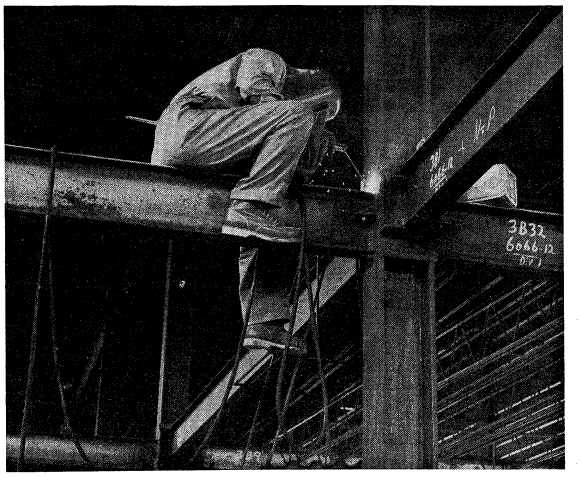
World War I brought an interruption to the boom in what had come to be called skyscrapers (the origin of the word is uncertain), but in the 1920s New York saw a resumption of the height race, culminating in the Empire State Building in 1931. The Empire State's 102 stories (1,250 ft. [381 m]) were to keep it established as the highest building in the world for the next 40 years. Its speed of erection demonstrated how thoroughly the new construction technique had been mastered. A depot across the bay at Bayonne, N.J., supplied the girders by lighter and truck on a schedule operated with military precision; nine derricks powered by electric hoists lifted the girders to position; an industrial-railway setup moved steel and other material on each floor. As the building rose, the hoist engines were lifted outside the building to successively higher stages. Initial connections were made by bolting, closely followed by riveting, followed by masonry and finishing. The entire job was completed in one year and 45 days.

The Empire State Building

The worldwide depression of the 1930s and World War II provided another interruption to steel construction development, but at the same time the introduction of welding (see photograph) to replace rivetting provided an important advance.

Joining of steel parts by metal arc welding had been successfully achieved by the end of the 19th century and was used in emergency ship repairs during World War I, but its application to construction was limited until after World War II. Since then it has gained steadily, especially in shop fabrication. Another advance in the same area has been the introduction of high-strength bolts to replace rivets in field connections.

By courtesy of American Iron and Steel Institute



Steel frame being welded.

Since the close of World War II, research in Europe, the U.S., and Japan has greatly extended knowledge of the behaviour of different types of structural steel under varying stresses, including those exceeding the yield point, making possible more refined and systematic analysis. This in turn has led to the adoption of more liberal design codes in most countries, permitting freer, more imaginative design made possible by so-called plastic design and by the use of the higher strength materials, with significant savings in cost. The introduction of the computer, by short-cutting tedious paperwork, made further advances and savings possible.

Resistance to earthquakes. While its interior was gutted by fire, the 16-story Call Building in San Francisco, erected in 1898, withstood the 1906 earthquake and was subsequently rehabilitated. Nevertheless, the construction of much higher buildings in areas subject to strong seismic forces was postponed until midcentury. That they can be designed to resist strong ground motion without damage was demonstrated by the 43-story Latino-Americano Tower during the 1957 earthquake in Mexico City. Similar performance was observed with comparable buildings in Los Angeles in 1971.

Stiff-wall framing. The 1950s saw a new development in steel-skeleton design that was, paradoxically, a return to an older, presteel building system—that of bearing walls. Exterior wall skeletons were built of closely spaced steel members arranged and proportioned with sufficient rigidity to function as the stiff sides of a box, capable of carrying all interior loadings, and resisting wind forces. The advantage of such an arrangement in modern office buildings is that it provides column-free interior space around the central service core. The office space can be organized by light partition walls to suit individual tenants' requirements.

The first such structure built was the IBM Building, erected in Pittsburgh in 1962. Its novel diamond-pattern exterior walls were designed to deliver their entire load to a pair of trellis-like supports under the centre of each exterior wall. Three different grades of structural steel were used on the diamond frame so that the wide range in stresses could be borne without disturbing the uniformity of the pattern. The structural members, which had a yield strength ranging from 36,000 to 100,000 lb/sq in. (2,531 to 7,030 kg/sq cm), were covered with stainless steel.

The success of the IBM Building led to the adoption of

the same bearing-wall principle elsewhere, most spectacularly in the 110-story twin towers of the World Trade Center in New York City, completed in the early 1970s. Unlike the Empire State Building, which they exceed in height by 100 ft (30 m), the towers maintain the same cross section (209 sq ft [19.4 sq m]) from ground level to roof. The 350-ft (107-m) TV antenna mast on the north tower carries the total height to just short of a third of a mile. Stiffness in the exterior walls is provided by closely spaced columns rigidly connected to deep spandrel beams, which are located between window units at successive floor levels. Steel in the columns ranges from ordinary grade at the top to heat-treated low-alloy at ground level, where extraordinary compressive strength is needed. Steel also is basic to the floor construction; corrugated steel forms, supported on steel joists spanning from the exterior walls to the interior service core framing, support the poured concrete slab. The slabs are mechanically anchored to the joists to act integrally with the joists in supporting the floor loads.

Specialized structures. Improved research facilities available since World War II have expanded and refined the understanding of steel as a structural material; combined with improved fabrication techniques and the computer, this has resulted in a proliferation of unusual applications of steel in construction. Three notable buildings at world's fairs of the 1950s and 1960s provide examples: the U.S. Pavilion at the 1958 Brussels, Belgium, exposition, in which the roof was supported by a wire rope system resembling the spokes of a wheel, stretched in tension between a ring hub and a rim made as a circular latticed girder; the 600-ft- (183-m-) high Space Needle of the Century 21 Exposition in Seattle, Washington, in 1962, in which three legs, built up of heavy rolled beam sections, surrounded a central core, providing access to a revolving restaurant and observation level at the top; and the 275-ft-diameter geodesic, spherical dome enclosing the U.S. Exhibition Building at Expo 67 in Montreal.

Other examples of unusual steel construction include the movable 575-ton (575,000-kg) feeder platform supported on cables 500 ft (152.4 m) above the dished reflector antenna at the ionospheric observatory in Arecibo, Puerto Rico; the 2.7-mile (4.3-kilometre) Sandia Peak Aerial Tramway near Albuquerque, New Mexico, with its 7,720-ft (2,353-m) catenary (suspension cable) between Tower No. 2 and the summit; the 630-ft- (192-m-) span Gateway Arch at St. Louis; and the Vertical Assembly Building at Cape Kennedy, Florida, used for assembly of space rockets, and enclosing more space than any other scientific building in the world.

BIBLIOGRAPHY. C. COWPER, *The Building Erected in Hyde Park for the Great Exhibition of the Works of Industry* (1851), an account of the construction of the Crystal Palace, including illustrative plates; G. EIFFEL, *Travaux Scientifique Exécutés à la Tour de Trois Cents Metres* (1900), an account of the building of the Eiffel Tower and the meteorological observations made from it; A.E.R. DE JONGE, *Riveted Joints* (1945), annotated bibliography of literature covering research on the behaviour of mechanically fastened structural joints; R. FLEMING, *Wind Stresses on Buildings* (1930), a discussion of the intensity of wind pressure on buildings and a review of various systems of wind bracing employed in multistory buildings; F. BLEICH and L.B. RAMSEY, *Buckling Strength of Metal Structures* (1952), a monograph dealing with the derivation of design criteria for compression elements and the stability of steel frames; S.P. TIMOSHENKO and J.M. GERE, *Theory of Elastic Stability* (1961), a textbook dealing with the stability of structures and their components subject to compressive stress; J.F. BAKER, M.R. HORNE, and J. HEYMAN, *Steel Skeleton*, 2 vol. (1954–56), a presentation of contemporary British practice, vol. 2 deals exclusively with plastic design; L.S. BEEDLE, *Plastic Design of Steel Frames* (1958), a textbook documenting design provisions based upon the postyield behaviour of continuous steel framing; Z.S. MAKOWSKI, *Steel Space Structures* (1965), examples of single- and double-layer grid construction, braced barrel-vault-type buildings, framed domes and suspended roof structures; generously illustrated with an extensive bibliography; W. MCGUIRE, *Steel Structures* (1968), a textbook covering designs of various kinds of steel structures and their component parts; E.H. and C.N. GAYLORD, *Structural Engineering Handbook* (1968), a comprehensive reference text covering all phases of structural engineering; *Bauen in Stahl 2*, (1962), a description, in French and German, of 147 structures of all types erected in 15 countries since 1956; profusely illustrated, presenting a well-rounded report on modern steel construction.

(T.R.H.)

Steele, Sir Richard

English essayist, dramatist, journalist, and politician, Richard Steele is best known today as principal author, with his friend Joseph Addison, of the essay periodicals *The Tatler* and *The Spectator;* but his stage comedies had a great effect on the sensibility of the 18th century, all but one of them retaining a considerable popularity for several decades and his work as a political pamphleteer and editor on behalf of the Whigs during the last years of Queen Anne's reign brought him national importance and a knighthood. Both as man and writer Steele is one of the most attractive figures of his time, much of his writing—easy, rapid, slipshod, but deeply sincere—reflecting his personality. "There appears in his natural temper," wrote his contemporary, the philosopher George Berkeley, "something very generous and a great benevolence to mankind." An emotional, impetuous, good-natured, and idealistic man, he always found it easier to get money than to keep it, and his career can be seen as in part shaped by the constant need to keep his head above the waters of debt.

By courtesy of the National Portrait Gallery, London



Steele, oil painting by Sir Godfrey Kneller, 1711. In the National Portrait Gallery, London.

Early life and works. Steele was born in Dublin early in 1672. His father, an ailing and somewhat ineffectual attorney, died when Steele was about five (a moving passage in *Tatler* No. 181 records what were probably his memories of this); and the boy was taken under the protection of his uncle Henry Gascoigne, confidential secretary to the Duke of Ormonde, to whose bounty, as Steele later wrote, he owed "a liberal education." He was sent to study at Charterhouse in 1684 and to Christ Church, Oxford, in 1689. At Charterhouse he met Addison, and thus began one of the most famous and fruitful of all literary friendships, which lasted until disagreements (mainly political) brought about a cooling and a final estrangement shortly before Addison's death in 1719. Steele moved to Merton College in 1691 but, caught up with the excitement of King William's campaigns against the French, left in 1692 without taking a degree to join the army. No doubt he intended to make this his career, or one of his careers; he was commissioned in 1697 and promoted captain in 1699, but, lacking the money and connections necessary for substantial advancement, he left the army in 1705.

Beginning of Steele's friendship with Addison

Meanwhile, he had embarked on a second career, as a writer. Perhaps partly because he gravely wounded a fellow officer in a duel in 1700 (an incident that inspired a lifelong detestation of duelling), partly because of sincere

feelings of disgust at the "irregularity" of army life and his own dissipated existence, he published in 1701 a moralistic tract, "The Christian Hero," of which ten editions were sold in his lifetime. This tract, asserting "that no Principles but those of Religion are Sufficient to make a Great Man," inevitably led to Steele's being accused of hypocrisy and mocked for the contrast between his austere precepts and his genially convivial practice; but, even if Steele occasionally deluded himself about his moral worth, he was unquestionably sincere, and this early work foreshadowed much of his later writing. Sensitive to the jeers of his fellow officers, however, Steele turned his literary talents to a very different kind of production and in the same year (1701) wrote his first comedy, *The Funeral*. Performed at Drury Lane "with more than expected success," it made his reputation and helped to bring him to the notice of King William and the Whig leaders. Late in 1703 he followed this with his only stage failure, *The Lying Lover*, which ran for only six nights, being, as Steele said, "damned for its piety." Sententious and ill-constructed, with much priggish moralizing, it is nevertheless of some historical importance as one of the first "sentimental comedies" and is of interest in that there is much of Steele in it—an emotional protest against duelling, a solemnly tender regard for "the fair sex," and a wry allusion to the gullible author's having recently squandered a good deal of money in an alchemical project. Steele was, indeed, always ready to entertain moneymaking ventures, having an indefatigable faith that something would turn up, and from 1717 he was to spend much time, energy, and money on a scheme for transporting live fish in specially constructed boats.

First comedy

A third play, *The Tender Husband*, with which Addison helped him (1705), had some success, but Steele continued to search for advancement and for money. In the next few years he secured various minor appointments, such as that of gentleman-waiter to the Queen's husband, Prince George, and in 1705 he married a widow, Margaret Stretch, who owned considerable property in Barbados. Little is known of this marriage, but it is difficult not to believe that Steele's motives were at least partly mercenary. Almost immediately the estate was entangled in his debts (he lost two actions for debt, with damages, in 1706), but, when, late in 1706, she conveniently died, Margaret left her husband with a substantial income. Steele's second marriage, contracted within a year of Margaret's death, was very different, though again he married a woman of means: Mary Scurlock, his "Dear Prue," was completely adored by Steele, however much he might at times neglect her. His hundreds of letters and notes to her provide a vivid revelation of his personality during the 11 years of their marriage. Having borne him four children (of whom only the eldest, Elizabeth, long survived Richard), she died, during pregnancy, in 1718.

Mature life and works. Steele's other important appointment in the early part of Queen Anne's reign was that of gazetteer—writer of *The London Gazette*, the official government journal. Although this reinforced his connection with the Whig leaders, it gave little scope for his artistic talents, and, on April 12, 1709, he secured his place in literary history by launching the thrice-weekly essay periodical *The Tatler*. Writing under the name (already made famous by the satirist Jonathan Swift) of Isaac Bickerstaff, Steele created the mixture of entertainment, and instruction in manners and morals, that was to be perfected in *The Spectator*. "The general purpose of the whole," wrote Steele, "has been to recommend truth, innocence, honour, and virtue, as the chief ornaments of life"; and here, as in the later periodical, can be seen his strong ethical bent, his attachment to the simple virtues of friendship, frankness, and benevolence, his seriousness of approach tempered by the colloquial ease and lightness of his style. Addison contributed some 46 papers and collaborated in several others, but the great bulk of the 271 issues were by Steele himself; and, apart from bringing him fame, it brought a measure of prosperity. The exact cause of *The Tatler*'s demise is uncertain, but probably the reasons were mainly political: in 1710 power had shifted to the Tories and Steele, a Whig, had lost his gazetteership and had come near to losing his post of commissioner of stamps. *The Tatler* had contained a good deal of political innuendo, some of it aimed at Robert Harley, the Tory leader, himself, and Harley may well have put pressure on Steele to discontinue the paper.

Launching of *The Tatler*

The Tatler's greater successor, first appearing on March 1, 1711, was avowedly nonpolitical and was enormously successful. *The Spectator* was a joint venture; Steele's was probably the more original journalistic flair, and he evolved many of the most celebrated ideas and characters (such as Sir Roger de Coverley), although later Addison tended to develop them in his own way. Steele's attractive, often casual style formed a perfect foil for Addison's more measured, polished, and erudite writing. Of the 555 daily numbers, Steele contributed 251 (though about two-thirds made up from correspondents' letters).

Of Steele's many later ventures into periodical journalism, some, such as *The Englishman*, were mainly politically partisan. *The Guardian* (to which Addison contributed substantially) contains some of his most distinguished work, and *The Lover* comprises 40 of his most attractive essays. Other, short-lived, periodicals, such as *The Reader*, *Town-Talk*, and *The Plebeian*, contain matter of considerable political importance. Steele became, indeed, the chief journalist of the Whigs in opposition (1710–14), his writings being marked by an unusual degree of principle and integrity. His last periodical and his last extended literary work was *The Theatre*.

Steele's political writings had stirred up enough storms to make his career far from smooth. He resigned as commissioner of stamps in 1713 and was elected to Parliament; but, as a consequence of his anti-Tory pamphlets "The Importance of Dunkirk Consider'd" and "The Crisis," he was expelled from the House of Commons for "seditious writings." Calmer weather, however, and rewards followed on George I's accession: Steele was appointed to the congenial and fairly lucrative post of governor of Drury Lane Theatre in 1714, knighted in 1715, and re-elected to Parliament in the same year.

Last years

Steele's health was gradually undermined by his cheerful intemperance, and he was long plagued by gout. Nevertheless, he busied himself conscientiously with parliamentary duties and, more erratically, with his part in the management of Drury Lane. One of his main contributions to that theatre's prosperity was his last and most successful comedy, *The Conscious Lovers*—one of the most popular plays of the century and perhaps the best example of English sentimental comedy.

In 1724 Steele retired to his late wife's estate in Wales and began to settle his debts. His closing years were quiet; but his health continued to deteriorate, and, on September 1, 1729, the genial and charming writer, whose burly figure, brown face, and brilliant eyes had long been part of the London scene, died at Carmarthen.

MAJOR WORKS

PERIODICAL ESSAYS: in *The Tatler* (1709–11); *The Spectator* (1711–12); *The Guardian* (1713); *The Englishman* (1713–14; second series, 1715); *The Lover* (1714); *The Reader* (1714); *Town-Talk* (1715–16); *The Theatre* (1720).

POLITICAL WRITINGS: "The Importance of Dunkirk Consider'd" (1713); "The Crisis" (1714); "Mr. Steele's Apology for Himself and His Writings" (1714). Some of Steele's periodical journalism was purely political; many of his essays and pamphlets were gathered in *Political Writings* (1715).

MISCELLANEOUS: "The Christian Hero" (1701).

PLAYS: *The Funeral* (performed 1701, published 1702); *The Lying Lover* (performed 1703, published 1704); *The Tender Husband* (1705); *The Conscious Lovers* (performed 1722, published 1723).

BIBLIOGRAPHY. There is no systematic bibliography of Steele's works. Aitken's *Life* has a bibliography, as does Blanchard's edition of *The Christian Hero* (see below). The major collections of Steele material are in the British Museum and the Bodleian Libraries; there are many documents relating to him in the Public Record Offices of London and of Dublin and in Somerset House, London. A definitive edition of the works is in progress. RAE BLANCHARD has edited *The Christian Hero* (1932), the *Correspondence* (1941, reprinted 1968 with additional letters), *Tracts and Pamphlets* (1944, reprinted 1967), *Occasional Verse* (1952), *The Englishman* (1955), and *Periodical Journalism, 1714–16* (1959). JOHN

LOFTIS has edited *The Theatre, 1720* (1962); and SHIRLEY STRUM KENNY, *The Plays* (1971). DONALD F. BOND has edited *The Spectator*, 5 vol. (1965), and is preparing an edition of *The Tatler;* the latter was edited by G.A. AITKEN, 4 vol. (1898–99). The most fully documented biography is G.A. AITKEN, *The Life of Richard Steele*, 2 vol. (1889); more up-to-date are the biographies by WILLARD CONNELY, *Sir Richard Steele* (1934); and CALHOUN WINTON, *Captain Steele: The Early Career of Richard Steele* (1964) and *Sir Richard Steele, M.P.: The Late Career* (1970). See also JOHN LOFTIS, *Steele at Drury Lane* (1952). Critical studies include WALTER J. GRAHAM, *The Beginnings of English Literary Periodicals: A Study of Periodical Literature, 1665–1715* (1926); F.W. BATESON, *English Comic Drama, 1700–1750* (1929); and A.R. HUMPHREYS, *Steele, Addison, and Their Periodical Essays* (1959).

(R.P.C.M.)

Steel Production

The name steel is applied to an enormous variety of alloys in which the principal ingredient is the chemical element iron. Carbon is the most important alloying element, although other alloying elements may be present in higher concentrations. The carbon content can range from approximately 0.03 percent to 1.5 percent and in a few cases may go as high as 2.25 percent. Iron alloys containing less than 0.03 percent carbon are called irons, and those containing carbon above 1.5 percent are usually called cast irons or, in the crude form, pig iron.

Three classes of steels

Broadly, there are three commercial classes of steels: carbon steels, low-alloy steels, and high-alloy steels. In addition to carbon, the carbon steels contain up to 1.65 percent of manganese and possibly small quantities of silicon, aluminum, copper, and other elements. The low-alloy steels also include a total of from 1 to 5 percent of common alloying elements such as nickel, chromium, molybdenum, tungsten, titanium, niobium (columbium), and often other elements. The high-alloy steels contain over 5 percent of at least one alloying element, principally chromium, nickel, manganese, molybdenum, or tungsten. The stainless steels constitute a major class of the high-alloy steels. Usually a steel with more than 5 percent chromium is called a stainless steel. By common practice, an alloy containing less than 50 percent iron is not considered to be a steel. Carbon steels are used extensively for automobile bodies, appliances, machinery, ships, containers, and the structure of buildings. The low-alloy steels are used most often in applications in which special properties are needed, but cost is of considerable importance. Examples are forged products for highly stressed parts of machines, aircraft landing gear, shafts, hand tools, gears, and high-strength structural steels. The high-alloy steels are used for their unusual properties: typical are the stainless steels, used when resistance to corrosion and oxidation and good appearance are required; *i.e.*, as parts of jet engines, chemical equipment, tableware and cooking utensils, and cutting tools.

The world production of all kinds of raw steel in the second year of the present decade was over 642,000,000 tons, (excepting Tables 1 and 2; and Fig. 13 amounts in tons refer to short, or net, tons of 2,000 pounds), with the United States, the Soviet Union, and Japan together producing almost 60 percent of the total, excluding that produced by China. Great quantities of iron ore, coal, fuel oil, electricity, natural gas, limestone, oxygen, and air are needed as raw materials by the industry. In the United States an average of 1,750 pounds of iron ore, 1,050 pounds of steel scrap, 1,600 pounds of coal, 800 pounds of limestone, 150 kilowatt-hours of electrical energy, and 16,000 gallons of water are required to produce one ton of raw steel. The equivalent amounts in metric units per metric ton of steel are 875 kilograms of ore, 525 kilograms of steel scrap, 800 kilograms of coal, 400 kilograms of limestone, and 66 cubic metres of water. The steel industry of a modern industrial nation constitutes an important segment of the economy.

Properties of iron

The physical and chemical properties of the element iron make it one of the most interesting in the periodic table (see TRANSITION ELEMENTS AND THEIR COMPOUNDS). The metal forms solid and liquid alloys with most of the metallic elements and forms compounds with most of the nonmetallic elements; those of particular interest in steel production are the oxides, carbides, nitrides, and sulfides of iron. Iron occurs in two crystallographic forms: so-called face-centred cubic structure, in which the atoms are arrayed at the corners of a cube and at the centre of each face of the cube; and body-centred-cubic structure, in which the atoms are located at the corners of the cube and one at the centre of the cube. The tremendous variety possible in the properties of steels arises principally because of the effects of atoms of elements such as manganese, chromium, nickel, molybdenum, and vanadium substituting for atoms of iron in the crystal lattice and by elements such as carbon and nitrogen, which can lodge interstitially in holes between the atoms of iron. The importance of these interstitial elements can be seen from the fact that in many commercial steels the carbon content must be controlled within 0.01 percent and nitrogen within 0.001 percent.

The general course of the chemical changes that iron undergoes in the conventional conversion of iron ore to steel is as follows. Oxygen is removed from the ore by reactions with carbon, carbon monoxide, and hydrogen. The resulting metallic iron is melted, and in this state it will dissolve carbon and other impurities. Most of these impurities are removed from the liquid iron in the steelmaking furnace by reacting the metal with oxygen; the process is controlled, however, to prevent the iron from being oxidized again to the oxide. After the liquid metal has been refined, the necessary alloying elements are added, and the steel is solidified. In the subsequent rolling, forging, pressing, and processing of the solid steel, care is again taken to prevent undue oxidation of the steel and loss of carbon from the surface.

This article is divided into the following sections:

I. History
  The Iron Age
  Mass production of steel
II. Modern steel technology
  Raw materials
  The basic steelmaking process
  Types of steelmaking processes
  Solidification and casting
  Rolling and working of steels
  Control of properties
III. Properties and uses of carbon and alloy steels
  Carbon steels
  Alloy steels
IV. Steel metallurgy
  Structural forms of iron
  Phases in steel
  Effects of carbon content on structure and properties of steel
  Surface hardening
  Research and development
V. World steel production

I. History

THE IRON AGE

Early uses of iron. Iron is found in the metallic state only in small quantities and only in a few places in the Earth's crust. Principally, it is associated with basaltic rock on the west coast of Greenland. It has also been found in very small quantities in a few places in middle Europe; in the Don district, in the southern U.S.S.R.; in Scotland; and in Canada. Occasional pieces of iron of meteoric origin also have been found in many places in the world. It seems probable that iron objects made from these sources entered ancient commerce before 4000 BC, though there is no evidence for this, because iron corrodes relatively rapidly, so that iron objects disappeared, whereas those of gold and silver have survived until today. Most probably, iron came into more general use in primitive times through the accident of iron-bearing stones being exposed to the hot charcoal in a fire. The action of the charcoal converted the stones to spongy pieces of iron, which were later recovered from the residue of the fire. The association of certain kinds of rock with charcoal and fire and the presence of metallic iron in the ashes may have been discovered at various times in many places. It would seem, however, that development

Native iron

of the technology by which iron was forged, welded, and worked into useful objects took many centuries and occurred in only a few locations. It is known that iron objects were used in ancient civilizations in China, India, the Near East, Egypt, and Greece. Iron is mentioned a number of times in the Old Testament, the earliest references being in Genesis and Deuteronomy. It is also mentioned in Homer's *Iliad*.

The Catalan forge. Iron weapons and implements were apparently made by hammering and welding sponge iron by 1000 BC. The sponge iron was produced in shallow, clay-lined, bowl-shaped hearths into which were placed charcoal and the iron-bearing stones. Intense heat was generated by a jet of air directed onto the charcoal through a nozzle, or tuyere. The Catalan forge, *c.* AD 700

From H.E. McGannon (ed.), *Making, Shaping and Treating of Steel*, 9th ed., copyright 1971 by United States Steel Corporation

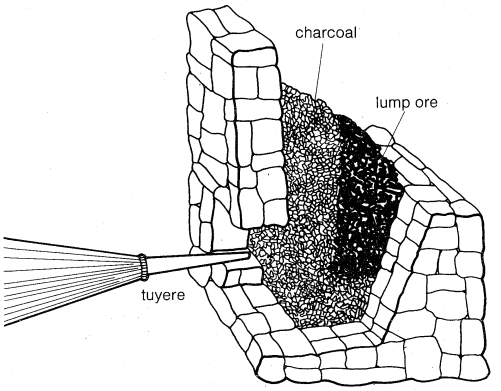


Figure 1: Catalan hearth or forge used for smelting iron ore until relatively recent times. The method of charging fuel and ore and the approximate position of the nozzle supplied with air by a bellows are shown (see text).

(see Figure 1), was a short shaft furnace for producing sponge iron directly from iron ore. A blast of air produced with a stream of falling water or with a bellows powered by hand, animal, or waterpower was introduced into the charcoal at the bottom of the hearth. It was possible to obtain much higher temperatures with this type of air supply than could be reached with more primitive furnaces that used natural draft or the force of the prevailing wind directed into the furnace opening. Though the temperature reached was not hot enough to melt the sponge iron (the melting point of iron is 1,538° C [2,800° F]), it did melt the impurities in the ore and the ash in the charcoal to form a liquid slag. After sufficient sponge iron was formed, the hot lump was withdrawn from the hearth and hammered vigorously to expel most of the slag, to weld the pieces of sponge together, and to shape the resulting piece of metal into a rod, bar, or plate. The metal of this type of product, called wrought iron, was relatively pure, having no more than about 0.1 percent impurities alloyed with the metal and up to 3 percent entrapped slag (see IRON MINING AND PROCESSING).

Production of steel objects. By 1200 BC it was possible to produce steel objects that could be hardened by quenching them in water or oil to develop a hard edge for cutting, picking, or chiselling. This required that the carbon content of wrought iron be increased by holding the sponge iron in contact with hot charcoal or other carbonaceous material away from the air blast for several hours. Alternatively, a wrought-iron object would be packed in a bed of hot charcoal in a clay box or jar. The whole assembly would be heated above a red heat for days so the carbon content of the metal surface was raised sufficiently to obtain the desired hardness when the object was quenched. This latter technique is called the cementation process, and the product is cemented steel.

Development of the blast furnace. With the passage of time, the sides of the smelting hearth grew higher and the bed of ore particles and charcoal deeper. Greater production and economy in fuel consumption was realized, because the gases from the hot region passed upward through the bed to preheat and prereduce the ore. As a consequence, the iron absorbed more carbon, so that it liquified and trickled down into the hearth to form a puddle of impure iron. This iron has a melting point in the range of 1,150° to 1,200° C (2,100 to 2,200° F) and could be cast readily into convenient shapes, such as kettles and plates. It is probable that cast iron was first produced in China during the 6th century AD. The tall furnace, known as a blast furnace, was in general use throughout Europe by the 14th century and was introduced into Britain late in the 15th. The first blast furnace erected on the North American continent was in Falling Creek, Virginia, in 1620. Early furnaces were about 20 to 26 feet (six to eight metres) high. The fuel was charcoal, and the air blast was supplied by bellows driven by a waterwheel. Sixteenth-century furnaces produced about one ton (900 kilograms) of iron a day. By the middle of the 17th century some English furnaces produced two to three tons a day. The product of the blast furnace was called pig iron, because liquid metal flowing from the furnace was conveyed into a sand bed in which was formed a long trough to lead the metal into numerous smaller troughs. The feeder trough was likened to a sow and the side bars to piglets, hence the name pig iron. As a result of the extensive use of charcoal in the production of pig iron, wrought iron, and steel, much of the forests in Great Britain had been destroyed by the middle of the 17th century, and until the end of the 18th century Great Britain had to import iron from Sweden, Russia, and her American colonies. It was only after the successful commercial use of coke in the blast furnace in 1740 that the production of pig iron returned to Great Britain.

Pig iron

In the 18th century and the early part of the 19th century, blast furnaces grew rapidly in size and capacity. With coke as the fuel, blowing capacities greater than were available from water-driven bellows were required. In 1776 James Watt's steam engine was applied to pump air for a blast furnace in England, and the replacement of charcoal by coke proceeded rapidly thereafter. An 1816 invention by which the hot waste gases from an industrial process were used to preheat incoming air for combustion was applied to the iron blast furnace in 1828 in Scotland. There resulted a saving of at least one-third in the fuel required to smelt iron.

Indirect process for making wrought iron. Although pig iron was cheap to produce and easy to cast, it was brittle, because it contained at least several percent of carbon and appreciable concentrations of silicon, manganese, phosphorus, and sulfur. Wrought iron could be produced from it by an indirect process, to which there were so many variations that only its general character can be described here. Pigs of cast iron from the blast furnace were melted gradually on a hearth filled with charcoal. The liquid collected in a puddle, which was exposed to a blast of air from a tuyere. The air oxidized the silicon and carbon out of the metal, and, without these elements, the melting point of the metal rose. As a result, the puddle became a pasty lump. When a large enough ball of metal and slag was collected, the mass was removed from the furnace and hammered to obtain a ball of wrought iron. Early in the 17th century a brick roof was put over the hearth, and heat for puddling the iron was provided from a combustion chamber that was separated from the hearth by a brick wall. The heat and flame from this combustion chamber were reflected, or reverberated, from the roof down upon the charge in the hearth. This type of furnace has come to be known as a reverberatory furnace. In 1784 Henry Cort of England was granted a patent for an improved method of producing puddled wrought iron in this type of furnace. During the latter part of the 18th century the use of wrought iron for machinery and implements expanded rapidly. In 1777 the first bridge made of wrought iron was erected in Coalbrookdale, Shropshire.

Reverberatory furnace

Crucible process. The crucible process for making steel reappeared in 1740. Pieces of cemented wrought iron and slag-making ingredients were charged into a fireclay crucible (see Figure 2). The crucible was tightly sealed and then placed in a heating furnace. In this way the fuel could be burned under highly oxidizing condi-

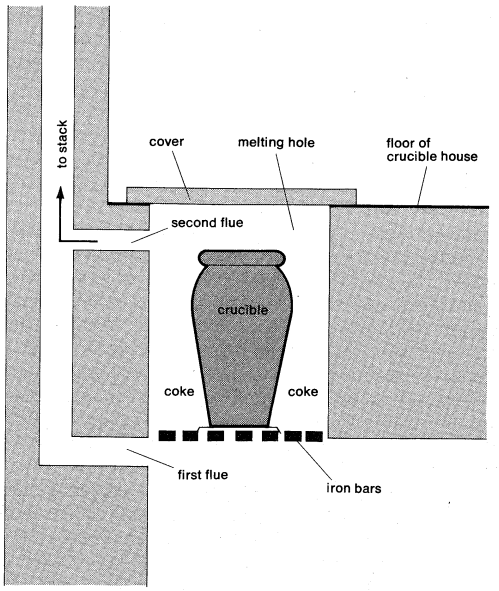


Figure 2: Position of crucible in melting hole in the crucible process.
From A.K. Osborne, *An Encyclopaedia of the Iron and Steel Industry;* Technical Press, London

tions, so that temperatures high enough to melt the steel were reached, but the charge was not oxidized. After several hours of heating, the crucible was withdrawn from the furnace and its contents cast in a mold of the desired shape. Because the steel had been melted, the internal structure of the product was much superior to the steel products obtained directly by cementation of wrought iron. A crucible could contain not more than 100 pounds (45 kilograms) of charge, so that castings of greater weight had to be obtained by combining the product from several of them. Silicon, the chief alloying element in the steel, was obtained by a reaction between carbon in the liquid and silica (SiO_2) in the crucible wall and in the slag. Carbon monoxide gas (CO) was also produced. The process was completed in 2½ to four hours, after which there was no further evolution of carbon monoxide. In 1801 a British steelmaker introduced manganese into the charge, and later the additions of nickel, chromium, and tungsten in the charge permitted the production of a wide range of alloy steels.

Early steelworking machinery. Objects of iron and steel have been worked by hand hammering since antiquity. With the development of the production of wrought iron on a large scale, water-driven hammers came into use. Water-driven mills with two rollers for slitting iron bars were first developed in Liège, Belgium, early in the 16th century. A mill of this type for the production of rods and bars used in the manufacture of nails was installed at ironworks at Saugus, Massachusetts, in 1648. Water-powered mills for rolling larger plates and sheets were used in Sweden and Great Britain early in the 18th century. Rolls with grooves for rolling bars of various cross sections, such as rounds and squares, came into use in the middle of the 18th century; perfection of a mill with grooved rolls in 1783 marks the beginning of modern rolling mills. In England in 1786, Watt's steam engine was first used to drive a slitting and rolling mill; in subsequent years, however, the steam engine was used more widely to drive forge hammers.

MASS PRODUCTION OF STEEL

Late 19th and early 20th centuries. The rapidly rising demand for iron and steel in the middle of the 19th century resulted from the use of these materials in the construction of railroads, ships, buildings, and machinery, required by the rapid industrialization of western Europe, Great Britain, and the United States. War and advancing military technology also demanded steel.

Improvements in steelworking machinery. Advances in machinery for working and rolling wrought iron and steel made it possible to produce a wide variety of useful products. The steam hammer that came into general use between 1850 and 1860 had a steam-driven piston that actuated the hammer directly, permitting rapid operation with precise control. Two rolling achievements of note were the production of an iron rail 62 feet (19 metres) long in 1851 at Port Talbot, Wales, and the installation of a mill at Dowlais, Wales, capable of rolling 18-inch- (46-centimetre-) deep H beams 50 feet (15 metres) long and rails up to 120 feet (37 metres) long. The universal mill with both vertical and horizontal rolls, which was invented in England in 1829, was first used in 1848. The first I beams were rolled in France in 1840. The so-called three-high mill, which first appeared in England in the early 1820s and came into general use after 1850, when it became possible to build bearings, gears, and drive systems of satisfactory quality and precision, has three rolls placed one above the other, and a bar or sheet can be passed back and forth through the mill. The rolling can be accomplished in both directions without the need for reversing the mill. The advantage of this mill over the two-high mill (see Figure 3) was that the piece did not have to be passed back over the mill by hand after each rolling operation.

From A.K. Osborne, *An Encyclopaedia of the Iron and Steel Industry;* Technical Press, London

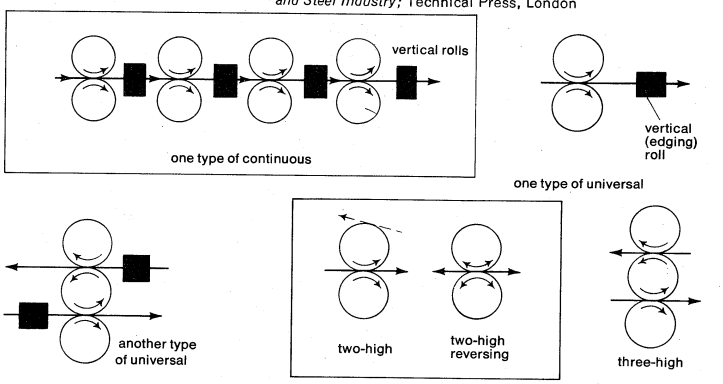


Figure 3: Representative roll arrangements.

Continuous rolling mill

The continuous mill for rolling rods for wire first appeared in England in 1861. It has several consecutive two-high stands of rolls that are alternately vertical and horizontal. A bar passing through the mill is in several stands at once, and its speed increases as its cross section is reduced. The multiple-stand continuous rolling mill for producing large coils of steel strip from hot slabs of low-carbon steel was first used in 1923 in the United States. The multiple-stand continuous mill for rolling these coils cold at high speed into thin sheets began to replace old hand-operated mills in the 1930s.

To reduce waste as well as the amount of rolling required and the size of mill needed, equipment and methods have been developed for casting steel continuously into bars, rods, and slabs. Employed successfully since 1950, these strand-casting methods have been used to make lengths of steel with round or square cross sections called blooms.

First techniques of mass production. By 1850, relatively large quantities of liquid pig iron were being produced by the blast furnace, but the existing processes for making wrought iron and steel, principally the puddling furnace and the crucible process, could produce metal in batches of not more than a few hundred pounds each. In addition, these operations involved heavy manual labour and the exposure of workers to high temperatures. This situation and the growing need for large tonnage of wrought iron and steel stimulated the invention of methods for the mass production of steel. Work on a pneumatic means for producing steel from liquid impure iron by blowing a blast of cold air up through the molten mass was begun by William Kelly of Kentucky, in 1846. In England, Henry Bessemer's efforts along the same lines led to the patenting of the Bessemer process in 1856. At first, Bessemer's and Kelly's process could not be used

Patenting of the Bessemer process

to make good steel from common English and American liquid crude iron. It was soon discovered, however, that the addition of manganese to the blown metal from the converter would make the steel ingots dense and easily workable by rolling and forging without cracking. In 1856 William and Friedrich Siemens, German-born scientists who also lived in England, invented a regenerative system for preheating the air supplied to a furnace so that flames of very high temperature could be produced. A chamber filled with open brickwork is heated for a time by the hot gases from the furnace. Then, for a period subsequently, the air supply for combustion is heated by passing it through this chamber. In this way the chamber is alternately heated and cooled in cycles; normally, a furnace is equipped with two regenerator chambers. In 1858 a patent was issued to the Siemens brothers for the use of regenerators with a shallow-hearthed furnace for melting and refining steel. The Siemens furnace used a grate for burning solid fuel, however, and the ductwork necessary to convey the gases from one end of the furnace to the regenerator chambers and the hot air to the grate at the other end was complex and inefficient. Pierre and Émile Martin in France in 1864 built a furnace that was fired by gas and placed a set of two Siemens regenerator chambers at each end of the furnace; one heated air, and the other heated the fuel gas. By firing the furnace with the gas alternately at one end and then the other, each set of chambers alternately was heated by waste gases and then heated the incoming air and fuel gas. As a consequence, it was possible to operate with a flame whose temperature was sufficiently hot to melt steel scrap. This furnace became known as the Siemens–Martin furnace, or, more commonly, as the open-hearth furnace. The regenerative system was also applied to the crucible process.

Open-hearth furnace

Native ores in Europe and in England contain considerable phosphorus, and pig iron produced from these ores, upward of 1 percent phosphorus, is unsuitable for steelmaking by Bessemer's process or the Siemens–Martin furnace. In 1876 S.G. Thomas and Percy Gilchrist, also in England, invented a modification of Bessemer's process by which satisfactory steel could be made from high-phosphorus pig iron. Instead of silica, a lining of lime was used; this permitted operation with a slag high in lime that could absorb phosphorus from the metal. Very shortly after this discovery, open-hearth furnaces were also built with their hearths lined with lime or magnesia instead of silica. Thus, the basic open-hearth furnace came into being.

The new methods for making steel were utilized rapidly in the commercial production of steel on the European continent, in Great Britain, and in the United States. The first commercial production of Bessemer steel in the United States was in Wyandotte, Michigan, in 1864, and in 1868 the first commercial open-hearth operation was started in Trenton, New Jersey. These processes produced steel in batches, or heats, that ranged in size from one or two to 15 tons or more. Thus, steel products in the form of rails, structural members, plates, and castings could be made cheaply and in relatively large quantities. The expansion of railroads in the United States following the Civil War depended largely on the availability of steel for rails. In the 1880s and 1890s buildings were built using steel members as the skeletons. Though the electric motor came into use for many purposes between 1880 and 1890, it did not completely replace the steam engine in the large rolling mills for many decades.

Use of the electric dynamo in steelmaking

As soon as the dynamo became a practical reality, chemists and metallurgists put electricity to work as an energy source for steelmaking. William Siemens is credited with having first used electric power to make steel (1878). The heat source was the electric arc. P.-L.-T. Héroult in France in 1899 operated the first commercial electric furnace for the production of steel. His design was the forerunner of the modern high-powered arc furnace. This type of furnace was enlarged in an installation in the United States in 1904 with a capacity of four tons (3,600 kilograms) per charge. The electric arc furnace is well suited to the production of high-quality special and alloy steels, and it rapidly replaced the crucible process for making these steels.

Development of fuels. As has been noted, fuels have been of great importance in the production of wrought iron, pig iron, and steel. Charcoal was virtually the only fuel employed from antiquity until early in the 16th century. Coal appears to have been used only occasionally before 1000, and heap coking of coals was practiced in the Middle Ages. The beehive coke oven was first used in 1620, but heap coking of coal for metallurgical purposes persisted. Attempts to develop ovens from which by-products could be obtained from the materials volatilized from the coal began in 1781. The modern retort oven was developed in 1861 and has been the chief means for producing coke for metallurgical use since about 1910. The beehive ovens continued to be operated for many years, particularly in times of high demand for coke. An important step in the development of fuels for steelmaking involved the use of gas as a fuel in the open hearth. This permitted all of the combustion to occur over the charge, a condition that could not be achieved with the combustion of solid fuel in a grate. Between 1880 and 1900, heavy oils, atomized with steam or air under pressure to form a long, nonturbulent flame over the hearth, also came to be used as fuels in the open-hearth furnace. The Bessemer converters and the basic oxygen process are said to be autogenous, because all the heat required is supplied by oxidation of elements in the charge, and no fuel is required. An important advantage of the electric furnace is that there is no possible contamination of the charge by impurities from the source of the energy at the generating station, such as can result from fossil fuels or nuclear power. Supplies of low-sulfur coal and fuel oils are being depleted rapidly, however, and the steel industry will face greater problems of contamination of steel with sulfur from fuels in the 1970s and 1980s.

Use of gas in open hearth

New steels. The growth of steel production in the world was rapid between 1870 and 1950. By 1910 the use of the acid Bessemer process had begun to decline in the face of competition of the basic open-hearth furnace, and the acid open hearth came to be used for the production of special steels for such items as forgings and bearings. The basic Bessemer process was used extensively in Europe for refining pig iron produced from certain low-grade (minette) ores. The electric-arc furnace became the principal method for producing alloy steels by 1920, and the use of the crucible process declined rapidly thereafter. Wrought iron continued to be produced in appreciable quantities, but there was a gradual decline in output in the face of competition from steel. The Aston process was developed in the United States for producing a "synthetic" wrought iron starting with low-carbon iron from an acid Bessemer converter and was used commercially until 1971.

The principal changes in the methods of producing steel in the early part of the 20th century were the rapid increase in the capacity and number of producing units of the iron blast furnace and of the basic open-hearth furnace. In 1950 a large blast furnace produced 1,500 to 1,800 tons of iron a day; and a large open hearth, 400 tons of steel in a heat time of eight to ten hours. The types of steels being produced expanded greatly, and the highly alloyed steels, such as the stainless steels (1915), came into use.

Post-World War II period. The enormous demand for steel created by World War II, the destruction of steel-producing facilities in a number of countries as a result of the conflict, and the economic changes in the immediate postwar era had a profound effect on the steel industry in most countries. The mines that supplied high-grade iron ores that could be used in the blast furnace with little or no processing were seriously depleted and could no longer supply the rapidly rising requirements for iron and steel in the postwar years. In consequence, it became necessary to process low-grade ores to obtain high-grade concentrates. Because these concentrates were very fine, they were agglomerated to obtain a granular product of uniform size (see ORE DEPOSITS for a description of iron-ore occurrence; see IRON MINING AND PROCESSING for in-

formation about mining and ore dressing or beneficiation).

Improvements in ore supply

The development of supplies of iron ores of carefully controlled size and composition has resulted in a very large increase in the output of a blast furnace since 1950. A furnace that produced 1,500 to 1,700 tons of iron a day in 1950 using an unprepared burden, for example, is expected to produce 3,500 to 4,000 tons a day today with an ore supply of good quality. Other improvements have aided in making this increase in productivity possible: increased pressure and temperature of the air blast, injection of fuel with the air, enrichment of the air blast with oxygen. Many mechanical improvements in equipment have also helped to increase the output. The availability of high-grade raw materials have made it practical to build very large furnaces with capacities of 10,000 tons of iron a day. Several of these furnaces were under construction in Japan in 1970.

Use of pure oxygen. Metallurgists have dreamed since Bessemer's time of using pure oxygen instead of air to refine liquid pig iron into steel. Although the fractionation of air to produce pure oxygen was developed in 1889, oxygen remained too costly for many years for it to be used as a primary oxidant in ironmaking and steelmaking. The first large-scale use of oxygen occurred in Germany immediately prior to World War II, following the development of a process for extracting oxygen from air in 1937. Immediately before and after the war, experimentation of the use of oxygen in steelmaking began in Germany, Switzerland, England, Sweden, and the United States. In 1949 a new process came into being as a result of experimental work in Austria. The furnace for the process is a roughly pear-shaped steel vessel lined with a basic refractory (Figure 4). The mouth of the vessel is

"Oxygen and Steelmaking Processes: A Comparison of Open Hearth, LD, Kaldo and Rotor Processes," by E.F. Kurzinski, *AISE Yearly Proceedings, 1960*, p. 99

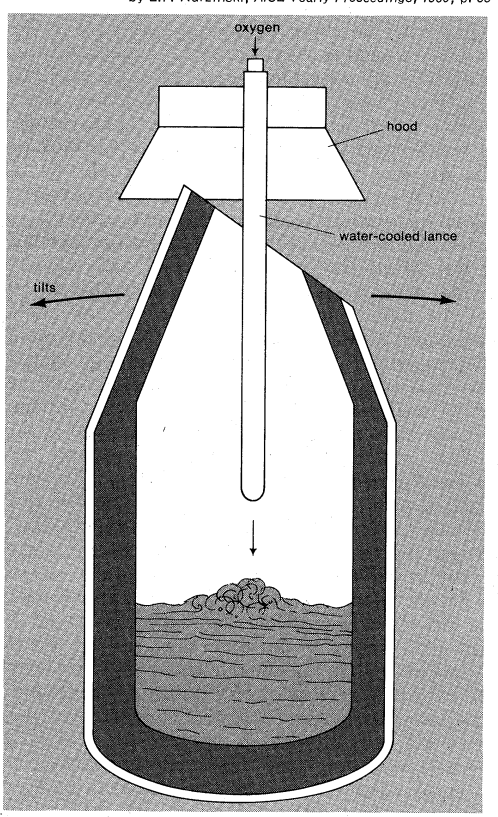


Figure 4: Elements in the L–D, or basic oxygen, furnace (see text).

open and faces upward. The liquid metallic charge placed in the vessel is oxidized by a supersonic jet of oxygen from a water-cooled lance that directs the oxygen vertically downward onto the charge of scrap and liquid hot metal. Depending on where it is located, this process is known as the L–D process, from the towns of Linz and Donawitz, in Austria, the basic oxygen furnace, the basic oxygen process, or the basic oxygen steel process. The first commercial oxygen furnaces had a capacity of 38 tons, but successive installations became larger, and furnaces with capacities of over 275 tons per heat have been in operation in Italy, the United States, and Germany since 1964. A plant with one furnace of this size in regular operation can produce over 2,500,000 tons of steel a year.

The basic oxygen furnace produces carbon steels at least the equivalent in quality to basic open-hearth steel. Many new installations of this process have been made throughout the world, and it is virtually the sole method used in making carbon steels in Japan. In 1970 it surpassed the basic open hearth in the production of raw steel in the United States, and it has been installed in a number of the major steel plants in the Soviet Union. It is also being operated extensively in Europe as a result of the greater use of low-phosphorus ores from overseas. Several modifications of the basic oxygen process are used to refine high-phosphorus pig iron on the European continent and in Great Britain.

Direct reduction. For several hundred years, metallurgists have argued the relative merits of the indirect and the direct processes for producing steel. The indirect process consists of producing a crude, relatively impure iron in the blast furnace and then refining it in the liquid form in one of the steelmaking processes, such as Bessemer, basic open hearth, or basic oxygen. The indirect process is commonly employed in virtually all modern steel plants throughout the world. Its detractors point, however, to the complicated processing sequence and the enormous capital investment for the facilities. In the direct process, relatively pure iron ore is reduced to the solid-metallic state by hydrogen, carbon, or carbon monoxide. The solid metal is then melted, alloyed, and cast into ingots or strand-cast bars and slabs for subsequent processing into finished products. If the reduced solid iron is pure enough, it can be compacted and formed into a bar of iron without the metal having been melted. The problems, however, of finding large quantities of iron ore of very high purity and of controlling the composition of the product have prevented large-scale use of this type of process. In the past century a large number of schemes have been proposed and patented for the direct production of steel. Many of them have been sound technically, but, except in unusual circumstances, none has been able to compete economically with the indirect route of the blast furnace and steelmaking furnace.

Interest in direct methods revived between 1965 and 1970 for a number of reasons. The initial cost of a single blast furnace and its auxiliaries with a capacity of 6,000 to 7,000 tons of pig iron a day is over $70,000,000. It is not practical in most countries to build a small unit with a capacity of 500 to 1,000 tons a day. In addition, relatively large supplies of high-quality ores suitable for direct processing have become available commercially, and the modern high-powered electric-arc furnace is well suited to economical melting of solid reduced ore. Finally, direct processing is suited to small-scale production of steel in the range of 500 to 2,000 tons a day away from the large-scale producing centres. It remains to be seen whether the improved technology for direct reduction will alter the century-old economic advantage of the indirect processing.

In the two decades since World War II the production of steel has spread to many countries of the world. The changing technology by which steel can be produced economically from scrap on a relatively modest scale of a few hundred tons a year has favoured this expansion. Also, a country often desires a steel industry as a mark of its industrialization and as a means of reducing its expenditures abroad.

Alloy steels. The use of alloying elements to obtain special properties in steels began on a commercial basis in 1861, when a patent was issued to cover the use of tungsten in tool steels. Chromium-bearing steels began to appear in the United States, Britain, and France in the

Steel used in construction

1870s. Chromium was used in tool steels; its first employment in constructional steels was in the Eads Bridge built in St. Louis, Missouri, 1867–74. This was the beginning of widespread use of the high-strength, low-alloy steels for constructional purposes. Steel containing 12.5 percent manganese and 1.2 percent carbon was found to be exceptionally tough and resistant to impact and wear. Invented by Sir Robert Abbott Hadfield in England in 1883, this steel is used for railway tires, railway crossings, and the tips and cutting edges of excavating equipment. Hadfield also assisted in the development of silicon-bearing steels for electrical machinery. These steels came into use just after 1900. Late in the 1880s there was a rapid development of the use of nickel as the alloying element in steels used in guns, armour plate, and projectiles in Britain, France, and the United States. Improved armour plate was obtained with the addition of chromium to nickel steels by the French in 1891. Soon after, the use of nickel–chromium steels spread to all types of forgings and machine parts. High-speed tool steels that retain a cutting edge at high temperatures and permit high cutting speeds were first developed in the United States about 1900. These steels contain tungsten, chromium, and vanadium. The appearance of the motorcar in considerable numbers about that time also accelerated the use of alloy steels for crankshafts, gears, springs, and shafts. Vanadium is used extensively in these steels because it imparts resistance to shock and fatigue. The appearance of the electric furnace not only led to improved methods for making alloy steels but also resulted in the production of numerous kinds of ferro-alloys required in the manufacture of these steels. Shortly after 1900, ferro-alloys containing silicon, chromium, tungsten, titanium, molybdenum, and vanadium became readily available to the alloy steelmakers. Stainless steels, which contain over 5 percent chromium, were first developed in Great Britain (1912). These steels have superior resistance to corrosion and were first used in cutlery. Between 1909 and 1912 the more familiar chromium–nickel stainless steels resulted from research work in Germany. The best known composition contains 18 percent chromium and 8 percent nickel; hence, it is known as 18–8 stainless.

Although many alloy steels were developed for munitions and military purposes, during wartime the shortage of materials made it necessary to economize on the use of alloying elements. In World War I a number of special alloy steels were developed that required reduced levels of alloying elements to obtain the desired properties. This was also true during World War II, when additional economies in the use of alloying elements were achieved by adding boron to steels. Boron is a potent agent, and 0.0005 percent of it permits the reduction of other alloying elements by half without loss of strength, ductility, and hardness. Subsequent to World War II, metallurgists throughout the world have been seeking ways of making and solidifying steels and processing them to reduce the amounts of expensive alloying elements that have been used in the past to impart special properties. To a large measure, the efforts have been successful, and in many places low-alloy steels or coated carbon steels are being used in place of more highly alloyed steels.

Protecting steel

Steel can be protected from corrosion and rusting by coating it with tin, zinc, chromium, lead, or copper. Tinning of iron was first practiced in Bohemia in the 17th century. The iron was cleaned in acid and then was dipped into a pot of molten tin to form a thin, adherent layer of tin. This method of coating iron and steel with tin, zinc, or lead containing some tin has been followed for centuries. Galvanized steel has a coating of zinc, and so-called terne plate has a lead–tin coating. To conserve tin, which was scarce in World War II, the electrolytic method was developed for making tinplate. A very thin coating of tin is applied continuously to the surface of a strip of sheet steel by the electrolysis of tin from an aqueous solution. The layer is then melted quickly and cooled to form a bright, adherent coating. The production of galvanized steel was made continuous just after World War II. A strip of sheet steel moves continuously through a furnace, which prepares the surface, and then the strip is passed through a pot of zinc alloy. Subsequently, the strip is cooled to solidify the zinc. Between 1960 and 1970 a wide variety of coatings have been developed for steel sheets; aluminum coatings by hot dipping or deposited from the vapour and a large variety of organic coatings are examples.

II. Modern steel technology

For information on the occurrence of iron-ore deposits, see ORE DEPOSITS. Extraction of iron ore and its treatment are described in IRON MINING AND PROCESSING. For additional information, see MINING AND QUARRYING.

RAW MATERIALS

Pig iron. Virtually all large steel plants in the world have blast furnaces for producing liquid pig iron, which usually constitutes over 50 percent of the metal-bearing part of the charge for the steelmaking furnace. In these large plants, a small quantity of cold pig iron may also be included in the charge for steelmaking, but hot metal is preferred, because no extra heat is required to melt it. Usually the cold pig iron is obtained from discarded ingot molds and the excess hot metal, which is cast into pigs from time to time. Pig iron may constitute up to 15 percent of the metallic charge of electric and open-hearth furnaces that operate on a completely cold charge, as it is an essential source of carbon for the system in these cases. Hot metal is transported from the blast furnace to the steel-producing shop in ladles of various types that are carried on special railway trucks. Common forms for these ladles are either a simple refractory-lined bucket (having a heat-resistant lining) with an open top or the submarine-shaped refractory-lined tank having an opening in the top central portion. There is also equipment for tilting the ladle so that metal can be poured from it and slag and refuse can be dumped out. The capacities of submarine ladles range from 100 to 300 tons, and iron will remain molten in them for periods as long as 24 hours. At the steelmaking shop, hot metal is poured from these ladles into an open-top ladle, which is used to transfer the metal to the furnace. In some plants, especially the older open-hearth shops and Bessemer shops, the metal will first go to a large refractory-lined tank or mixer that serves as a reservoir. A mixer usually has a capacity of 1,000 to 1,500 tons of hot metal. Liquid metal is taken from the mixer as needed by the steel furnaces.

Steel scrap. Steel scrap is the second most important source of iron in the charge to the steel furnace. Generally speaking, scrap has constituted between 48 and 50 percent of the total tonnage of metallic raw materials used in steel production throughout the world in recent years. There are three primary sources of scrap: revert scrap is the material discarded within the steel plant in the production of steel products such as plates, sheets, tubing, and wire from the raw cast steel; industrial scrap is the discarded steel from the plants and factories that process steel into useful objects, such as automobile bodies, machinery, and appliances; and reclaimed scrap is the ferrous materials obtained from refuse, worn-out objects, old buildings, and other discarded objects. Revert scrap varies greatly in size and thickness, because it consists of discarded ingots, crops removed from the ends of bars, billets and blooms during rolling, and clippings and trimmings from all kinds of products as they are cut to the required dimensions. It includes rejected products that may be found unsatisfactory at any point in the production process. Approximately 30 percent of the raw steel produced by the steel furnaces is recycled for remelting within the steel-producing plant. Thus, it is imperative for economic reasons that the steel-melting facilities be able to consume this scrap. Where possible, the steel plant segregates scrap from special alloy steels to use when making new steel having a similar composition.

Sources of scrap

Industrial scrap from factories and plants is generally of very good quality. Usually it is clean, its composition is known, and it is baled, or compacted into bundles of a size suitable for charging to the steelmaking furnace. The amount of this scrap produced varies greatly, depending upon the nature of the operation in the factory.

Many new manufacturing techniques have been developed, however, that tend to reduce the amount of steel lost as scrap. In consequence, the fraction of the finished steel produced by the steel industry that returns as industrial scrap to the steel plant also will probably be reduced in the years ahead. The physical condition and composition of reclaimed scrap varies widely. It may be contaminated with undesirable substances (*e.g.*, paints, coatings, cement, and dirt) containing zinc, copper, tin, sulfur, arsenic, or silica that cause difficulties in steelmaking and may be deleterious to many grades of steel. An element such as copper or nickel, for example, may be valuable in certain kinds of steel but quite undesirable in others.

Metallized ore. Metallized ore, the solid metallic product obtained by the direct reduction of iron ore, is becoming an important raw material, particularly for the electric arc furnace. It is of great value in the charge for high-grade steels because of its freedom from contaminants such as copper and tin. It has the disadvantage that the worthless rock (gangue) it carries from the ore increases the slag volume of the melting furnace. The reduced product must be processed and handled so it does not reoxidize extensively between the time it leaves the reduction system and the time it is melted in the steel furnace. In some cases the spongy metallic product is compacted under high pressure to form dense briquettes. These briquettes can be handled and stored under cover without reoxidation being a serious problem. The usual practice has been to transfer the reduced product, either hot or cold, directly to the furnace for melting with a minimum of delay. Provisions for a cover of an inert gas may be included in the handling facilities and storage bins to prevent contact of the metallized ore with air and moisture. Metallized ore may constitute part or all of the initial charge to a furnace, and it may also be fed continuously as the furnace operates. In the latter case, special handling facilities are required.

Fluxes. Limestone, lime, and fluorspar (CaF_2) are important as fluxes for forming a slag of the correct composition and consistency in the steelmaking furnace. The amounts used vary with the different steelmaking methods and will be considered later when each process is described. Lime and limestone must have low levels of sulfur, in the range of 0.04 percent, and silica of under approximately 2 percent to be satisfactory. They must also be crushed and sized by screening to eliminate dust and large pieces. The most desirable burnt lime is as completely calcined as possible without becoming hard and dense due to its being overburned. Similarly, fluorspar should be coarsely granular to prevent loss by dusting when it is shipped, transported within the steel plant, and added to the furnace. Relatively large quantities of fluorspar are required in oxygen steelmaking to assist in the formation of the slag, and there has been a worldwide shortage from 1968 to 1971.

Preparation of fluxes

Liquid iron. The cylindrical-shaped cupola furnace has been used since Bessemer's time to melt pig iron and scrap to provide liquid iron high in carbon for steelmaking. High-quality coke is used as a fuel. During the early years of the Bessemer process, all of the iron from the blast furnace was pigged, and this cold iron remelted in a cupola to provide liquid iron for steelmaking. Several decades elapsed before liquid iron directly from the blast furnace could be used. Since World War II the hot-blast cupola has been used to melt scrap to provide a supply of liquid pig iron for oxygen steelmaking in steel plants where the demand for steel is insufficient to support a blast furnace.

Oxygen. Enormous quantities of gaseous oxygen are used in steelmaking throughout the world. Produced by the fractionation of air (see GASES, INDUSTRIAL AND DOMESTIC), the gas is usually 99.5 percent pure to avoid contaminating the steel bath with nitrogen. In most cases the gas is transmitted by pipeline from the producing plant to the steel furnace. The basic oxygen furnace consumes more oxygen per ton of steel made than does the basic open hearth, and, as the former has replaced the latter, the consumption of oxygen in steelmaking has risen rapidly.

Refractory materials. Large quantities of costly high-grade refractory materials are consumed in steelmaking. Every steel plant must have facilities for storing and handling these refractories and for relining ladles and converters and for rebuilding furnaces. The consumption of refractories by the basic open-hearth process is much greater than it is for the basic oxygen process, because the furnace structure is more complex. As the basic oxygen process has replaced the basic open-hearth process, the requirement of basic refractories in steelmaking has declined markedly.

THE BASIC STEELMAKING PROCESS

Virtually all the world's annual production of 642,000,000 tons of steel is made in a process that begins with the production of an impure liquid bath from a charge that is principally steel scrap, hot metal, pig iron, and, perhaps, a small amount of ore. The bath is then refined to obtain the desired carbon content and to remove impurities. Finally, alloying elements are added to the refined bath to adjust its oxygen content and to bring its composition to the formulation required for a given steel product. The major impurities in the bath removed during refining are carbon, silicon, manganese, phosphorous, and sulfur. These are oxidized to form a gas containing carbon monoxide and carbon dioxide from the carbon and a liquid slag containing the oxides, silica (SiO_2), manganese oxide (MnO), and ferrous oxide (FeO), which react with lime (CaO) in the charge. Undesirable materials introduced with the charge that also enter the slag are sand, dirt, bits of cement, small quantities of blast-furnace slag, and gangue (worthless rock) in the ore. Materials such as lime, limestone, and fluorspar (CaF_2) may be added to assist in the formation of the slag and to give it the desired composition.

Impurities in the liquid bath

Types of furnaces. Contemporary steelmaking furnaces can be classified broadly in terms of their shape and the way in which heat is provided. The open-hearth furnace has a broad, shallow hearth to contain the pool of metal and slag. Heat is supplied from a large, luminous flame that plays over the surface of the slag. The electric-arc furnace also has a shallow hearth; heat is supplied from arcs struck between graphite electrodes and the metal bath. The converter is a pear-shaped vessel into which an oxidizing gas is injected. Heat to support the process is provided by reactions between the gas and materials in the charge. An induction furnace is also used on a small scale: the scrap charge is contained in a cup-shaped crucible, and heat for melting and refining is supplied from electric currents induced in the metallic charge from an external coil in which an electric current is circulated.

Slags and refractories. Steelmaking processes may also be differentiated on the basis of the type of slag necessary for refining the steel and the type of refractory used in the furnace. Broadly, there are two types of oxides used in slags: acids, such as silica (SiO_2) and phosphorus pentoxide (P_2O_5), and bases, such as lime (CaO) and magnesia (MgO). Ferrous oxide (FeO) and alumina (Al_2O_3) are essentially neutral in steelmaking slags, and manganous oxide (MnO) tends to be basic. Acidic steelmaking slags contain an appreciable concentration of silica with very little lime, and basic slags contain an appreciable concentration of lime and up to about 20 percent silica. The decision whether to use an acid or basic slag depends on the levels of sulfur and phosphorus to be contained in the charge to the furnace. If the amount of either is greater than that which can be tolerated in the final steel product, it is necessary to use a basic slag, because it is able to absorb both of these elements. The nature of the slag also determines the type of refractory materials that are used to contain the slag and metal. An acid slag will attack and rapidly dissolve basic refractories, while a basic slag will destroy acid refractories. Silica is the principal refractory used in acid steelmaking furnaces. Basic refractories are composed chiefly of pure magnesia (MgO), calcined dolomite ($CaO \cdot MgO$), or lime. In some cases, chromite ($FeO \cdot Cr_2O_3$) may be used as a hearth lining if steels high in chromium are to

Acidic and basic slags

be made. It is usual to choose a refractory with a low solubility in the slag. Only a few percent of magnesia, for example, will dissolve in basic steelmaking slags. The pure oxides from which refractories are made have melting points well above the operating temperatures of the steelmaking system. In most processes, the composition of the slag must be adjusted so that it is completely liquid at the operating temperature of the furnace.

Amount of nitrogen. The amount of nitrogen introduced into the steel during processing also plays an important part in the choice of a steelmaking process. Actual levels are very low; 0.005 percent or less is desired in sheet steel, and values above 0.009 percent are desired only in selected types of steel. (The effect of nitrogen on the properties of steel is described below.) Nitrogen is introduced into liquid steel during melting and refining if the liquid metal is exposed to air or other gas that contains nitrogen. The rate of absorption rises rapidly as the temperature of the liquid metal increases and is particularly rapid when the metal is exposed to an electric arc operating in air or when a jet of air or of oxygen carrying small concentrations of nitrogen impinges on the metal surface. The practical steps necessary to minimize the uptake of nitrogen by the metal are to keep the metal covered with a slag that is impervious to nitrogen, to exclude nitrogen from the furnace atmosphere and be certain that the oxygen used in refining has a very low nitrogen content, and to obtain a vigorous boiling action of the bath by the formation of carbon monoxide as bubbles of these gases carry nitrogen out of the system.

Refining reactions. Liquid impure iron and steel can be refined rapidly by reactions with oxygen from an oxygen jet, from iron oxides such as ore added to the furnace, and from iron oxide dissolved in the slag. The reactions between ferrous oxide in the slag and silicon, manganese, and phosphorus in the crude iron can be written:

$$2FeO\ (slag) + Si\ (metal) = SiO_2\ (slag) + 2Fe\ (iron)$$
$$FeO\ (slag) + Mn\ (metal) = MnO\ (slag) + Fe$$
$$5FeO\ (slag) + 2P\ (metal) = P_2O_5\ (slag) + 5Fe.$$

The first two reactions occur very early in the refining process, and usually all but minute quantities of silicon and manganese go into the slag, particularly if it contains more than about 10 percent ferrous oxide. The oxidation of carbon by iron oxide in the slag is written:

$$FeO(slag) + C(metal) = CO(gas) + Fe$$
$$2FeO(slag) + C(metal) = CO_2(gas) + 2Fe.$$

The gas formed will contain only a few percent carbon dioxide. Pure liquid iron at 1,600° C (2,900° F) in the presence of pure liquid ferrous oxide will dissolve 0.24 percent oxygen. Since solubility rises very rapidly as the temperature increases, the steel bath will absorb oxygen from ferrous oxide in the slag and from a jet of oxygen gas impinging on the surface. As oxygen is added to the bath, it will react with carbon to form carbon monoxide and carbon dioxide. A steel bath at 1,600° C containing 1 percent carbon will have to contain approximately 0.002 percent oxygen to be able to eliminate carbon as carbon monoxide in a bubble of gas passing through the metal. As the carbon content of the bath is reduced during refining, the oxygen content of the bath necessary to drive carbon out into the gas phase rises rapidly. At 0.1 percent carbon, a minimum of 0.02 percent oxygen in the steel is needed, and, at 0.05 percent carbon, 0.04 percent oxygen is required. During refining of the steel bath the actual oxygen level is about twice the minimum. Since the solubility of oxygen in solid steel is very low, the oxygen in the liquid steel evolves very rapidly, chiefly as carbon monoxide, during solidification. Liquid steel containing 0.2 percent carbon and 0.02 percent oxygen may evolve over 100 times its own volume of hot carbon monoxide as it solidifies. Unless the release of gas is controlled, sound ingots and castings of the steel cannot be obtained. Control is effected after refining by adding deoxidizing agents such as aluminum, silicon, titanium, and manganese to the liquid; these combine with most or all of the oxygen. The products of the deoxidation reactions are small particles of the oxides of these elements. Some will float out of the steel, but many will remain in the solidified steel. If a metal serves as an alloying agent and will also deoxidize the steel, only a small excess need be added for deoxidation, because the amount consumed by the reaction with oxygen is usually only a few hundredths of a percent of the weight of the steel.

Removing sulfur from the bath

Sulfur can be removed from the steel bath only if a basic slag is present. Very little sulfur is carried off in the waste gases. The chief requirements for control of sulfur are that the ratio of lime to silica in the slag be greater than 2.8, that the slag be fluid and hot, and that the slag and metal be stirred together actively. Strongly reducing conditions in the furnace also favour desulfurization. The concentration of sulfur in a basic slag ranges between 0.05 and 0.20 percent in normal commercial oxidizing steelmaking operations such as the open hearth and the basic oxygen furnace. Since the final steel should contain from about 0.01 to 0.025 percent sulfur, control of this element is a major problem of the steelmaker.

TYPES OF STEELMAKING PROCESSES

Acid Bessemer process. The converter developed by Bessemer and Kelly is lined with a siliceous refractory material and operated with an acidic slag and, therefore, cannot be used if phosphorus must be removed from the charge. Its use grew rapidly in the United States, where pig iron low in phosphorus was produced from plentiful supplies of low-phosphorus ores. On the other hand, it found little use in Europe and Great Britain, where most steelworks produced hot metal high in phosphorus from local ores high in phosphorus. Even in the United States, however, the open-hearth process enjoyed advantages because of superior control of sulfur and the avoidance of nitrogen in the steel from the air blast.

The acid Bessemer converter is either a pear-shaped or a cylindrical vessel with a slightly constricted nose that has an opening at the top (see Figure 5). Waste gases leave

By courtesy of United States Steel Corporation

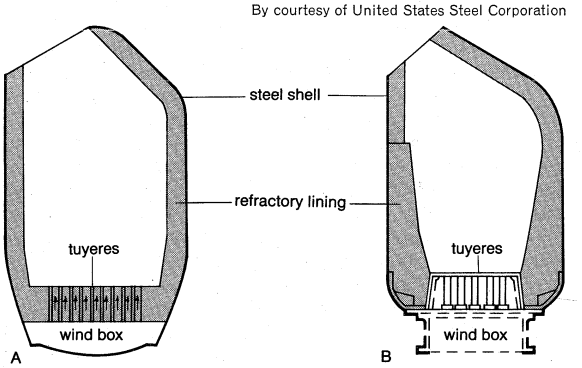


Figure 5: *Bessemer vessels.*
(A) Bottom-blown acid converter; (B) bottom-blown basic converter.

through this opening. The vessel also has a perforated bottom for admission of the air blast. The furnace is mounted in a steel ring around its middle, and this ring is supported by two horizontal shafts, or trunnions, fixed at a right angle to the longitudinal axis of the vessel so it can be tipped for charging and tapping. All charge materials enter through the mouth of the vessel, and finished steel is poured from the mouth as well. Compressed air for the blast is led through a tube in one trunnion and then into a steel box on the bottom of the vessel, from which it passes into the vessel through holes (tuyeres) in the bottom refractory lining.

The refractory lining of the vessel walls, usually composed of sandstone or micaceous schist, will usually last over a thousand heats. The bottom refractory, however, is subject to severe wear and survives only 25 to 35 heats. Accordingly, the bottom of the vessel is constructed for ready removal to be relined and replaced. Spare bottoms are kept on hand and can be installed in 30 minutes.

In the United States a Bessemer shop was usually equipped with two or three vessels that stood in the open. One vessel was available for relining and repair while the

other one or two were in use. A vessel was usually ten to 12 feet in diameter by 20 feet high (approximately three by six metres) and produced from 25 to 30 tons of steel each blow. Ladles of hot metal and chutes for charging scrap were carried by overhead cranes or on railcars. Ladles for the finished steel and pots for the slag were mounted on railcars while being filled. Usually the ladle of steel was transferred by an overhead travelling crane to the teeming platform, where the steel was poured into ingot molds. The operation was controlled by an operator in a pulpit so positioned as to permit an unobstructed view of the vessel. Hot metal was stored in a large mixer with a capacity of between 800 and 1,200 tons. Such a reservoir was needed because hot metal came from the blast furnace infrequently and in large batches but was used in small quantities and at much more frequent intervals. The mixer also provided a supply of hot metal that was relatively uniform in composition and temperature. The shop included facilities for scrap preparation, relining of the vessel and bottoms, storage and weighing and ferro-alloy additions, and a compressor to provide air blasts at a pressure of 20 to 25 pounds per square inch.

The hot-metal reservoir

The usual hot metal for acid Bessemer operations contains 1.1 to 1.5 percent silicon, 0.4 to 0.7 percent manganese, about 4 percent carbon, 0.09 percent maximum phosphorus, and 0.03 percent maximum sulfur. A typical charge is approximately 90 percent hot metal and 10 percent scrap. Sometimes small quantities of high-quality iron ore or roll scale are added to decrease the blow time. When this is done it is also necessary to reduce the amount of scrap because of the extra heat required to reduce the iron oxide. The furnace is charged in the horizontal position, the air blast is turned on, and then the furnace is rotated to the vertical position, a procedure that prevents plugging of the tuyeres. The cold-air blast oxidizes manganese and silicon and some iron to form a slag in the first four minutes of the blow. During this time, the flame emerging from the mouth of the vessel is short and transparent. Then the carbon is oxidized, producing a bright flame about 30 feet (ten metres) long. This flame results from carbon monoxide burning in the surrounding air. When most of the carbon is burned out of the steel, the flame drops. An operator learns from experience to judge the carbon content and temperature of the steel by the character of the flame. A blow lasts 12 to 15 minutes, and the final slag contains iron oxide, manganese oxide, and silica, the amount being approximately 7 percent of the weight of the steel made. Additional scrap, iron ore, cold pig iron, or steam in the blast can be used to cool a hot heat; the temperature of a cold heat can be raised by blowing with the vessel turned on its side. This oxidizes iron into the slag, which generates considerable heat, but reduces the amount of steel produced from the charge. Deoxidizing agents and alloying additions to bring the steel to the desired composition are added to the ladle as the steel is tapped from the furnace into the ladle. For a long time, high-carbon steels such as rails were made by recarburizing the steel with hot metal. The practice led, however, to a product of poor quality.

The basic Bessemer process. Developed to produce steel from the low-grade ores of Europe and similar ores in Britain, the basic Bessemer, or Thomas, process has had an existence in those areas similar to that of the acid Bessemer process in the United States. Production of Thomas steel rose rapidly until 1910 and, except for the period of high demand during World War II, declined as a result of competition from basic open-hearth steel. The process is suffering further decline as a result of the increased importation of high-grade ores that also are low in phosphorus.

In most external aspects, the basic Bessemer steel shop is similar to the acid Bessemer plant. There are essential differences in the way the process operates, however, because a basic refractory and a basic slag are employed.

The basic and acid Bessemer vessels are similar in shape, except that in the basic vessel the nose is usually concentric with the longitudinal axis. The capacity of basic furnaces and the relatively deep bath requires a blast pressure of 35 pounds per square inch. The lining is a plastic mixture of tar and calcined dolomite or lime that is rammed into place. The wall and nose lining lasts up to 400 heats and the bottom refractory to about 40.

Thomas process in operation

The operation of the Thomas process varies so widely from plant to plant that only the general pattern can be described: the vessel is charged with hot metal, the composition of which is 1.5 to 2 percent phosphorus, 0.6 to 1 percent manganese, 0.2 to 0.4 percent silicon, about 3.8 percent carbon, and a maximum of 0.05 percent sulfur. It should be noted that the regular basic Bessemer process cannot operate successfully on pig iron containing less than 1.5 percent phosphorus, because insufficient heat is generated to meet the needs of the process. The weight of lime charged is approximately 15 percent of the weight of the finished steel. This lime is charged with the hot metal. Only about 10 percent of the metallic charge is steel scrap, and it is usually added late in the blow, because much of the heat is generated by oxidation of the phosphorus occurring late in the heat. The reactions proceed much as they do in the acid process, with silicon and manganese being oxidized first, followed by oxidation of carbon. The flame drops when the carbon reaches about 0.04 percent. About three minutes before the flame drops, scrap is added to the vessel, and the blow is continued for two to three minutes more. During this time, the principal reaction is oxidation of phosphorus into the slag. At the end of the blow, the carbon, manganese, and silicon levels in the metal are very low, and the phosphorus content is about 0.04 percent and nitrogen 0.01 percent. The carbon must always be brought to a low level to obtain a dephosphorizing basic slag with 15 to 20 percent ferrous oxide. Thus, it is necessary to add to the ladle large amounts of ferromanganese, other carbon-containing additions, and a deoxidizer as the heat is tapped to obtain steel of the desired final composition.

The final weight of the basic Bessemer slag is about 20 to 22 percent of the weight of the steel made. This slag contains 16 to 18 percent phosphoric acid and up to 50 percent lime, so that it has been sold profitably as a fertilizer. The recent production of less expensive fertilizer from phosphate rock has meant the loss of revenue to the steel plant from the sale of this slag.

Low-nitrogen steel

A number of special practices have been developed to permit the production of steels low in nitrogen by the basic Bessemer process. These consist of keeping the bath cool and shortening the afterblow by adding roll scale or ore (up to 2.5 percent of the metal weight) just before the flame drops, by first blowing half of the charge to completion in order to develop a slag containing considerable iron oxide and then adding the balance of the hot metal and completing the blow, and by equipping the vessel with tuyeres on the sidewall so that the air has a shorter path through the metal and is in contact with the bath for a shorter time. Use of a blast of oxygen and hydrocarbon gas permits refining of hot metal containing little or no phosphorus. Oxygen is used in some plants to enrich the air blast. This practice permits the use of hot metal with less phosphorus, because, with a lower level of nitrogen in the blast, less heat is carried from the system by the waste gases. Alternatively, the use of oxygen permits the percentage of scrap that dilutes the phosphorus in the hot metal and also shortens the afterblow.

Steels produced by the basic Bessemer process are used extensively except where high ductility and formability are required in low-carbon sheet steels. With special practices in the afterblow, particularly blowing with oxygen and steam, low-carbon steels, with low levels of phosphorus and nitrogen, can be made that are competitive in properties with open-hearth steels.

The basic oxygen process. Though the basic oxygen furnace was developed in form from the basic converter, the chemistry of the process is more closely related to that of the basic open-hearth process. The basic oxygen furnace is a large, cylindrical steel vessel with a conical nose and a solid bottom. The opening in the nose is usually concentric, and a taphole is placed about halfway up the conical portion. The refractory lining of the walls and nose cone is made of a mixture of tar and calcined dolomite. The refractory on the bottom may also be made of

tar–dolomite brick or plastic mix rammed in place. A supersonic jet of oxygen is projected down onto the central portion of the charge by a large water-cooled lance inserted vertically through the mouth of the vessel. The vessel is mounted on trunnions so it can be tilted for charging, tapping, and repair of the refractories. Two or three vessels are usually installed together in a row, and one vessel is relined while the other one or two operate. The building housing the furnaces is tall to accommodate bins and weighing equipment for delivering lime, limestone, calcined dolomite, and spar to the vessel as required. The lance-support system and replacement lances are also located above each furnace. Waste gases and heavy red fumes from the furnace are collected in a hood, passed through cleaning equipment, and discharged through a tall stack. Often the hood is water-jacketed so the heat of the hot gases and that resulting from the combustion of the waste gases with air is used to generate low-pressure steam: a hood system has been developed in Japan that is sealed to the mouth of the vessel, and the waste gases are collected for use as a gaseous fuel in other areas of the steel plant.

Charging the vessel

The vessel is charged with hot metal from a large ladle carried by an overhead crane. Scrap may also be charged from a chute carried by the same crane or may be added from a large box mounted on a very large tilting device that travels on rails. The ladle for the steel and a pot to collect the slag are carried under the furnace by a car that travels on rails. This car moves the steel ladle and slag pot out from under the furnace on the side opposite to the charging side. The ladle of steel is then carried by another overhead crane to the place where the steel is poured into ingots or into a strand-casting machine. There are several positions from which the furnace is operated. The controls for the furnace, oxygen lance, waste-gas system, and bins for additions to the furnace are located in a small, well-protected building on the level of the furnace trunnions and where the operator can see the mouth of the vessel during blowing. There are usually auxiliary control stations for tilting the vessel on the charging side and on the tapping side.

The charge to a furnace is usually 27 to 33 percent scrap, with the balance being hot metal. The amount of scrap will be a little higher if the metal is hotter and the level of silicon is high. Scrap in the charge has been increased up to 40 percent by preheating the scrap and by the use of an auxiliary fuel such as silicon carbide. The furnace will operate successfully on a broad range of compositions of hot metal—3.5 to 4.3 percent carbon, 0.2 to 2.0 percent silicon, 0.1 to 2.5 percent manganese, up to 0.45 percent phosphorus, and up to 0.06 percent sulfur. When making steels with sulfur below 0.025 percent, however, it is necessary that the sulfur in the hot metal be less than 0.04 percent. The operation of the process is not particularly sensitive to the kind of steel scrap in the charge. The oxygen required is approximately 0.07 tons per ton of steel made (1,750 cubic feet per net ton [55 cubic metres per metric ton]). Gaseous oxygen is delivered to the lance at a pressure of ten to 13 atmospheres gauge (150 to 180 pounds per square inch). The amount of lime and calcined dolomite required varies depending upon the silicon content of the iron, but an average figure is 8 percent of the weight of the finished steel. Fluorspar (CaF_2) is used almost universally as a fluxing agent to help dissolve the lime in the slag and to produce a thin, fluid slag. The amount used also varies widely and increases as the level of manganese in the hot metal decreases. An average figure for U.S. practice is 0.7 to 0.8 percent of the weight of steel. The slag produced is 10 to 15 percent of the weight of the steel. The scrap is usually charged to the vessel first, and then the hot metal is poured in. The oxygen lance is lowered to within five to nine feet (1.5 to three metres) of the metal surface, and the oxygen jet is turned on. The blowing rate depends primarily upon the size of the furnace, the rates being approximately 0.035 ton of oxygen a minute per ton of steel capacity; *i.e.*, a 200-ton furnace is blown at a rate of 0.7 ton of oxygen per minute. Usually within seconds after the oxygen is turned on, flame and fumes appear at the mouth of the vessel, and the refining reactions begin. Shortly thereafter the desired amounts of lime, calcined dolomite, fluorspar, and ore, if used, are added. Within four to six minutes an early slag is formed by oxidation of iron and of impurities in the charge. After eight minutes, oxidation of carbon begins and continues until the flame drops. If the reaction becomes too violent, slag and metal may be ejected from the furnace, and the operator adjusts the lance position and oxygen flow to correct the condition. Usually the amount of the charge and the blowing rate are designed to give a blow time of from 15 to 20 minutes. In most furnaces, the flame drops when the carbon is about 0.06 percent. If the heat of steel is to have a higher concentration of carbon, the blowing is stopped earlier. When the blow is ended, the furnace is turned down, a thermocouple is inserted into the bath to measure its temperature, and a sample of metal is taken for chemical analysis. Often a sample of the slag is also taken. The carbon content of the bath is determined within a minute or two, and the contents of metals such as manganese, phosphorus, sulfur, copper, tin, and nickel can, if needed, be determined by a spectroscope in a few minutes. The ferrous-oxide content of the slag and the lime–silica ratio may also be obtained. If there is too much carbon, sulfur, or phosphorus in the steel or if the temperature is too low, the heat is reblown for an interval of from one-half to three minutes. If the steel is too hot, limestone or scrap is added. As the steel is poured into the ladle, ferro-alloys and deoxidizers are added to bring it to the desired composition.

Heat times

The time from the beginning of charge of one heat to the beginning of charge of the next heat ranges from just under an hour for large furnaces to less than 30 minutes for small furnaces. There is vigorous competition among furnace crews, steel plants, and nations for the fastest furnace; a good performance is 30 heats a day for one furnace, but an average of 50 heats a day for a month has been achieved. The composition of a typical slag from a low-carbon heat is 55 percent lime, 25 percent ferrous oxide, 16 percent silica, 2 percent magnesia, and a few percent each of manganous oxide and phosphorus pentoxide. The sulfur content would range from 0.05 to 0.20 percent.

The reactions among silicon, manganese, and carbon proceed in much the same manner as in the acid and basic Bessemer converters. Silicon and manganese oxidize rapidly to silica and manganous oxide. Some ferrous oxide is formed, and these oxides react with the lime, calcined dolomite, and fluorspar to form a slag within four or five minutes after the fluxes have been added. As soon as a slag has formed, some phosphorus is oxidized to phosphorus pentoxide and is absorbed by the slag. After approximately eight minutes of blowing, only a trace of silicon is left in the metal, and most of the manganese that will oxidize is in the slag. At this point the reactions with carbon become much more vigorous, and the slag and metal are highly agitated and mixed by the release of very large volumes of gas. The high speed of the basic oxygen process results from this agitation. When the carbon in the bath has reached 1 percent, the phosphorus in the bath is usually under 0.05 percent, so satisfactory steel with a carbon level as high as 1 percent can be made. As the carbon drops, the ferrous oxide in the slag rises, and the lime becomes fully dissolved to produce a fluid, basic slag that will dissolve sulfur and more phosphorus. Nitrogen is not absorbed by the metal in the process, because high-purity oxygen containing less than a 0.1 percent of nitrogen is used, and the vessel nose is designed to prevent any air from entering the converter when the metal is being agitated violently. Levels in the final steel between 0.003 and 0.005 percent nitrogen are obtained.

A great deal of attention has been devoted to finding means for automatic control of the process. Instruments have been developed for determining the temperature and carbon content of the bath and the amount of carbon in the waste gases during the blow. Computer programs have been developed that calculate the charge for a heat and determine such things as the lance position, oxygen delivery rate, and time to stop blowing. Although much

has been learned, a wholly complete and automatic control system has not yet been devised.

The original furnaces installed at Linz and Donawitz, Austria, in 1952 and 1953 had capacities of 39 tons. Since then the size of furnaces has grown rapidly, and many vessels with capacities of as much as 200 and 300 tons are now in operation in Europe, North America, Australia, Great Britain, Russia, and Japan. The basic oxygen furnace is rapidly replacing the basic open-hearth furnace and the Thomas converter for producing carbon steels. Virtually all production of carbon steels in Japan is by oxygen converter. The first oxygen converter was installed in the United States in 1956, and, by 1970, 50 percent of the steel made was by the process. Large plants have also been installed in Great Britain, Italy, France, The Netherlands, the Soviet Union, and India. Installed capacities of basic oxygen furnaces in 1970 in millions of tons were: Japan, 105; United States, 78; European Economic Community, 75; Soviet Union, 23; and the United Kingdom, 23. The total capacity for the world is reported to be 336,000,000 tons, with an additional 116,000,000 tons being planned.

Two-slag operation

A special two-slag operation is carried out when the basic oxygen furnace is used to refine high-phosphorus hot metal, which previously required the use of the Thomas process. An early slag that is high in phosphorus is formed by carrying powdered lime into the vessel along with the oxygen stream in a special lance. A fluid slag is formed, the blow is interrupted, and this slag is poured off. The blow is continued, and more lime is added, to form a second slag typical of normal operations. The first slag can be sold as fertilizer because of its high phosphorus content. In some cases, the second slag is retained in the vessel after the steel is tapped and becomes the first slag in the next heat. This two-slag practice is used in Luxembourg, Belgium, France, and Great Britain.

The open-hearth process. *Open-hearth furnace.* The open-hearth furnace has a large, rectangular, shallow hearth that contains the molten pool of metal and the liquid slag floating on its surface (see Figure 6). Walls of

From H.E. McGannon (ed.), *Making, Shaping and Treating of Steel*, 9th ed. (copyright 1971 by United States Steel Corporation)

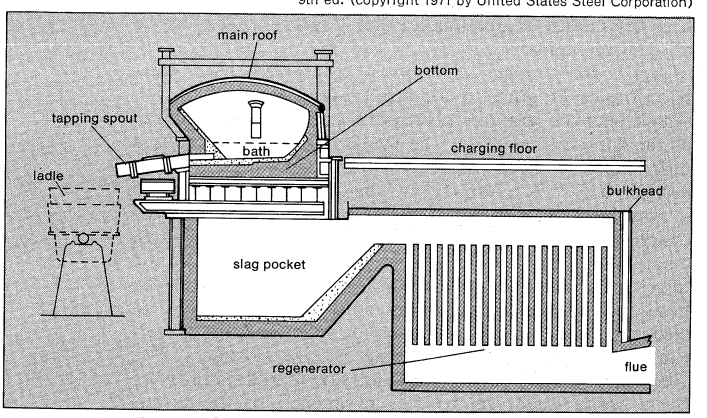


Figure 6: Open-hearth furnace (not to scale). Upper portion shows cross section through taphole of furnace; bottom portion shows cross section through slag pocket and regenerator.

brick and an arched brick roof confine the large flame that is the source of heat in the process and reflect heat from the flame down onto the hot steel and slag. This brickwork is held in place by a framework of heavy steel members. Brick ducts at each of the short sides of the rectangle—*i.e.*, the ends of the furnace—permit the passage of air and gases between the furnace chamber and the brick regenerator chambers located one floor below the hearth level. A large water-cooled burner is mounted at each end of the furnace and, when in operation, projects natural gas or atomized heavy fuel oil, pitch, or tar down the furnace chamber. This spray of fuel is burned by the stream of hot air from the regenerator. A jet of oxygen gas may also emerge from the burner under the fuel jet to increase the rate at which the fuel is burned. The furnace is charged through a number of openings on the front side of the furnace. These openings are closed by water-cooled, refractory-lined doors that slide up and down. It is probable that the name open hearth is derived from the fact that these openings extend the full length of the hearth to make the furnace chamber accessible for charging, working of the heat, and making repairs. Equipment for servicing the furnace operates on the charging floor at hearth level in front of these doors. The materials to be charged are brought to the furnace in large steel boxes carried on small railway cars. A charging machine that also runs on rails is equipped to pick up a charging box, thrust it through an open door, dump its contents on the hearth, and return it to the small car. Hot metal is poured into the furnace from a ladle carried by an overhead travelling crane. The metal is poured in through a movable spout inserted in one of the open doors.

The burner

Draining the furnace

A taphole for draining steel and slag from the furnace is built into the bottom of the furnace on the side opposite the doors. This hole is plugged with refractory, and when the heat is to be tapped the plug is partially dug out and then blasted open with a small explosive charge. The metal flows through a short trough called a launder and falls into a large ladle. The slag also flows into this ladle and floats on top of the liquid steel. When the ladle is full of steel, the slag cascades into a cast-steel slag pot set beside the ladle. The full ladle is transported by another overhead travelling crane to a stand where ingot molds are filled, or teemed. In some shops the ladle is moved in this fashion to a strand-casting machine.

Most modern furnaces are equipped with one or more oxygen lances that project down through the roof. A lance can be lowered so that its tip is close to the slag and the jets of oxygen from numerous small holes in the tip impinge on the surface of the metal bath. A lance also can be withdrawn completely from the furnace so as not to interfere with other operations, such as charging. These lances are smaller than those used in the basic oxygen furnace and deliver oxygen at a much lower rate to avoid heavy splashing of slag and metal. The waste gases leaving the furnace carry large quantities of fume, particularly when oxygen is being used. Gas-cleaning systems have been installed on many modern furnaces to collect this fume before the gases are discharged into the atmosphere.

The regenerators

The regenerators for each furnace are two large chambers, each of which may be as big as the furnace itself, located below the furnace and under the charging floor. Each chamber is partially filled with refractory brick stacked to form many small vertical flues. Air from a blower passes into the lower part of one of the chambers. It then turns and passes up through the flues. The checker brickwork is hot, and the air is heated from 1,200° to 1,550° C (2,200° to 2,820° F) in the passage. The air then passes up into the furnace, where it burns fuel. The gases from the flame and from the reactions in progress in the furnace pass out of the opposite end of the furnace and down into the other regenerator chamber. The ductwork has several abrupt changes in direction and cross section to cause particles of slag, soot, and iron oxide to settle out of the stream of gas before it enters the regenerator chamber. This very hot gas heats the brick flues in the chamber but still retains some heat when it leaves the regenerator chamber. To remove this heat, the gas is passed through a waste-heat boiler. It then goes to the gas-cleaning system and out the stack. The gas flow through the regenerator system and furnace is reversed at approximately 20-minute intervals to keep the air for combustion hot.

The capacities of furnaces range from 200 to 650 tons per heat. Eight to 12 furnaces are usually installed in a long row so that auxiliary equipment can service several furnaces. Hot metal is usually brought from the blast furnace in submarine-type (enclosed) ladles. It may be put in a mixer and transferred to the furnace; or the submarine ladle may serve as a reservoir, and the hot metal may be transferred directly to the open-hearth furnace as needed. Because of its complex shape and regenerator chambers, an open-hearth furnace requires large amounts of costly refractories, and considerable plant space

and staff are devoted to the repair and reconstruction of the furnaces. Most furnaces are basic, and the hearth is formed by ramming a plastic magnesite (MgO) mixture into the desired form. In modern shops the walls and roof are made of chrome–magnesite ($Cr_2O_3 \cdot MgO$) bricks, each of which has a thin steel casing. This steel oxidizes in use and functions as mortar to seal the bricks in place. The hearth requires frequent repairs but will last 3,000 to 6,000 heats. The roof will last between 500 to 1,000 heats; its life depends on the care of the operator running the furnace. Each time a new roof is put on, the furnace is repaired, and the regenerators are cleaned out. The whole structure may be rebuilt when the hearth is replaced.

The furnace is able to melt virtually all kinds of scrap, ranging from discarded ingots and odd pieces that formed in the ladle from a cold heat to very lightweight clippings and trimmings left in the processing of steel and miscellaneous iron-bearing waste obtained from refuse.

Production of a heat. The production of a heat of steel starts with an inspection of the furnace and sealing of the taphole with a refractory plug. The charge usually consists of limestone, steel scrap, iron ore, and hot metal. The weight of the limestone is determined by the weight of silicon in the hot metal and normally ranges from 6 to 9 percent of the weight of steel made. The hearth is first covered with a light layer of plate scrap, and then limestone is spread over the hearth. Iron ore, if used, is spread over the limestone. The amount of ore required ranges from 0 to 25 percent of the weight of steel made and is determined principally by the percentage of hot metal and cold pig iron in the metallic charge. If more than 45 percent hot metal is used, some ore will be needed, and the amount is increased as the percentage of hot metal increases above this level. Charging of these materials may take up to 2.5 hours. The flame is kept on during this time to heat the charge, and much of the scrap is melted by the time charging is finished. During melting, some of the scrap is oxidized, and an early slag high in iron-oxide forms. The hot metal is poured in as soon as possible after charging is complete. The desired composition of the hot metal is 0.8 percent silicon, 1 percent manganese, less than 0.03 percent sulfur, less than 0.15 percent phosphorus, and about 4 percent carbon. A wide range of compositions is possible. If there is ore in the charge, the hot metal reacts with the hot ore and early slag causing a vigorous boiling action of the metal and slag because of the large amount of gas (principally carbon monoxide) generated. It is desired that the ore boil be vigorous so that an appreciable portion of slag is flushed from the furnace through a special port or out one of the doors. If there is no ore in the charge, no slag will be flushed from the furnace. The ore-boiling period will take up to 2.5 hours, during which virtually all of the silicon, much of the manganese, and approximately 60 percent of the phosphorus in the hot metal will have been oxidized into the slag. When the ore boil is almost completed, calcination of the limestone in the charge begins, and a heavy boiling action results from liberation of carbon dioxide from the limestone. The lime boil lasts one to 1.5 hours. At its termination, the slag is heavy and lumpy because of the presence of particles of undissolved lime. The metal is all molten, however, and relatively cold. The charge to the furnace is designed so the level of carbon in the bath at this time is 0.2 to 0.4 percent above the carbon level desired in the steel when the heat is tapped. Much heat is required to finish melting the scrap and to calcine the limestone so the flame is maintained in the furnace during the lime boil. The oxygen lances may be lowered and used to assist in melting the scrap; their action also speeds the ore and lime boils. With the lances in operation, the action in the furnace may become so vigorous and the rate of gas evolution so high that the fuel may be turned off to prevent overtaxing the waste-gas system.

The ore boil

The refining period begins at the end of the lime boil. At this point the bath is cold, the slag is lumpy and raw, and the levels of carbon, sulfur, and phosphorus in the bath are higher in the steel than is desired at the end of the heat. A sample of the bath is taken for chemical analysis to determine the levels of carbon, copper, nickel, tin, phosphorus, and sulfur. Continued heating, use of the oxygen lances, and judicious additions of burnt lime and small amounts of fluorspar developed a fluid basic slag suitable for the removal of sulfur and phosphorus. Iron oxide in the slag and oxygen from the lances react with the carbon in the bath to form carbon monoxide. The heat from these reactions and from the flame raises the temperature of the bath. A skillful operator will bring the bath to the desired temperature and carbon levels simultaneously and will have both sulfur and phosphorus levels below the specified limits.

Refining the steel

Heat times range from seven to nine hours and depend on the kind of steel being made, the use of oxygen, and the type of charge. The heat time does not vary appreciably with the size of the furnace. The amount of slag produced depends upon the amount of hot metal in the charge and on the silicon content of the hot metal. The weight will be from 10 to 15 percent of the final weight of the steel bath. The composition of the slags will be very similar to the range given for the basic oxygen furnace. The fuel composition will range from 6 to 12 percent by weight (fuel oil) of the steel made.

Many of the older basic open-hearth plants have been closed down, and most of the rest were expected to cease operation in the 1970s in the face of competition from the basic oxygen furnace. It would appear, however, that existing modern, well-equipped shops will continue to operate for years to come. The reason is that the open hearth enjoys the advantages over the basic oxygen furnace of greater flexibility in the amount of scrap that can be used in the charge and the ease with which it can make all kinds of carbon steels. It also has the advantage over the electric furnace of being able to produce steels that are low in nitrogen.

The electric process. Steelmaking processes employing electricity as the chief source of heat are utilized in the production of most highly alloyed and specialty steels. The advantages of electric furnaces in steelmaking are absence of contamination of the steel with elements, such as sulfur, that are present in most fuels; flexibility of the system, which permits special slags to be made and removed easily; excellent control of temperature and very high temperatures obtainable; and less loss of expensive alloying elements.

Advantages of electric furnaces

Electric arc furnace. The most commonly used electric furnace (Figure 7) for steelmaking employs an arc struck between graphite electrodes and the metal charge as the source of heat for the system. The furnace has a dish-shaped circular hearth, vertical sidewalls, and an arched roof. It is mounted on a large rocker assembly so it may be tilted forward for tapping and backward for charging, working the heat, and removing slag. Three cylindrical graphite electrodes (two are shown in the figure) are inserted through holes in the roof and are positioned at the apexes of an equilateral triangle. The electrodes are carried on arms projecting from a vertical mast located on one side of the furnace. The electrodes can be moved up and down individually by these arms. Each electrode is connected to the terminal of a large transformer by a heavy cable and a (copper) bus bar. Electrical energy is supplied to this transformer, usually at 22,000 volts, from another transformer and a high-tension line leading to the power generating system. An elaborate set of terminals and switching equipment on the furnace transformer permit energy to be supplied to the furnace in a range of from 50 to 700 volts. Modern furnaces are equipped with a mechanism that lifts the roof, electrode-support system, and electrodes and swings them clear of the furnace, to permit charging from the top by a large steel bucket with a drop bottom. Most furnaces are also equipped with a water-cooled oxygen lance that is inserted into the furnace chamber through a port in the sidewall or in the roof. An electric furnace also has two openings in the sides that are closed by water-cooled sliding doors. One port is opposite the taphole, and the other is 90° from it on the side opposite the electrode mast. These openings permit access to the furnace for adding materials, for removing the slag, and for "working the heat." Machinery for making additions

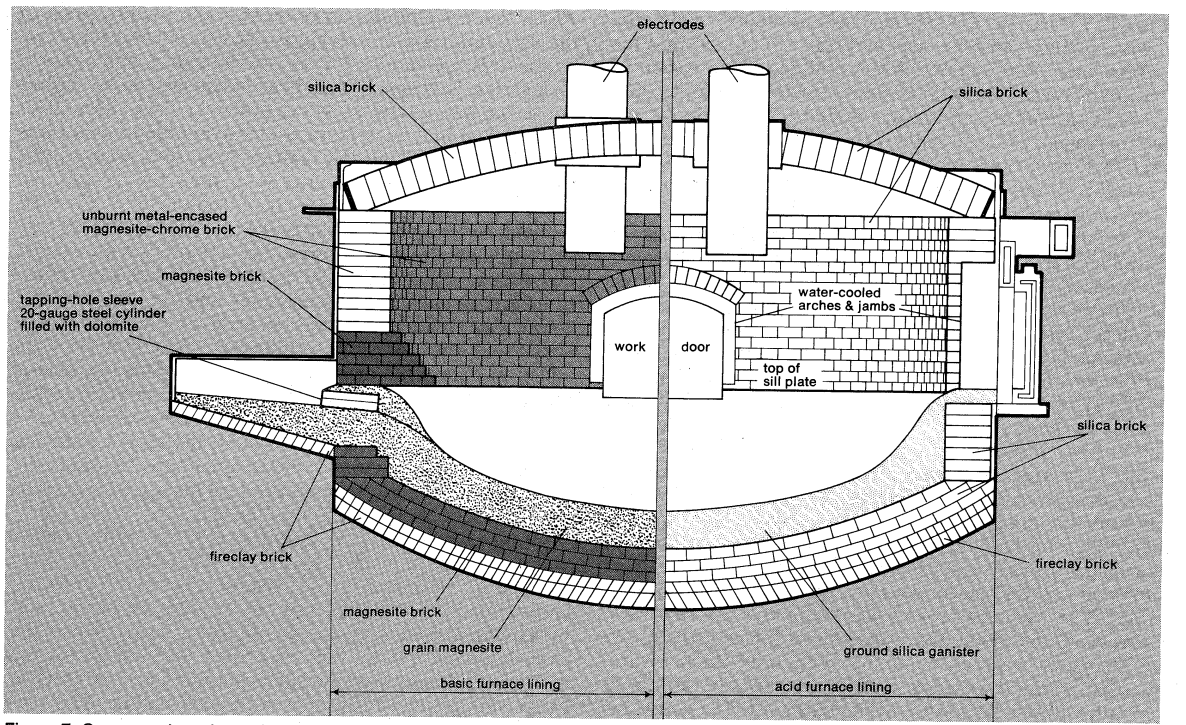


Figure 7: Cross section of a typical electric-arc furnace, showing construction details for (left) basic furnace lining and (right) acid furnace lining (see text).
From C.E. Sims (ed.), *Electric Furnace Steelmaking;* the Metallurgical Society of AIME

to the furnace and for service and repairs operates on the charging floor at the hearth level. Large furnaces are usually located with the hearth level 20 to 30 feet (six to nine metres) above ground level. The support structure is located below the furnace, and the slag pots and the steel ladles are carried on railcars at ground level. Charging buckets are carried by an overhead crane. The steel ladles and slag pots are carried by another overhead crane on the tapping side of the furnace. A plant usually has several furnaces in a line, with the centre line through the tap hole and front door of each furnace at right angles to the line of furnaces. The overhead cranes travel parallel to the furnace line.

Virtually all of the electric-arc furnaces employed for steelmaking have basic hearth refractories and utilize basic slags. The hearth is usually a layer of rammed grain magnesite (MgO) over magnesite brick, or chrome–magnesite materials ($Cr_2O_3 \cdot MgO$) are used when steels high in chromium, such as the stainless steels, are being made. The sidewalls are chrome–magnesite brick, and the roof is made of high-alumina or high-silica brick.

The electrodes

The electrodes are cylindrical lengths of graphite connected together by threaded graphite nipples. As an electrode is consumed, an additional section is added to the end outside the furnace, and the assembly is slipped a distance through the support clamp. Though electrode size is determined by furnace size, furnaces of more than 100 tons capacity usually employ 610-millimetre- (24-inch-) diameter elements; these are the largest practical size commercially available. Either carbon or graphite electrodes may be used. Carbon electrodes are cheaper, but the best quality graphite electrodes are required to meet the severe operating conditions of high-speed, large-capacity modern furnaces. Electrodes are consumed by oxidation, by the action of the arc, and by breakage. The weight of electrodes consumed is 0.4 to 0.75 percent of the steel produced. A modern 200-ton furnace is powered with a 50,000 kilo-volt-ampere transformer, and each electrode carries up to 60,000 amperes at full power. A furnace of this size would have a shell 7.3 metres (24 feet) in diameter.

Considerable fume and smoke are generated when the arc furnace is in operation; therefore, to maintain air quality a furnace must be equipped with gas-collecting and cleaning equipment.

Single-slag oxidizing practice. An oxidizing practice is employed for producing plain-carbon and low-alloy steels in electric-arc furnaces of all sizes. Many recently installed large furnaces use this method to produce ordinary carbon steels in competition with the open hearth. Prior to charging, the furnace is inspected, the electrodes are raised, the roof is swung away from the furnace and the scrap is discharged from a large bucket containing the charge. Some heavy scrap is placed on the bottom of the charge, and light scrap is positioned to fall against the furnace walls to protect them from exposure to the arcs during melting. Hot metal is rarely included in the charge. The burned lime, ore, and fluorspar necessary for the heat are added on top of the scrap. The roof and electrodes are repositioned on the furnace, and the electrodes are lowered, and the power is turned on at about one-third the rated capacity. The electrodes, on automatic control, strike arcs against the scrap and burn holes down through the pile. The power is then turned off, and the electrodes are raised slightly and then lowered again with a higher power setting. This is repeated until full power is reached. The arcs are then deep into the pile of scrap, so that all of their energy is absorbed by the pile. This method also produces a pool of steel in the hearth for the arcs to play on, which also expedites melting. If the charge is mostly light scrap, a second bucket of scrap is added to the furnace after the first is almost completely melted. The full charge is melted in from 60 to 120 minutes. At this time it is necessary to reduce the power to 20 to 30 percent of normal rating of the furnace to prevent severe damage of the roof and wall refractories because of their exposure to direct and reflected radiation from the arc. The bath is cold, and the level of carbon is 0.1 to 0.3 percent above the desired final value for the steel. A sample of the metal is withdrawn for composition analysis. The slag contains undissolved lime and is high in iron oxide from oxidation of the scrap that occurred during melting. The operator then adds ore, lime, and fluorspar as required to assist in obtaining a basic, oxidizing fluid slag, and he may also add the necessary alloying elements—nickel, molybdenum, and copper—that do not oxidize. These may be added either as the metal or as the oxide, because the oxides are reduced by reaction with carbon in the bath. The oxygen lance is also employed to speed the refining reactions and to assist in the formation of a fluid slag. About 30 to 60 minutes are required to reach the desired temperature and carbon level in the

Melting the scrap charge

bath. If a heat of low-carbon steel is being made, the furnace is tilted, and the steel and slag are tapped into the waiting ladle. Although additions and deoxidizers such as ferromanganese, ferrochromium, ferrovanadium, and ferrosilicon and aluminum are added to the ladle as the steel is tapped. The full ladle is then moved to the pouring stand, where a train of ingot molds is waiting or where the steel is strand cast (explained below).

Two-slag practice. Many alloy and plain carbon steels that must be manufactured to very high standards of cleanliness and quality are made exclusively in the electric furnace using what is called a two-slag practice. First the charge is melted and refined with the oxidizing practice described above. At the end of the refining period, the oxidizing slag is removed by tilting the furnace back and letting the slag drain into a pot beneath the furnace. Removal of the slag is assisted by rabbling (stirring the surface by hand) to drag as much as possible of the slag out of the furnaces. Some furnaces are equipped with induction stirrers that induce currents in the pool toward the door and thereby carry the slag to the door. Small amounts of graphite, ferro-silicon, and aluminum are then added to the bath to deoxidize and recarburize it. A new slag, formed by adding burned lime, silica sand, and fluorspar to the furnace, contains less than 1 percent iron oxide; it is very basic and very effective in removing sulfur from the metal. Once the slag is formed, the alloying elements, such as titanium, zirconium, and chromium, that normally oxidize readily can now be added to the bath without serious loss. The bath is stirred gently by rabbling or with the induction stirrer, and then the heat is tapped into the ladle. The steel produced in this way is very clean, because the bath is quiet so that the particles of slag and dirt float to the surface, and levels of sulfur below 0.005 percent can be reached.

Stirring to remove slag

The level of nitrogen in electric-furnace steels tends to be much higher than in open-hearth or basic-oxygen steels, because the bath is exposed to the atmosphere in the region of the arc. By contrast, there is virtually no nitrogen in the furnace gases of the basic oxygen furnace, and in the open hearth the bath is isolated from the atmosphere by a layer of slag. Electric-furnace steels may contain between 0.006 and 0.012 percent nitrogen. Ordinary carbon steels made with the single-slag practice generally stand at the lower end of this range.

The oxidizing slags employed are similar to basic open-hearth slags, except that the iron oxide content usually is lower at a given level of carbon in the bath. The reason for this is that the oxygen level in the furnace atmosphere is kept low because of the reducing action of the hot graphite electrodes in the furnace chamber. Accordingly, phosphorus removal presents a more serious problem in the electric furnace than in the open-hearth or basic oxygen furnace. On the other hand, the phosphorus content in the original charge to the electric furnace is usually very low, so the limitation is not a serious one. The high temperature of the arc makes it possible to obtain a very hot slag; it is relatively easy to produce very basic slags that are fluid if needed. This is advantageous when working with the reducing slag, which is low in iron oxide and silica. The ratio of lime to silica in these slags ranges between 2.5 and 4.5. A typical reducing slag has the following composition: lime, 60 to 65 percent; silica, 18 to 22 percent; magnesia, 7 to 9 percent; calcium carbide (CaC_2), 1 to 2 percent; ferrous oxide, 0.5 percent; sulfur, 0.3 percent; and small amounts of alumina and manganous oxide.

Production of stainless steels. The electric furnace is used extensively in the production of steels high in chromium, especially the stainless steels that contain over 5 percent chromium. To economize on chromium, many plants use a large percentage of stainless scrap in the charge. At ordinary steelmaking temperatures around 1,600° C (2,900° F), oxidation of the bath to eliminate carbon to the low levels necessary in many types of stainless steels results in most of the chromium being oxidized into the slag and therefore wasted. By copious injection of oxygen into the bath, however, the temperature of the system can be raised as high as 1,850° C (3,360° F), which favours elimination of carbon and avoids much of the chromium loss. The slag is deoxidized and removed. The bath is cooled to 1,600° C subsequently, and a deoxidizing slag is formed. The heat is finished as described above for the normal two-slag practice. Two alternatives to the use of high temperatures are available to avoid loss of chromium, which also causes extensive wear of the furnace refractories. The first is to melt the charge and oxidize it modestly to reduce the carbon level of the bath without serious loss of chromium. After the bath has been refined, it is transferred to a ladle in a vacuum chamber in which the metal can also be heated by induction. The application of a vacuum and the use of some oxygen gas reduces the carbon level of the bath as desired, and the necessary additional chromium is then added to obtain the desired composition, and the steel is ready for pouring. A second alternative, coming into use in the United States and Japan, involves bubbling a stream of argon and oxygen through the metal, reducing the carbon in the bath to the desired level without the need for very high temperatures. This method avoids excessive loss of chromium into the slag.

Avoiding chromium loss

Vacuum processing. If liquid steel comes into contact with moisture or water, hydrogen from the water may dissolve in the metal in the ensuing reaction. The usual levels encountered range from one to ten parts per million by weight. On cooling, the solubility of hydrogen in the steel diminishes. If the steel is cooled too rapidly for the hydrogen to migrate to the surface and escape, pockets of hydrogen gas may form at weak points within the structure, and small internal ruptures will result. The problem is particularly acute with very large ingots of electric-furnace steel cast for big forgings used in turbines and electric generators, and is also generally present with alloy steels containing from 2 to 10 percent nickel used for smaller forgings. The presence of these ruptures make a casting unusable. Moisture in the furnace additions, moist runner and ladle refractories, and moisture on the ingot-mold walls contribute hydrogen to the steel. By carefully drying all additions and refractories and by cooling the ingots very slowly in special furnaces, the problem of hydrogen in forgings can be minimized but not eliminated.

Internal rupturing

A number of techniques have been devised in which the liquid steel is exposed to a vacuum in a chamber, and the hydrogen is pumped out to a level ranging from 1.5 to two parts per million. If the steel then does not come into contact with moisture until it is solidified, the problem of internal rupturing can be avoided. One method of preventing such contact is to place the ingot mold in a large steel chamber connected to vacuum pumps. The chamber is evacuated, and the ladle containing the steel is set on top of the chamber. The nozzle is sealed to a fitting closed with a thin metal disc. When the stopper rod is raised, the steel stream cuts through the disc, and the stream and the steel in the mold are exposed to the vacuum. Several alternative methods are also employed.

Vacuum processing is also used to eliminate oxygen from liquid steel. Carbon and oxygen react to form carbon dioxide gas at a low pressure maintained by the vacuum pumps. Under this condition carbon itself acts as a very strong deoxidizer, and no products of deoxidation are left to contaminate the steel. With a good vacuum and a clean system, the level of oxygen dissolved in the steel can be reduced to as low as ten to 20 parts per million, a level that may not be reached with strong deoxidizers such as aluminum.

SOLIDIFICATION AND CASTING

The quality of the surface and interior and the physical properties such as strength, resistance to impact, and corrosion resistance of a steel product are influenced by the way the liquid steel is solidified. In the usual practice, molten steel is poured into a cast-iron mold from a hole in the bottom of the ladle, the hole is closed by a stopper on a rod extending up through the metal, and the rod is protected from high temperatures and action of the liquid metal and slag by a refractory sleeve. A recent development is a sliding refractory valve with all parts outside the ladle and below the nozzle. Because the cast-iron

mold is colder than the liquid steel, solid steel forms as a shell on the mold wall and bottom. Freezing proceeds inwardly from the wall as heat from the hot steel flows into the mold.

The refined steel in the furnace contains an appreciable amount of dissolved oxygen, as described above. As the liquid cools and freezes, carbon in the steel and dissolved oxygen will combine to form a large volume of gaseous carbon monoxide. If the evolution of this gas is uncontrolled, the metal will boil up and overflow from the mold, and the steel ingot will have a porous, spongy interior. This phenomenon caused serious difficulties in the early days of the Bessemer process until it was discovered that adding manganese to the steel to combine with the oxygen would prevent formation of the gas, so the ingots were sound and satisfactory. Subsequently, it was found that other elements, such as silicon, aluminum, titanium, and calcium, all of which form stable oxides, could also be used to deoxidize the steel. Furthermore, as liquid steel solidifies and cools, its volume shrinks 7 percent. Finally, carbon and many alloying elements and impurities such as sulfur and phosphorus tend to migrate, or segregate, into the unfrozen liquid during solidification, with the result that the final bit of liquid to freeze is much more impure than the portion frozen first.

Types of ingots and their manufacture. The internal structure of an ingot can be controlled by regulating the rate at which carbon monoxide gas is formed during solidification. If all the oxygen dissolved in the steel is removed by the addition of a deoxidizer, no gas will be generated, and a shrinkage cavity will form in the upper central portion of the ingot. This ingot is a killed ingot (Figure 8A), and the cavity is called "pipe." The steel immediately surrounding the lower portion of the pipe is relatively impure because of the segregation of impurities and carbon. As a result, the upper portion of this ingot is unfit for use, and as much as 30 percent must be discarded. This loss can be minimized by casting the steel in tapered molds with the large end up. Another method is to install a brick or clay reservoir on top of the ingot mold that maintains a small pool of liquid metal during freezing of the body of the ingot. The metal in this "hot top" flows down to compensate for shrinkage of the metal in the mold. The top of the ingot is cropped and discarded; loss is only 10 to 15 percent. Killed steels are employed when a dense, uniform product is desired, as is the case with alloy steels used in forging, tubing, and plates for pressure vessels. Most high-grade alloy steels are killed steels. A "semikilled" ingot (Figure 8B) results when the addition of the deoxidation agents is limited to allow the generation of a small amount of gas during the latter stages of freezing. If properly controlled, this gas forms small, well-distributed blowholes to compensate for the shrinkage of the metal, and no pipe appears. These blowholes collapse and weld shut when the ingot is rolled. Since it is not necessary to discard the top portion of the ingot, semikilled steels are less costly to produce than killed steels. Semikilled steels are used for structural members, rails, some types of plates, welded pipe, and ordinary bars and common sections, such as angles and tees. Ingot C in Figure 8 is an example of a poorly controlled semikilled ingot in which pipe formed, gas pressure ruptured the top, and metal bled out. With a greater amount of oxygen in the steel when it is poured, it is possible to cause small bubbles of carbon monoxide to form at the interface between the liquid and solid metal. These bubbles grow, detach themselves, and rise toward the top of the ingot. In rising they sweep the liquid along and generate a flow of metal up the walls and down in the central portion of the ingot. If the circulation is gentle, blowholes will be present in the solidified rim, and if it is very vigorous a dense rim forms. After a period of rimming, the top of the ingot is sealed, and the resulting interior structure is similar to that of a semikilled ingot. Ingots D through H in Figure 8 show the types of structures that can be obtained. With ingot D and E, the rimming action is modest, and the level of the liquid rises because of the large volume of the blowholes in the rim. Within one to two minutes after the mold has been filled, the steel rises and freezes against a heavy cast-iron or steel cap placed on top of it that seals the ingot and stops evolution of the gas because of a rapid rise in internal pressure. This type is called a capped ingot, or a bottle-top ingot, because of its shape. Ingot E is a better structure, because the rimming action was stronger, and the blowholes in the rim are farther from the surface of the ingot. Plain carbon sheet steels containing between 0.04 to 0.20 percent carbon are often cast as capped ingots. Ingots F, G, and H are called rimmed ingots, because the oxygen level in the liquid steel is high enough to sustain a strong rimming action for from 12 to 16 minutes that produces a rim from three to four inches (75 to 100 millimetres) thick. After a rim of the desired thickness is formed, the ingots are usually sealed by placing a heavy steel plate on top of the liquid surface, and the top of the mold is flooded with water. Of the three, ingot G has the best structure for most purposes. In practice, the rimming behaviour is controlled by adding gas-forming agents to the steel to stimulate the action or small amounts of aluminum to reduce the action. Plain carbon steels containing between 0.04 and 0.25 percent carbon can be rimmed successfully. This type of steel has a soft, pure rim and is used when good surface, high ductility, and good formability are desired. Large tonnages of the product are used in automobile bodies.

From G. Derge, *Basic Open Hearth Steelmaking*

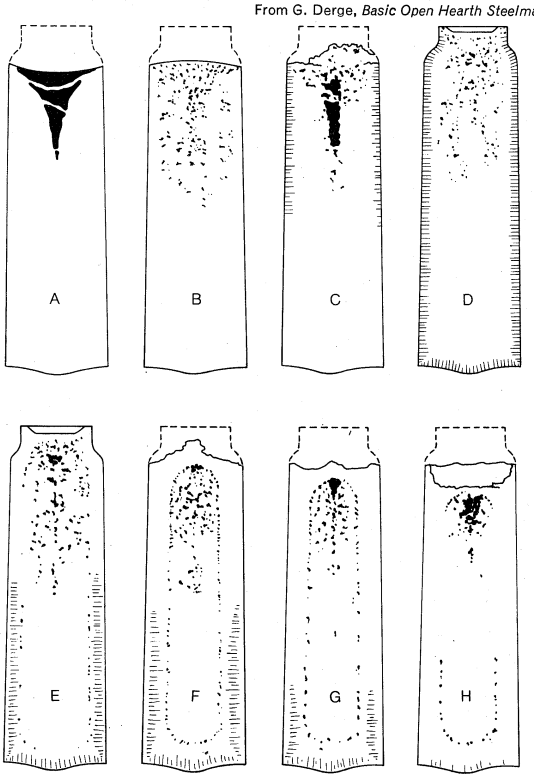


Figure 8: *Ingot structures.*
(A) Killed; (B,C) semikilled; (D,E) capped or bottle top; (F,G,H) rimmed (see text).

Minimizing loss of steel

Rimmed ingots

Control of nonmetallic particles. In addition to controlling the structure, control over the nature and composition of nonmetallic particles in steels is needed. In most steels it is desired to eliminate the particles formed by the reaction of the deoxidizer with the oxygen. Furthermore, small particles of slag may be emulsified in the steel, and bits of refractory may be entrained as a result of the flow and churning of the metal when the heat is tapped from the surface. Most of these particles float to the surface of the metal and combine with the slag in ten to 30 minutes between the end of tap and the beginning of pouring of the ingots. Not all do, however, so many special deoxidizers and pouring methods are used to keep the remaining particles small and well distributed throughout the ingot. Free machining steels are excep-

tions where many small uniformly distributed particles of certain types are desired. These particles lubricate the cutting tool, prevent welding of the cut chip to the tool tip, and cause the chip to break into short pieces. Additions of lead, sulfur, manganese sulfide, selenium, and tellurium to the steel improve the machining properties of steels in this manner.

Variation of ingot size. The size of ingots varies enormously depending upon both the type of steel and its ultimate use. Ingots of highly alloyed tool steels may weigh only a few hundred pounds each. Steels for structural, bar, and rod products usually weigh from two to eight tons (1,800 to 7,300 kilograms). Ingots from which slabs are rolled for large plates and large coils of sheet steels may weigh from 15 to 40 tons. These larger ingots are blocks of steel that may be 90 inches tall by 84 inches wide and up to 38 inches thick (2.3 by 2.1 by one metres). Very large ingots for large, thick plates and forgings for turbine shafts, drive shafts, and pressure vessels may weigh as much as 300 tons each; these are cast in enormous cast-iron molds and take weeks to solidify. Forging ingots are from three to five times thicker than the thickness of the product forged or rolled from it to ensure adequate working of the steel. A greater reduction in thickness is desired for plates, structural products, rails, and most other products.

Ingot casting. In large steelmaking shops, two or three ingot molds are carried on a small ingot car; several cars are connected together so that one heat is carried on a train. These ingot molds are filled from the top. If particularly good surface quality is required, ingot molds may be filled from the bottom. Usually a number of molds are stood in a cluster on a large car. The steel is poured from the ladle into a tall, slim, trumpet-like tube. The metal flows from this trumpet through hollow clay brick to the bottom of each mold, filling all simultaneously and more slowly, with less splashing, than is the case with top-pouring single molds. After the ingots have cooled sufficiently, the train of cars is moved to a stripping device, which seizes each mold and frees the ingot from the mold by the action of a powerful plunger. The molds are set aside for cleaning and reuse, and the train of hot ingots is moved to the rolling mill. There they are inserted in a reheating furnace, where they are "soaked" for four or five hours to bring them to a uniform temperature in preparation for rolling. These furnaces are often called soaking pits. Six to a dozen ingots are stood on end in each pit; a gas or oil flame passing above the ingots supplies the heat. The air for combustion is heated in regenerators similar to those described for the open-hearth furnace. The furnace is covered, and a slightly oxidizing atmosphere is maintained so a crust of oxide forms on the exterior of each ingot. Care is taken so the ingots are not heated above the temperature at which any part of the internal structure becomes liquid, because the presence of liquid will result in severe cracking and rupturing of the steel during rolling.

"Soaking" hot ingots

Strand casting. Rapid progress is being made in the use of methods for casting steel in a long, continuous bar, or strand. Once called continuous casting, this method is now called strand casting. The process has long been used in casting nonferrous metals such as copper and aluminum, but many problems peculiar to steel had to be solved. Many highly specialized machines, control methods, and deoxidation methods have been developed for the process. Only the essential aspects of the method need be dealt with here. It has been possible to cast square and rectangular sections of steel for a number of years, but the problems of surface quality, internal structure, and reliability of operation when casting slabs from 150 by 760 millimetres to 280 by 2,000 millimetres (six by 30 inches to 11 by 80 inches) have been resolved only since 1968. Facilities for strand casting the whole output of large basic-oxygen shops and electric-furnace shops are being installed in plants in many countries in the world.

Advantages of strand casting

There are two important advantages of strand casting over ingot casting. A greater yield of useful product from the liquid steel is obtained, because the large pipe in the cast product is avoided and less steel is lost by cropping the ends of the bar of steel during rolling. Also the structure and composition of the bar is uniform along its length, which is not so with a bar rolled from an ingot.

The essential characteristics of the strand-casting system are shown in Figure 9. The steel is poured from the ladle in the customary manner. It flows into a shallow refractory-lined basin (tundish) and then through a nozzle into the water-cooled mold. Flow into the mold is regulated by a stopper in the tundish. The tundish is used to permit precise regulation of the flow of metal into the mold and to allow additional time for nonmetallic particles in the steel to float out. Several molds are usually fed from one tundish. In recently built systems, the metal from the tundish is conducted through a ceramic tube and is discharged below the surface of the metal in the mold. This avoids splashing and further oxidation of the metal by contact with the air. The surface of the metal in the mold may also be covered with a powdery layer of material used for insulation. The layer in contact with the metal melts to form a thin slag, which dissolves particles of oxide that float on the metal and also coats the bar of steel as it is cast. Heat is extracted from the liquid steel by the water-cooled mold; a thin layer of solid steel forms on the mold wall. In most of the strand-casting machines, the mold oscillates rapidly up and down to aid in the separation of the metal shell from the mold wall. Below the mold the steel bar is sprayed vigorously with high-velocity streams of cold water to speed solidification.

The casting operation is started by inserting a large steel bar of the cross section to be cast up into the mold. The liquid metal is poured into the mold and freezes around a large projection on the bar. This bar is withdrawn at the desired rate, and the flow of metal into the mold is regulated to maintain a constant level. The steel bar is supported by several sets of rolls after it emerges from the mold, and some of the rolls are driven at a controlled rate to regulate the speed of withdrawal of the bar. Considerable time is required to complete solidification, and the tapered liquid core may be 6.3 to 13 metres (21 to 43 feet) long. The bar is cut into lengths by a saw or an oxygen torch below the tip of the liquid region. The speed of casting is very fast for small bars: *e.g.*, 10 by 10 centimetres or 15 by 15 centimetres (4 by 4 inches or 6 by 6 inches). Casting speed ranges from one to two metres per minute (three to six feet per minute) for wide slabs that are 230 to 280 millimetres (nine to 11 inches) thick. To reduce the height of the casting machine, some machines bend the bar 90° in the cooling section (Figure 9) so that it comes out horizontally.

From J. Szekely and N.J. Themelis, *Rate Phenomena in Process Metallurgy*, copyright 1971; used with permission of John Wiley & Sons, Inc.

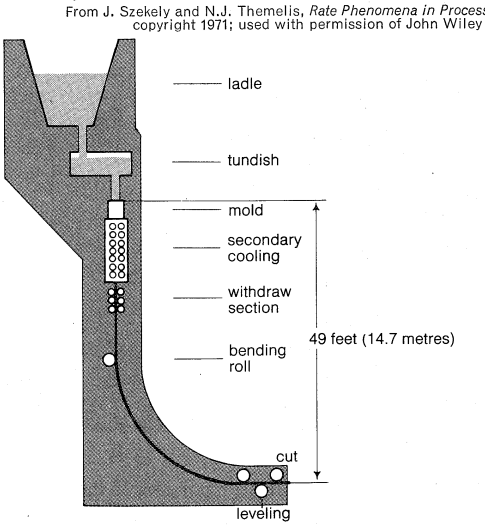


Figure 9: Continuous casting machine.

Other machines have a curved mold that casts a curved bar, in which case the bar is straightened as it emerges from the cooling sections and is horizontal as it leaves the machine. Strand casting avoids the need for large, expensive mills for rolling ingots.

ROLLING AND WORKING OF STEELS

The rolling process. The mechanical properties of a cast steel bar or ingot are improved materially by kneading and working the piece in a press, hammer, or roll stand when it is red-hot and plastic. The nonuniform, coarsely crystalline cast structure is converted to a uniform structure having a fibrous appearance when fractured; there are usually attendant increases in toughness and ductility. The steel is also altered to the desired form. During the working the steel is kept sufficiently hot for the small grains, or crystallites, produced to regrow to form a new structure. Massive gun tubes, drive shafts, and turbine rotors are forged slowly in enormous hydraulic presses. Axles, crankshafts, and heavy engine parts are forged by large steam hammers, while tools and small machine parts are worked by rapid-acting drop hammers.

The rolling mill. Most steel ingots and strand-cast bars are formed in rolling mills. As noted above, rolling mills vary widely in mechanical details depending upon their purpose. The working and shaping of the red-hot piece of steel is accomplished by passing it back and forth through one stand or by sending it through one stand of rolls after another. A bar has its cross section reduced and a sheet is made thinner with each pass through a mill. The rolls of a stand may be grooved to produce a bar with a particular cross section, such as a square bar, a rail, or a beam, or may be cylindrical for rolling flat pieces, such as strips or plates. Though a mill may consist of only one stand, more often several are employed, either in tandem (the steel passes straight from one stand to the next) or side-by-side (the steel bar is looped around to go from one pass to the next). In the latter case, the rolls of each stand may have a number of sets of grooves so that a bar may pass through a stand several times. In the past, a wide variety of arrangements of stand locations have been used in mills.

Mill stands

Modern mills usually have the stands in tandem; the steel passes directly and rapidly through them in a straight line from heating furnace to cooling beds or coiler. The products of rolling mills vary widely in weight and shape, but, in virtually all cases, each piece is uniform in cross section from end to end. Special mills are used to roll steel pieces that vary in cross section along their length. All modern mill stands are driven by powerful electric motors, and the mills are controlled by electronic equipment.

Primary mill. The first reduction of an ingot takes place in a "primary" mill, which may be called a blooming mill if it converts ingots of approximately square cross section into large square or rectangular bars, called blooms. The mill has two large, grooved rolls driven by electric motors of as much as 30,000 horsepower, and the ingot is passed back and forth as the mill is reversed. In each pass the spacing between the rolls is reduced so the bar is reduced in cross section and correspondingly lengthened. The grooves in the rolls also permit rolling blooms with special shapes to be converted into structural shapes, sheet piling, and other forms. This type of mill is called a reversing two-high blooming mill. Some blooming mills are three-high (Figure 3) and roll only in one direction. The ingot can be passed back and forth by alternately using the opening between the middle and upper rolls and then that between the middle and lower rolls. A two-high reversing mill is often used to roll large ingots into slabs for plates and for the production of coils of sheet steel. In a modern steel plant these larger ingots may also be rolled into slabs on a universal mill, which has two stands. One of these stands is a two-high reversing mill for rolling the flat faces of the slab. Adjacent to this stand is another mill, which simultaneously rolls the edges of the slab. Such an installation, with its heating furnaces, shears, controls, and auxiliary equipment, may cost $60,000,000, but it is very fast and can easily roll 4,000,000 or 5,000,000 tons of steel a year. Each mill is equipped with soaking pits to heat the ingots before they are rolled and a shear to crop off the ends of the slab and to cut it into appropriate lengths. Many mills are equipped with a scarfing machine that burns off one to three millimetres (0.04 to 0.12 inch) of each surface of the slab or bloom to obtain a surface free of defects. The burning is accomplished by many jets of gaseous oxygen impinging on the surface of the hot steel.

Second stage of rolling. The blooms and slabs produced by the primary mills receive additional hot work in another mill to convert them into hot-rolled products or to process them further in preparation for cold rolling. The bars and slabs produced by strand casting are introduced into the rolling sequence at this point. Usually the steel is not hot enough to be rolled directly after it leaves the primary mill. Also, it is usually desired to let the steel cool down so it may be inspected and defects on the surface cut out by machining, chipping, or scarfing. The steel is then heated in reheating furnaces, fired by fuel oil or gas. The furnace hearth is inclined and is as wide as the blooms or slabs are long; the pieces of steel are pushed sideways into the furnace. The charge rests on water-cooled skids, and, as one cold piece enters, a hot piece ready for rolling is forced out the other side of the furnace.

Reheating furnaces

Rolled and worked materials. *Rails, bars, and structural steels.* Rails, I beams, H beams, sheet piling, angles, large rounds, and similar products are rolled on a mill having stands with grooved rolls to shape the piece to the desired form. Since special rolls are required for each type of section, the stands are designed so the rolls may be quickly replaced. In modern mills, whole stands may also be changed, inserted, or removed when there is a change in the product being rolled. In a common rolling sequence, a reheated bloom receives an initial pass in a special stand called a scale breaker to remove the scale of iron oxide formed during reheating. The bloom then receives more passes in a two-high reversing stand with grooved rolls. The bar is then brought to its final shape by a number of passes in several subsequent stands. If 12 or more passes are required, it may be necessary to reheat the steel in a second furnace at an intermediate stage. Very close attention is given to the final passes to ensure that the final product is correctly shaped. The hot pieces of steel are sheared to length and cooled on tables. With alloy steels, it may be necessary to control cooling or even to cool slowly in a special furnace. The bars, rails, and beams are then straightened, trimmed, and given a final inspection. For H beams special, massive stands with four rolls arranged to roll the faces of the flanges and each side of the web of the beam are used simultaneously.

Rolling sequence

Mills used for rolling small, simple sections such as round and square bars, angles, and rods also vary widely in the design of the mill and the type of stand used. Blooms may first be reduced to obtain billets having a cross section of under roughly ten centimetres (four inches) square. These billets are then reheated for further rolling. Economical operations require high speed; the most modern mills are continuous, with the stands set in a line. The first several stands are spaced apart sufficiently so the billet is in only one of them at a time, and subsequent stands are closely spaced so the steel passes continuously through all stands simultaneously. To avoid buckling of the steel bar between stands, the speed of the stands must be coordinated and precisely controlled. The steel emerges at a speed of several thousand feet a minute and is either cut to length by a flying shear and cooled on beds or coiled in automatic coilers. These continuous mills may have several of the stands set with their rolls at right angles to the others, or square bars may be twisted 90° as they pass from stand to stand to work all corners of the bar and avoid corner cracks. Some continuous rod mills are equipped to roll four to six bars simultaneously. A very wide range of alloy compositions and types of bars and rods are rolled, and a mill has extensive facilities for inspecting, cutting, straightening, surface preparation, heat-treating, storing, packaging, and shipping.

Plates. The plate mill appears to be very simple in comparison to a structural or rail mill. In some mills, however, the stands and equipment must be massive to be able to roll plates of alloy steels up to five metres (16 feet) wide and smaller plates that may be up to 40 centi-

metres (16 inches) thick. The mills for producing large plates will be described only briefly here. Long plates less than approximately three centimetres (1.2 inches) thick and two metres (80 inches) wide are usually produced on a continuous-type mill, as described below. The production of large plates is a specialized operation because of the many different alloy compositions and sizes that must be made. The mill consists of a storage yard for slabs, several heating furnaces, a scale-breaking stand, the main mill stand and its drive system, run-out tables, a shear, a straightener, furnaces for heat-treating and annealing, and an area for inspection, cutting, trimming, and finishing of the plates.

Plate mill

The main mill stand is usually a four-high reversing mill in which two smaller cylindrical rolls are the work rolls. Slabs of not more than approximately two metres (80 inches) wide can be produced by the primary mill. When making wide plates, it is necessary to give the slabs several passes in its long dimension and then turn the slab 90° and cross-roll it in a few final passes to increase its width. This sequence is required to obtain the desired width, because, in passing through the mill, a slab or bar increases its length but changes its width only slightly. It is found that the ductility of steel is greater in the direction of rolling than it is at right angles to this direction. Hence, cross-rolling is required to avoid the development of the highly directional properties that result when steel is rolled only in one direction.

Continuous-strip mills. Much of the steel from the large integrated steel plants in the world today is produced in the form of long, wide strips, either flat or wound in coils. One large coil usually is obtained from a single slab that may be as much as 250 millimetres (ten inches) thick and 11 metres (36 feet) long and may weigh up to 40 tons. This slab may be strand cast or rolled from a single ingot. Sheet steel for skelp (pipe production), narrow, lightweight plates, and narrow and wide continuous-strip steels are produced on the hot strip mill. Such a mill consists of several heating furnaces, a scale breaker, four roughing stands, a shear, six finishing stands, a run-out table, and coilers. The roughing stands are equipped with edge rolls to control the width of the strip. Except for the scale breaker, the stands are four-high with smaller work rolls backed by two massive rolls to stiffen the work rolls. Each stand of a wide strip mill is driven by an electric motor rated at 3,000 to 6,000 horsepower. The scale breaker is a two-high stand followed by sprays of very high pressure water to carry away the scale loosened by the light reduction of the slab as it passes through the stand. Removal of all scale is essential if a sheet with a surface of good quality is to be obtained. All of the mill stands are set in a line, and the roughing stands are set sufficiently far apart so the slab is in only one of them at a time. The mill may have a slab squeezer, which is a large press located after the first roughing stand to flatten the slab and true up its edges. On leaving the last roughing stand, the strip is about 25 millimetres (one inch) thick. After its ends are squared in a shear, it passes on into the finishing train. The long strip of steel is in all finishing stands simultaneously; the openings of the rolls in the stands are set to obtain the desired final thickness of the strip. These settings and the speed of the stands are controlled automatically. It is remarkable that the stands are only about five metres (16 feet) apart and the strip is travelling several thousand feet a minute as it leaves the last stand. The temperature of the steel must also be carefully controlled to obtain a satisfactory product. The strip then is cooled by water sprays, after which it may be coiled. The thicker strips may be transferred for cooling to another roller table, where they are cut up into long, narrow plates. The final thickness of the hot-rolled strip produced on the continuous hot mill ranges from three to ten millimetres (0.1 to 0.4 inch). A large modern mill producing strips up to two metres (80 inches) wide can roll 4,000,000 to 6,000,000 tons of steel a year.

Removing scale

The product from the continuous hot strip mill has a tightly adhering oxide coating, or scale, covering its surface. This scale may be desired on light plate and sheets taken directly off the mill. For some products, however, the sheets may be pickled to remove the scale and then coated with oil. Pickled products may be sold flat or in coils; wide strips may be split into several narrow strips suitable for making such items as pipe. Much of the product from a wide continuous hot strip mill must be cold rolled to obtain the desired surface quality and physical properties. The coils are first pickled continuously in long tanks to remove the scale. They are then cold reduced 40 to 90 percent in thickness at high speed in powerful mills consisting of four to six four-high stands that are set closely spaced and in tandem. During rolling the strip is kept under tension and driven reels pay out and take up the strip. The product of the mill may be annealed (heat-treated to remove internal strains) in batch furnaces or may be uncoiled and processed continuously through a long annealing furnace. Often coils of either hot- or cold-reduced steel are rolled on single- or double-stand four-high mills to obtain precisely controlled physical and mechanical properties and also a special surface finish. The product is oiled and inspected and may be sold in coil form or may be cut up into sheets. Many steel plants process cold-rolled strip steel into coated products. Coils of tin-plated steel are produced continuously by passing the steel through a bath in which a coating of tin is electroplated onto the surfaces, and this coating is fused to obtain a bright surface. Galvanized sheet is produced continuously by passing the strip through a long furnace to reduce any oxides on the surface and to soften the strip and then passing it into and out of a tank of molten zinc. After cooling, the familiar bright, spangled galvanized steel is obtained. Many other types of coatings, such as paint, chromium, aluminum, and organic coatings, are now being applied to steels.

Pickling to remove scale

Pipe. Pipe in all sizes from six millimetres to 1.3 metres (¼ to 50 inches) in diameter is produced from flat-rolled strip and plate steels. The methods of production vary widely; some of the more common are illustrated in Figure 10 A, B, C, and E. The smaller sizes are produced from narrow strips called skelp by method A. Lengths of skelp six metres (20 feet) long are pointed at one end and are heated flat in a furnace to a white heat.

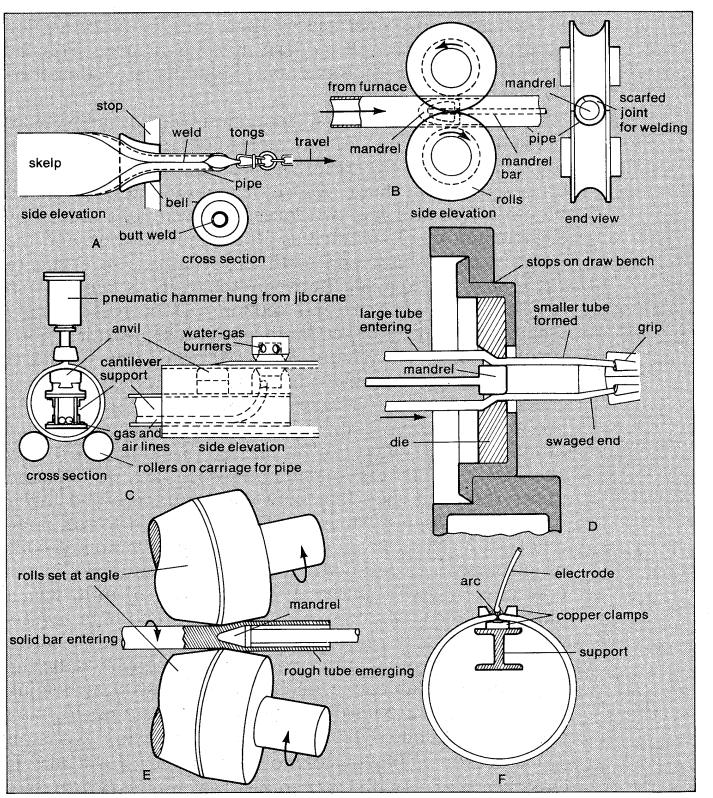


Figure 10: *Various methods of manufacturing pipe.*
(A) Butt welding a two-inch pipe; (B) lap welding; (C) hammer finishing and gas welding of 30-inch pipe; (D) drawing seamless pipe; (E) Mannesmann method; (F) arc welding of 30-inch pipe.

At this temperature the steel is coated with a shiny layer of molten iron oxide and welds easily under pressure. The strip is drawn rapidly through a welding bell, which simultaneously forms the strip into a tube and forms a butt weld. Smaller sizes of pipe with inside diameter ranging from three to seven millimetres (0.1 to 0.3 inch), are made by lap welding (B). The edges of the strip of skelp are first bevelled in a set of rolls, and then the strip is formed into a tube by pulling it through a ring. The tube is then heated to the welding temperature and rolled over a mandrel (core) by a special set of rolls, as shown. The pressure on the overlapping edges from the roll and mandrel causes the steel to weld. The butt-welded or lap-welded pipe is straightened, de-scaled, cut to size, and threaded, if required, tested, and coated with oil, paint, or plastic coating. The steel used in making these pipes is plain carbon steel containing not more than 0.1 percent carbon. The higher carbon and alloy steels do not weld readily by such simple means.

Large-diameter pipe

Large-diameter pipe is produced by forming long plates into tubular form in special presses. First the edges are bent slightly, then the strip is formed into a U shape, and finally the U is bent into an O. The weld is made either by gas (C) or electric (F) means, as shown. After welding, the pipe may be rerolled for shape, heat-treated or annealed if required, cut to size, pressure tested, and given a protective coating. In some plants, pipe of many sizes is produced continuously from strip by electric resistance welding. The strip is formed cold into a tube, with the edges tightly pressed together. Welding is accomplished by passing a heavy current across the joint, which rapidly heats the metal to the welding temperature. The current is supplied by two copper rollers, one on each side of the joint. Because high strength, good corrosion resistance, and other special properties are desired in large pipes, they are made of intermediate carbon and alloy steels.

Seamless tubing. Seamless tubing is usually employed if expected internal pressures are high, service is severe, and reliability is of great importance. Tubing for nuclear reactors, oil wells, and high-pressure chemical equipment are examples. The diameter of the finished tube may be up to 450 millimetres (18 inches), with wall thicknesses as high as 150 millimetres (six inches). The tube is made from a round billet with a total weight of steel that is slightly greater than that of the tube to be made. The billet at a red heat is first pierced by being rotated rapidly between two rolls whose axes are set at a slight angle to each other (Figure 10E). The rotation sets up forces in the billet, causing the centre to rupture. The action of the rolls also forces the billet to advance slowly over a mandrel set between the two rolls. The result is a length of rough tubular blank. After piercing, the tube is reheated and the inner and outer surfaces are smoothed by rolling it over a mandrel as in lap welding (Figure 10B). Subsequently, the tube is straightened, trimmed, and inspected. Tubes may be reduced in size and their strength improved by cold drawing over a mandrel held in a die, as shown in Figure 10D. If a large reduction in diameter or wall thickness is desired, the tubes may be annealed several times during the processing. Tubes produced in this way are certain to be of good quality as only the best quality steel can be pierced successfully. Seamless tubing and other complex shapes are also produced by extruding a hot billet of steel through a die in a powerful hydraulic press. To produce a tube, a billet with a hole down its axis is placed in the cylinder of the press, and the steel is forced to flow out the annulus between a die and a mandrel attached to the piston of the press. The billet is contained in a woven glass sock when it is placed in the press. This sock melts, and the liquid glass acts as a lubricant between the billet and the cylinder of the press. This method was developed in France. Extruded tubes are then processed similarly to the method used for pierced tubes. Steels can also be extruded cold.

Wire and cold-drawn rods. Steel wire is produced by drawing a hot-rolled rod continuously through a number of dies. The reduction in each die ranges from 10 to 40 percent. The rod is first pickled to remove scale and is then given a light coating of lime, phosphate, or copper for lubrication. During the drawing, the dies are also lubricated with soap. Because the steel is drawn cold, the wire becomes stronger and harder with each draw. To prevent breakage, the wire is annealed frequently. Wire is produced in a large number of steel compositions, diameters, stiffness, or tempers, and coatings, depending on its end use, which may be in steel cables, coil springs, nail fencing, or household screening. Hot-rolled rods may be cold reduced slightly by drawing them through several dies. This improves the surface finish and increases the strength of the steel.

CONTROL OF PROPERTIES

Satisfactory production of steel requires that the temperature, composition, and properties of the metal and the processing methods be kept under close surveillance at every step. The compositions of iron ores, sinter, pellets, hot metal, blast-furnace slag, liquid steel, and steel-furnace slag are determined from samples taken during processing. Analyses are obtained by conventional chemical methods and with the use of emission spectroscopy and X-ray fluorescence (see INSTRUMENTATION). The gas content of steel is determined by collecting and analyzing the gases obtained when a sample of steel is remelted under vacuum or in a stream of argon gas. Temperatures are measured by thermocouples inserted in the furnace or in the liquid metal or by optical devices sighted on the steel or into the furnace.

Measuring steel properties

The composition and properties of the solid steel are measured at many stages in processing, from the ingot and strand-cast bar or slab to the final hot- and cold-rolled products. The quality of cast and hot-rolled sections of steel is indicated by the appearance of a fractured surface and by the texture of a cut surface as revealed by etching with acid. Traditionally, the grain size and toughness of a steel bar are judged from the appearance of the fractured surface of the steel. A trained eye is needed for this method to be effective in controlling the production of steel; it has been superseded by many newer and more quantitative techniques.

The general structure and texture of a roughly polished surface cut through a rolled or forged piece of steel is revealed by etching the surface with a hot solution of hydrochloric acid. Lines of flow of the steel from hot working, areas of segregation, and collapsed blowholes and pipe can readily be seen. The distribution of sulfur in the cast or rolled structure can be observed by making a sulfur print of a smoothly ground surface of the steel. A sheet of photographic paper moistened with sulfuric acid is held in contact with the steel for approximately one minute. On being developed, the print shows the texture of the steel and the locations where sulfur is segregated. The structure of hot- and cold-rolled steel products is observed under the optical microscope to determine if the processing has produced the texture and phases necessary for the properties desired of the steel. Many test methods have been developed and are used regularly to determine the hardness, tensile strength, ductility, toughness, and drawability of the steel during processing. These measurements and metallographic examination of the steel are essential in controlling the quality and properties of the steel being made.

Sulfur print

Design of the furnace. The size of a furnace used in heating steel for processing depends very much upon the nature of the product. Furnaces for heat-treating tool steels are small; the batch furnaces for annealing coils of strip steel are refractory-lined boxes of sufficient size to contain six or eight coils in a charge; and furnaces for heating and annealing strip steels and steel parts on a conveyor are block-long horizontal tunnels or may have several vertical sections through which the steel is looped. These furnaces have zones for heating, cooling, and holding the temperature constant so that the steel receives the desired heat-treating cycle automatically. Much attention has been given recently to the design and construction of furnaces of this type for many purposes. Gas, oil, and electricity are used as sources of heat. If reactions between the steel and combustion gases causing carburiza-

tion, decarburization, scale formation, or pickup of sulfur must be avoided, electrical heating is often used. Furthermore, the steel can be isolated from combustion gases by burning fuel oil or gas within tubes of heat-resistant steel or silicon carbide that pass through the furnace. Heat is transferred from these tubes to the furnace charge by radiation and convection of hot gases within the furnace. Mechanical details of a furnace also depend upon the nature of the charge. Furnaces used for heat-treating and annealing large forgings will have large manipulators on wheels or an overhead crane for moving the steel to the furnace and to and from the press, forge, or hammer. Furnaces for treating many steel parts may resemble a tunnel the bottom of which consists of refractory-covered trucks mounted on small wheels that travel on rails. The charge is placed on these trucks, which are pushed through the furnace at the desired rate. Bars and billets may be moved through long or circular furnaces on rollers or conveyors. To conserve heat, many high-temperature furnaces are equipped with recuperators or regenerators for preheating the combustion air by waste heat extracted from the exhaust gases.

Furnace atmosphere. The atmosphere within a furnace is often important to the purpose of the furnace. Soaking pits for heating ingots and the reheating furnaces on primary mills are usually operated with a strongly oxidizing atmosphere to obtain a dry, brittle oxide coating on the steel. This is to assure that the oxide flakes off easily early in the rolling process so as not to blemish the surface of the finished piece. When heating strip steel or annealing tool steels for long periods, the surface must be preserved, so a nonoxidizing atmosphere is obtained by careful control of the ratio of water vapour to hydrogen and the ratio of carbon dioxide to carbon monoxide in the furnace atmosphere. In bright annealing at comparatively low temperatures (650° C [1,200° F]), a small excess of carbon monoxide over carbon dioxide in the atmosphere is employed to obtain a bright metallic lustre on the steel surface. Moisture content of the gas must be kept low, however, to prevent oxidation of the surface during the cooling cycle. Atmospheres that protect the steel from oxidation during heating to higher temperatures must be much more strongly reducing; these are obtained by combustion of natural gas or other gaseous hydrocarbons with a deficiency of air to produce the desired gas composition. Regulation of the ratios of carbon dioxide to carbon monoxide and water vapour to hydrogen permit a gas to be used to add or subtract carbon from the steel. If the steel is to be case carburized (externally hardened by addition of carbon), methane (CH_4) is added to the gas. The addition of both methane and ammonia (NH_3) produces a hard surface on the steel that is high in both carbon and nitrogen. The composition of the gases can also be adjusted in many ways: by partial combustion of a fuel gas, as mentioned earlier, and by cooling a partially burned gas and absorbing carbon dioxide from it or by chilling it sufficiently to condense out most of the moisture.

Non-oxidizing atmosphere

Cooling rate. Rapid extraction of heat from steel to cool it quickly from one temperature to another is important in developing certain desired properties. The cooling rate of the steel is determined by the nature of the quenching medium and the thickness of the piece being treated. An enormous variety of liquids has been employed since antiquity, but relatively few are now utilized. Often used mediums, in order of decreasing severity of quench, are cold water from high-pressure jets onto hot steel, a bath of iced brine, a bath of circulating cold water, a bath of oil (the oil must have a high flash point to be safe), a bath of boiling water, circulating liquid lead, a blast of air, still air, burial in a pit of lime, and slow cooling in a furnace or refractory-lined box.

Quenching media

The cooling rate required to harden a steel is reduced with increased levels of carbon and the presence of alloying elements. Hence, a less severe quench is used with alloy steels. With a given quench medium, however, the cooling rate in the interior of a piece of steel decreases with increased thickness of the steel. Thus, the greater the thickness of the piece, the more severe the quench must be to harden the steel throughout. With carbon steels, air cooling may be sufficient fully to harden thin strips, but water quenching may be inadequate for heavy gears and forgings. Accordingly, alloyed steels are used for the heavier sections both to obtain hardening throughout and to avoid the severe quench rates that cause distortion and cracking. The quenching medium usually has its temperature controlled by a cooling system and is circulated to assure uniformity of quench from piece to piece and over the surface of each piece. Mechanical devices and conveyors are employed to save labour and provide rapid and consistent introduction of metal parts into and removal from the quenching medium. Uniformity and consistency of the quenching operation are essential to obtaining good control of the parts being processed.

Metallography. The nature and structure of a steel can be observed by examining the polished and etched surface of the piece of steel in an optical microscope. The sample is given a high polish and then etched carefully to differentiate the various phases present. The choice of etchant is determined by the composition of the steel and the aspect of its structure under investigation. The magnification employed is also determined by the purpose of the investigation and ranges from one to 1,500 times. Since the latter figure is the approximate limit of ordinary light microscopy, electron microscopy is used if greater resolution is desired. The structures observed may be photographed for further study and to provide a record of the structure of the steel. Considerable attention is being devoted to the use of closed-circuit television and a computer to determine automatically the nature and amounts of various phases present in an image from the microscope. In the long run, this technique should reduce the tedious work of identifying and counting the nonmetallic particles present in steels, but more work is needed to make it practical.

As was pointed out in the section on *Steel metallurgy*, there is a close connection between the heat-treatment given a steel and the structures obtained. Thus, the microscope is important in determining whether the processing was satisfactory. Microscopic examination of a steel, coupled with measurements of tensile strength, ductility, fatigue strength, impact resistance, and other special physical measurements, is used to establish control of processing of all kinds of steels. Hardness is determined by measuring the size of indentation left by a steel ball, diamond cone, or diamond pyramid that is pressed into the surface of the steel with a specified force. Tensile strength is measured by subjecting a standard steel specimen to an increasing load until fracture. Toughness or impact strength is measured by the amount of energy absorbed by the instantaneous fracture of a standardized test piece when struck by a large swinging hammer (see also MATERIALS TESTING).

III. Properties and uses of carbon and alloy steels

CARBON STEELS

Steels can be classified in a number of ways. One of the most common methods is based on the structure produced during solidification; types, already described above, are killed, semikilled, and rimmed steels. Only a few additional comments are needed here. Killed steels are used in applications requiring a dense, uniform structure. Virtually all alloy steels are produced as killed steels, because a high level of deoxidation is necessary to assure control of density and composition. Steels for strand casting must be killed for satisfactory handling during solidification. Thus, a number of kinds of steels that have been cast as semikilled or rimmed ingots must have their compositions and deoxidation practices adjusted for strand casting.

Chemical classification of steels

The primary basis for classifying steels is by chemical analysis. Most major countries have their own methods of classification, which may indicate levels of carbon, manganese, phosphorus, sulfur, or silicon in a steel, depending on type. If other elements, such as aluminum, lead, and copper, are required, their levels also must be specified when a steel is made.

Carbon is both the cheapest and the most effective alloy-

ing agent in steels. Low-carbon steels, containing between 0.08 and 0.15 percent carbon, are used largely where low cost, good surface, good ductility and toughness, weldability, and ease of forming and working are desired. Examples include the sheet steels used for automobile bodies, welded tubing, all kinds of steel containers, and wire and case-hardened machine parts. Mild steels, containing from 0.15 to 0.25 percent carbon, are easily machined and stronger but less ductile than the lower carbon grades. These are used in structural members in buildings and bridges, in truck bodies and railroad equipment, and in boilers, drop forgings, and bolts. Steels in the medium-carbon range, from 0.25 to 0.35 percent carbon, are harder and stronger than mild steel and are used for large tubes, ship structures, shafts, and a variety of machine parts. Steels containing 0.35 to 0.65 percent carbon are used for large forgings for heavy machinery. Steel rails and wrought-steel wheels that require good strength and ductility and good response to heat-treatment are made from steels containing 0.65 to 0.85 percent carbon. These steels respond well to heat-treatment. Steels with 0.85 to 1.2 percent carbon are used in the heat-treated condition when ranges of hardness and ductility are required. Uses include springs, cutting tools, and rock drills.

Ordinary carbon steels constitute the bulk of the steels produced throughout the world. In the United States, for example, in the years just after 1970, 88 to 90 percent of the output of raw steel was carbon steels, 10 percent was alloy steels, and 1.0 to 1.2 percent was stainless steels. Furthermore, flat-rolled products, sheet, strip, and plate, constituted slightly less than half of the total production.

ALLOY STEELS

A very wide variety of compositions of steels is possible with many alloying elements that may be used singly or in combination. It has been found generally that a combination of alloying elements gives better results in cost and properties than is obtained with the use of one element.

Alloying metals. *Nickel and chromium.* The most important alloying elements in steels are nickel and chromium. Nickel dissolves in ferrite (see below *Steel metallurgy*) and imparts strength to the steel but does not increase toughness. Because of this and because of its cost, nickel is no longer used extensively as the sole alloying element.

Chromium and nickel in combination

Chromium at low levels increases the ability of a steel to be hardened by heat-treatment and joins with carbon to form hard, finely dispersed carbides; at levels above 5 percent it imparts corrosion resistance. A number of steels containing from 0.30 to 1.60 percent chromium are used for grinding rolls, roller bearings, and projectiles. Chromium and nickel are often used together or in combination with small quantities of other elements, such as vanadium, molybdenum, and boron. Chromium and nickel are used together because they increase the toughness and strength of a steel and improve its ability to be hardened. The other elements mentioned above aid in these improvements and reduce the amounts of nickel and chromium necessary. The chromium–nickel steels are used extensively for heat-treated machine parts, axles, gears, and armour plate. Rustless irons and stainless steels contain 12 to 18 percent chromium and up to 8 percent nickel. Heat-resisting alloys contain up to 25 percent each of nickel and chromium.

Manganese. Manganese is contained in all steels to ensure good hot-working properties; at levels above 1 percent it is considered an alloying element. Four steel compositions containing 1.75 percent manganese are used in farm implements and automobiles. Higher manganese killed steels are employed in ship construction where brittle fracture is to be avoided. Manganese is an important element in steels used in high-strength castings, rifle barrels, and heat-treated forgings. The tough Hadfield steel, containing 14 percent manganese, is described below in the section on *Steel metallurgy*.

Vanadium. Vanadium is a good deoxidizer. A few tenths of a percent yields steel with a sound structure, a fine grain, or texture, and good response to forging. High-speed steels and tools for hot work contain 0.5 to 2 percent vanadium.

Molybdenum. Molybdenum is an effective alloying addition at low concentrations (0.2 to 0.3 percent). Molybdenum increases the depth of hardening, improves the resistance to creep at moderately elevated temperatures, and permits good control of grain size.

Boron. Boron at a level of 0.0005 percent reduces by half the amounts of nickel, chromium, and molybdenum normally necessary to achieve a desired strength and toughness. Boron was first employed extensively during World War II to economize on scarce alloy elements; its use has continued since that time. Because of its great potency and the very low concentration required, particular skill and care are required when producing, processing, and heat-treating the boron steels.

Advantages of boron

Tool steels. Ordinary high-carbon tool steels must be used carefully to avoid overheating the tip of the tool, which destroys the cutting edge. Modern high-speed steel can be used at six or seven times the cutting speeds possible with carbon-tool steels, and the cutting edge can be allowed to reach a red heat. These steels contain tungsten, chromium, and vanadium and may also contain either molybdenum or cobalt. When forged and annealed, they are soft and can be machined easily into complex forms. The usual heat-treatment first involves preheating to 850° C (1,560° F), followed by rapid heating to 1,300° C (2,400° F), cooling in oil or air to room temperature, and finally a reheat to 600° C (1,100° F), where the steel is held for 30 minutes before it is cooled to room temperature.

There are many compositions of alloy steels used for special purpose tools. Steels containing 1 to 5 percent tungsten are used for woodworking tools. Tool steels that can be hardened with very little distortion contain 1 percent each of carbon, tungsten, chromium, and manganese. Die steels used for forging contain 0.3 percent carbon, 12 percent tungsten, 3 percent chromium, and 0.4 percent vanadium. The tool-steel industry is highly competitive; many proprietary compositions exist.

Stainless steels. The three principal types of stainless steels are austenitic, ferritic, and martensitic (see below *Steel metallurgy* for the identification of these structures). The most familiar is austenitic, which contains 18 percent chromium and 6 percent nickel, is very ductile, and has good resistance to corrosion and staining. It is used in cooking utensils, food-processing equipment, equipment for the chemical industry, hospital equipment, cutlery and flat tableware, and architectural trim. To reduce the cost, manganese may be substituted for some of the nickel. If superior corrosion resistance is required, small amounts of niobium (columbium) may be added. A number of special grades of austenitic steel have been developed for their unique properties.

Martensitic stainless steel, which contains chromium but little or no nickel, is readily formed when annealed; its strength, hardness, and other mechanical properties may be improved by heat-treatment. Martensite steels are used for equipment in the chemical and oil industries. Ferritic stainless steels are similar in composition, except for a low carbon content, and have better corrosion resistance. They are employed for automotive trim and interior architectural trim. Stainless steels containing up to 1 percent carbon are used in cutlery.

Ferritic stainless steels

Alloying agents. Various pure metals and ferro-alloys are added to liquid steel to obtain a desired composition in the final product and to control the levels of oxygen and sulfur in the steel. Pure aluminum, cobalt, copper, chromium, lead, manganese, nickel, and titanium are used when extra-pure alloying elements are required. Generally, aluminum, cobalt, copper, and lead are added in the pure form; the other metals are more expensive as pure metals, however, than when obtained as ferro-alloys. Hence, these ferro-alloys, which are alloys of iron and one or more of the alloying elements, are the principal alloying agents used today. Most contain appreciable levels of carbon, although a number of grades are produced with little or no carbon for use in special situations. An example is low-carbon ferrochromium, which is used in producing extra low-carbon stainless steels; *i.e.*, the steel contains not more than 0.03 percent carbon.

Ferromanganese. Ferromanganese is usually produced in a blast furnace from high-grade manganese ores low in phosphorus, silica, and iron. The usual product contains between 74 and 82 percent manganese, 7.5 percent carbon, not more than 1.25 percent silicon or 0.35 percent phosphorus, and the balance iron. The operation of the blast furnace is somewhat more difficult when producing ferromanganese than when producing hot metal for steelmaking; in consequence, older and smaller furnaces are usually used. The capacity is reduced by about two-thirds, and the coke rate is doubled. Approximately 20 percent of the manganese is lost to the slag. Ferromanganese can also be produced in submerged-arc electric furnaces (described below). Cost of the process is prohibitive, however, except where power is inexpensive, as in Norway.

Ferrosilicon. Ferrosilicon is used extensively in steelmaking both as an alloying agent and as a deoxidizer. Several grades are produced in the submerged-arc electric furnace: 50, 75, 80, and 90 percent silicon, with the balance principally iron. Ferrosilicon contains less than 0.5 percent carbon and less than 0.1 percent each of sulfur and phosphorus. Phosphorus levels must be kept low to prevent the formation of poisonous phosphine (PH_3) when the alloy is exposed to moist air or water. The submerged-arc electric furnace used to produce ferro-alloys is a steel shell having a shallow hearth usually lined with carbon or carbon blocks. Three large carbon electrodes project vertically down into the hearth with their tips well immersed in the furnace charge. Furnaces may or may not have roofs. The electrodes may be preformed or may be made in place by filling thin steel cylindrical shells with a tar–coke mix. This mix is baked as it passes into the hot zone of the furnace. Heat for the process is generated by currents from the electrodes passing through the charge. The charge, consisting of pure quartzite, coke, and iron turnings, is fed in around the electrodes. Silicon is reduced to the metal at a location low in the furnace near the electrode tips, and the charge is heated in part by hot carbon-monoxide gas rising from the reaction zone. The molten iron–silicon alloy collects in the hearth and is periodically drained from the surface.

Submerged-arc furnace

Ferrochromium. Many grades of ferrochromium are produced; the usual commercial alloys contain approximately 65 to 72 percent chromium and between 4.5 and 8.5 percent carbon. This high-carbon product is usually produced in a so-called three-phase Heroult-type electric-arc furnace having a carbon hearth. The charge is high-grade chromium ore, iron or iron ore, and a flux. The reductant is usually anthracite coal. Action of the arc heats the charge, and the oxides are reduced by the carbon. Molten ferrochromium collects in the hearth. When the reactions are complete, the metal is cast into slabs, and when cold these slabs are broken and the pieces crushed and sized for use. Ferrochromium low in carbon is produced by melting high-carbon chromium, chromium oxide, lime, and fluorspar in an electric furnace. The reaction between chromium oxide and carbon in the ferrochromium eliminates the carbon as carbon monoxide. The lime and fluorspar combine to form a slag. The molten metal contains from 0.2 to 2 percent carbon and is processed as described above. Ferrochromium with an intermediate carbon content can be produced in a submerged-arc furnace, but ferrosilicon and ferrochromium with a high level of silicon instead of carbonaceous materials are used as the reducing agents.

Ferrovanadium. Ferrovanadium is difficult to produce, because the metal is highly reactive. It is manufactured from a charge of ore, carbonaceous material, and fluxes in a small graphite-lined arc furnace. The alloy contains between 35 and 55 percent vanadium.

Ferrotungsten. Ferrotungsten is produced in a small, single-electrode arc furnace in which the hearth acts as one electrode. The charge consists of tungsten ore, carbonaceous material, iron ore, or iron and fluxes. The tungsten ore is easily reduced, but the alloy formed has a high melting point. During the heat, a chunk of solid alloy builds up on the hearth. When sufficient alloy has collected, the furnace is shut down and dismantled. The lump of alloy is crushed, cleaned, and prepared for shipment. The usual grade of ferrotungsten contains 70 to 80 percent tungsten and a maximum of 0.6 percent carbon.

Nickel. Nickel is often charged to the steel furnace as nickel oxide, which is easily and rapidly reduced to metal by carbon in the charge. Because nickel does not oxidize into the slag, nickel-steel scrap is usually charged to a heat when a nickel-bearing steel is made. Recently, metallic iron pellets containing 1 percent nickel have been used as a source of nickel when making low-alloy steels.

Other alloys. Alloys of many types are marketed for use in steelmaking. These contain barium, boron, calcium, niobium (columbium), molybdenum, magnesium, phosphorus, sulfur, selenium, tantalum, titanium, zirconium, etc. These metals may be in simple combinations with iron or with another metal, such as calcium silicide, or they may be in complex combinations as proprietary mixtures sold for special purposes.

IV. Steel metallurgy

STRUCTURAL FORMS OF IRON

Pure iron has three structural forms. The structure that is stable from room temperature to 911° C (1,672° F) is called alpha. The iron atoms in this structure are arrayed in a regular fashion on a lattice made up of simple cubes in which iron atoms are at the cube corners and one is at the centre of each cube. Accordingly, the alpha form of iron is said to have a body-centred cubic-crystalline structure. Between 911° C and 1,392° C (2,538° F), the gamma form is stable. In gamma iron the crystalline lattice is composed of identical cubes in which iron atoms are located at the corners and in the middle of each cube face; *i.e.*, the structure is face centred cubic. Between 1,392° C and 1,538° C (2,800° F) the stable structure is again the body-centred cubic form, which is designated delta iron. Iron melts at 1,538° C, and in the liquid there is no well-ordered arrangement of the atoms.

The physical and mechanical properties of a metal are influenced strongly by the nature of the atomic arrangement, and it is found that metals having the face-centred cubic structure are particularly ductile. The properties of a metal are also influenced significantly by defects in the lattice structure; much research has been directed to understanding the nature of these defects. Among the important defects are (1) a vacancy—*i.e.*, one atom is missing from a lattice point; (2) an iron interstitial, in which an iron atom is wedged between the regularly spaced atoms; and (3) large-scale effects called dislocations, one form of which is an extra plane of atoms that extends only partway through the crystal. The crystal structures of iron and steel have been explored by projecting beams of X-rays, neutrons, or electrons through thin foils or powdered specimens and observing the way that these beams are scattered and bent (diffracted) because of the regularity and irregularity of the atomic structure. Many changes in the properties of iron and of metals in general, such as hardening by cold work, brittle fracture, and plastic flow under stress, have been explained in terms of defects.

Structural defects

PHASES IN STEEL

Three phases may be found in unhardened steels: ferrite, austenite, and cementite. Ferrite is the body-centred cubic form of iron (alpha) in which some carbon has dissolved; the maximum amount at equilibrium is 0.025 percent at 727° C (1,341° F). Austenite is the face-centred form of iron (gamma) in which some carbon can dissolve; at 727° C the maximum amount is 0.76 percent, rising to 2.1 percent at 1,147° C (2,097° F). Cementite is a compound of iron and carbon having the formula Fe_3C. Cementite can decompose into ferrite and graphite below 727° C and into austenite and graphite above that temperature. The reactions are so slow, however, that they are not important in the processing of most steels. The iron–cementite phase diagram (Figure 11) shows the regions of stability of the phases found in slowly cooled steels. Below 727° C these steels, containing between 0.025 and 2 percent carbon, consist of ferrite and cementite. As has been noted already, with increasing carbon

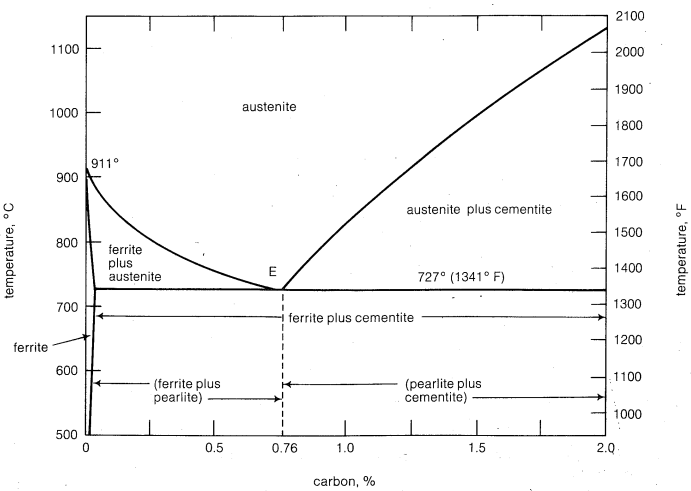


Figure 11: Steel portion of iron—cementite diagram (see text).

content, the amount of ferrite decreases and the amount of cementite increases.

A 0.76 percent carbon steel. The significance of other areas of the phase diagram can be illustrated with steels of three different compositions, beginning with a steel containing 0.76 percent carbon. Below 727° C this steel contains ferrite and cementite. When the steel is heated to just above 727° C (point E), the ferrite dissolves the cementite, and the whole structure is changed to austenite. The temperature at which austenite begins to form on heating is called the Ac_1 temperature (*c* for *chauffage*, or "heating"). Initially the austenite crystals are very small, but either holding the steel at this temperature for a long time or heating it to a higher temperature and holding for a shorter time causes some of the crystals to grow at the expense of others and the crystal size to increase. If this steel is then cooled to a little below 727° C, the austenite will decompose, forming alternate platelets of ferrite and cementite, in a structure called pearlite because of its appearance (see Figure 12A). This reaction of one solid phase decomposing to form two different solid phases is called a eutectoid change, and the temperature at which the austenite decomposition is complete is designated Ar_1 (*r* for *refroidissement*, or "cooling"). Because the reactions are somewhat sluggish, for a given composition, Ac_1 is slightly higher than Ar_1.

The distribution of the ferrite and cementite in pearlite is dependent on the rate of cooling and will be described later.

A 0.3 percent carbon steel. The second steel, one containing 0.3 percent carbon, will also consist of ferrite and cementite below 727° C. When heated just above that temperature, austenite containing 0.76 percent carbon forms by a reaction between the cementite and some of the ferrite. The resulting structure contains only ferrite and austenite. With increased temperature the ferrite dissolves in the austenite until at approximately 810° C (1,490° F) only austenite containing 0.3 percent carbon remains. The temperature at which the transformation to austenite is complete is given the designation Ac_3 (there is no Ac_2).

All carbon steels containing between 0.025 and 0.76 percent carbon undergo this same change on heating; the temperature (Cr_3) at which the steel becomes all austenite rises, however, as the carbon content decreases. This temperature range (see Figure 11) within which the structure changes on heating from ferrite plus cementite to ferrite plus austenite and then to all austenite is known as the critical range for carbon steels. The crystals of austenite again grow in size as the temperature of this steel is increased above 810° C. A steel containing less than 0.76 percent carbon that is slowly cooled from a

Critical range for carbon steels

D. Lloyd Thomas

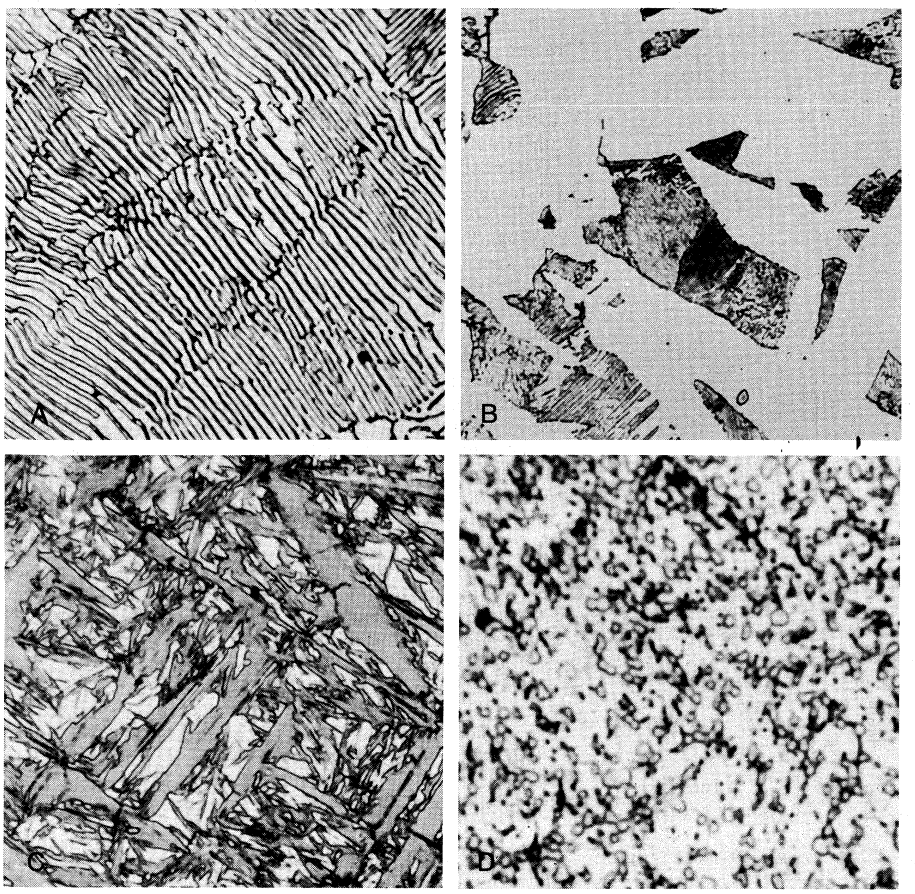


Figure 12: *Microphotographs of various steels.*
(A) Pearlite × 1,000; (B) ferrite (light) and pearlite (dark and lamellar) × 600; (C) martensite × 850; and (D) tempered martensite × 2,000.

temperature above the critical range will begin to form ferrite at the Ar_3 temperature. Very small crystallites of ferrite will form first at the boundaries of the austenite crystals. This ferrite is low in carbon so that the remaining austenite gradually becomes enriched in carbon, reaching 0.76 percent at 727° C (1,341° F).

A 1.1 percent carbon steel. The third steel contains 1.1 percent carbon. If cooled slowly, it, too, will contain only ferrite and cementite below 727° C. If this structure is heated to just above 727° C, the ferrite will react with some of the cementite to form austenite containing 0.76 percent carbon. The structure then will contain only austenite and cementite. With further heating the cementite will decrease, and at 860° C (1,580° F) the steel contains only austenite. When the steel is cooled from a temperature at which it is all austenite, cementite begins to form just below 860° C at the austenite grain boundaries. On further cooling, the amount of cementite increases, and the austenite is depleted in carbon until it reaches 0.76 percent at 727° C. The austenite then decomposes into pearlite, and the final structure is pearlite and cementite or ferrite and cementite, pearlite consisting, as noted earlier, of ferrite and cementite. In steels containing more than 0.76 percent carbon, the temperature at which the solution of cementite in austenite is complete is designated Ac_{cm}, and the temperature at which cementite begins to form from austentite is designated Ar_{cm}.

EFFECTS OF CARBON CONTENT ON STRUCTURE AND PROPERTIES OF STEEL

Slowly cooled steels. At temperatures below 727° C, slowly cooled steels containing more than 0.025 percent carbon contain ferrite and cementite. Ferrite is soft and ductile, and cementite hard and brittle. The amount of cementite seen under the microscope increases with the amount of carbon in the steel. A steel with 0.3 percent carbon contains approximately 94 percent ferrite and 6 percent cementite, whereas at 1 percent carbon these amounts are 85 percent ferrite and 15 percent cementite. In a slowly cooled steel the austenite remaining at 727° C will transform into hard and strong pearlite that consists of alternate layers of ferrite and cementite (Figure 12A).

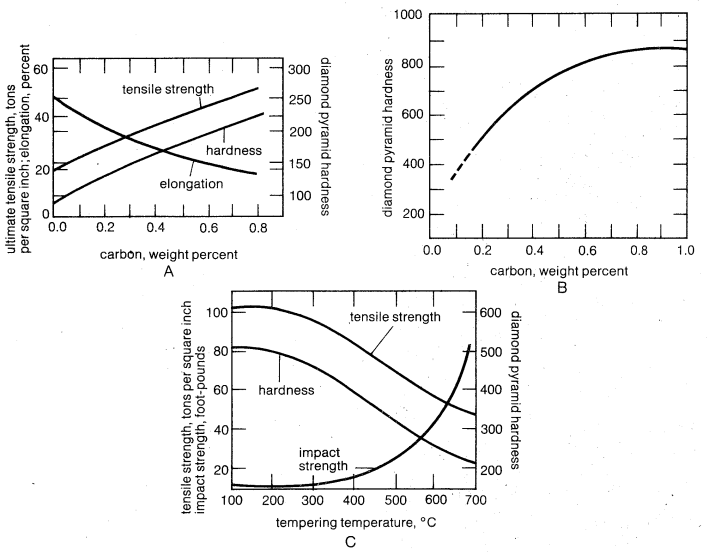


Figure 13: *Effects of composition and heat treatment on mechanical properties of steels.*
(A) Mechanical properties of annealed steel containing various percentages of carbon. (B) Hardness of martensite as a function of carbon content. (C) Effects of tempering temperature on mechanical properties of tempered martensite.

With increased carbon content up to 0.76 percent, slowly cooled steels contain an increased amount of pearlite and less ferrite (Figure 12B); strength and hardness increase, but ductility and toughness are reduced (see Figure 13A). Steels having maximum ductility and toughness are low-carbon steels that have been slowly cooled or annealed below 727° C (1,341° F). The final structure of a steel will be coarse if the austenite crystals were coarse at the beginning of cooling. Hardness and tensile strength are not affected materially by the temperature at which the steel is held before cooling, nor do they depend upon the size of the austenite crystals. The finer the final structure, however, the tougher will be the steel when subjected to shock at and below room temperature. Because recrystallization takes place both on heating (Ac_1 to Ac_3) and cooling (Ar_1 to Ar_3), a coarse structure can be converted to a fine structure by reheating the steel to just above its critical range of temperature and then slowly cooling it.

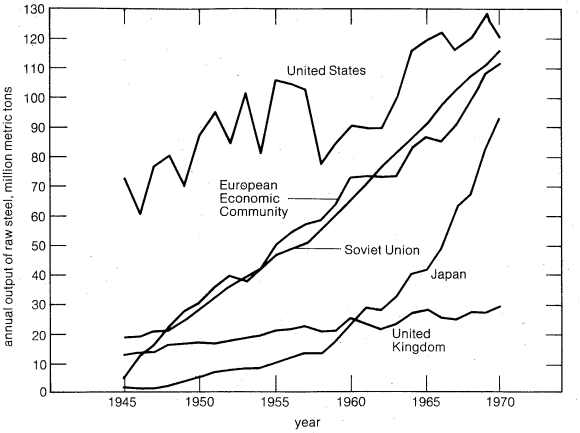


Figure 14: Growth of raw-steel production of major producing countries since World War II.
By courtesy of the American Iron and Steel Institute

A steel containing more than 0.76 percent carbon when cooled from a temperature where it is all austenite will first precipitate cementite at the austenite grain boundaries. With further cooling the structure will consist of grains of austenite surrounded by cementite. Subsequently, the austenite will form pearlite, and the final structure will consist of islands of pearlite in a continuous network of cementite. Though steel formed this way is strong, it is also very brittle, so the cementite is dispersed by kneading the steel. It is important that these steels not be reheated above 760° C (1,400° F) in subsequent heat-treatment; otherwise, the embrittling network will form on cooling.

Hardened steels. *Carbon steels.* The temperature and the nature of the decomposition of austenite into pearlite in plain carbon steels depends strongly upon the rate of cooling. With increased cooling rates the Ac_1 temperature can be lowered progressively to about 550° C (1,022° F), and the pearlite will become so fine that it cannot be resolved under the light microscope. Accompanying this change is increased hardness and strength of the pearlite. With a further increase in the cooling rate, the transformation of austenite to pearlite may be suppressed altogether. A broad range of heat-treatments are used to take advantage of these effects. A steel is annealed by cooling it slowly in a furnace from just above its critical temperature range. The result is coarse pearlite, and the steel is soft. A steel cooled in air from just above its critical temperature range is said to be normalized. The result is a finer pearlite, and the steel is harder, stronger, and less ductile than it would be if annealed. By quenching the steel very rapidly in cold water or brine, the formation of pearlite may be partially or completely suppressed, and austenite can be retained down to about 200° C (400° F). At that temperature, the austenite begins to form a structure called martensite (Figure 12C and D). The transformation from austenite to martensite is complete at about room temperature. The minimum cooling rate necessary to suppress the formation of any pearlite is called the critical cooling velocity for the steel. Quenching to form martensite is often used because of the special properties obtained when the steel is annealed.

Suppression of pearlite formation

Martensite results when face-centred cubic austenite transforms to the body-centred structure without cementite being precipitated. Carbon is retained in the structure, which is stretched slightly so that it is no longer cubic;

Forming martensite

thus, martensite can be thought to be ferrite supersaturated with carbon. This supersaturation results in severe distortion of the lattice, which increases with increasing carbon content of the steel. The hardness of martensite increases to a maximum as the carbon content of the steel increases. To obtain a completely martensitic structure in a steel containing less than 0.8 percent carbon, the steel must be quenched from a temperature above its critical temperature range. This is necessary to suppress the formation of both ferrite and cementite. Steels containing more than 0.8 percent carbon are quenched from 760° C (1,400° F), usually immediately after forging, thus producing a structure of cementite dispersed in a matrix of martensite. The cooling rate necessary to obtain martensite in plain carbon steels very often causes distortion of the part being heat-treated. In severe cases, cracking may also result, especially in a part with a complex shape. Thus, even with the most severe quench in brine or cold water, a maximum thickness of a plain carbon steel that can harden fully to martensite is only about 20 millimetres (0.8 inch), because the cooling rate decreases from the surface to the interior of a part.

Low-alloy steels. Nickel, chromium, manganese, and molybdenum in steels reduce the critical cooling rate necessary to obtain martensite. These elements are said to increase the steel's ability to harden. On the other hand, they do not increase the hardness of the martensite that forms. The use of these alloying elements singly or in combination reduces the tendency for distortion and cracking because a less drastic quench is required to obtain martensite. A quench in oil or air is usually sufficient. Alloy steels containing nickel, chromium, and molybdenum in sections 150 millimetres (about six inches) thick can be made martensitic throughout by an oil quench. Because alloy steels are much more expensive than carbon steels, they are used only when their special properties are required.

Hardened and tempered steels. Martensite is usually much too brittle and too hard for most uses, and residual internal stresses are present in the steel as a result of the rapid cooling necessary to form the martensite. Reheating the steel to a selected temperature in the range of 200° to 600° C (400° to 1,100° F) and then cooling it in air will relieve the internal stresses and decrease the hardness and strength of the steel. A marked increase in toughness is obtained, however. This operation is called tempering. The temperature and time for the tempering operation must be carefully selected on the basis of the type of steel and the final properties desired. A tempering temperature near 680° C (1,260° F) causes the greatest loss in hardness and strength but imparts the greatest toughness as measured by impact strength (see Figures 12D, 13C). Tool steels are tempered in the range of 200° to 250° C (400° to 500° F) to relieve internal stresses without the loss of hardness. Spring steels are tempered at 350° to 500° C (650° to 900° F) to obtain good strength without a serious loss of impact properties. Machine parts subject to impact loadings may be tempered in the range of 500° to 600° C (900° to 1,100° F).

In tempering, martensite decomposes into ferrite and cementite. The amount of decomposition increases as the tempering temperature increases up to 400° C (750° F). In the decomposition, carbon forms small particles of cementite, and the highly stressed martensite relaxes to ferrite. The product is a structure consisting of very fine particles of cementite imbedded in a matrix of ferrite. The structure obtained in tempering at a low temperature is so fine that it can be resolved with the electron microscope but not with the optical microscope. The structure is coarser with higher tempering temperatures, and those obtained at 600° C (1,100° F) can be resolved in the optical microscope at 1,000 magnifications (see Figure 12D).

Mechanical properties of tempered alloy steels

The mechanical properties obtained after tempering alloy steels containing nickel and chromium are much superior to those of the carbon steels. This is because these alloying elements strengthen the ferrite matrix. At the same strength, alloy steels are much tougher than carbon steels. The highly alloyed steels are softened much less by tempering than are carbon and low-alloy steels. High-speed tool steels, for example, are quenched from 1,300° C (2,400° F) and tempered at 500° to 600° C (900° to 1,100° F) with little if any loss in hardness. This resistance to softening at high temperatures makes these steels useful as cutting tools where very high tool temperatures are encountered at high cutting speeds.

The phenomenon known as temper brittleness occurs when steels containing relatively high levels of phosphorus, manganese, or chromium are tempered in the range of 400° to 550° C (750° to 1,000° F). It also occurs if these steels are tempered at a higher temperature and then are cooled slowly through this range. The effect can be avoided by tempering at a temperature above 550° C, followed by a quench from the tempering temperature. Molybdenum in the steel reduces the tendency toward temper brittleness.

SURFACE HARDENING

In many instances steel parts such as gear teeth, the teeth on power shovels, and the tires of railroad wheels are subject to severe wear or abrasion at the surface as well as shock and dynamic loading. The needs for high hardness at the surface and a tough interior are conflicting requirements, but they can be met by treating the surface to obtain a hard "case" while the interior of the steel part remains tough. Common methods for accomplishing this are described briefly below.

Carburizing. Carburizing is the modern form of an old cementation process. The part, made from a carbon or alloy steel low in carbon, is heated to about 900° C (1,650° F) in a bed of carbonaceous material or a carburizing atmosphere. Carbon diffuses into the surface, the amount of carbon added and the depth of penetration depending upon the carburizing medium, the temperature, and the time of exposure. The resulting case will contain about 0.8 percent carbon and be fully martensitic when the part is quenched to room temperature. Because of its lower carbon content, the core will not be fully hard and will be tough. A variety of heat-treatments after carburization are used, depending upon the combination of properties desired for the case and core.

Nitriding. A steel containing small amounts of chromium, molybdenum, and aluminum will form a very hard, thin case on the surface when exposed to an atmosphere of ammonia. This case consists of nitrides of aluminum and chromium formed when nitrogen obtained from the decomposition of the ammonia diffuses into the steel. The depth of the case increases with the duration of the nitriding process. The procedure is first to quench and temper the steel part at 550° to 650° C (1,000° to 1,200° F) to obtain a tough case. The hard case is then obtained by exposing the part to an atmosphere of ammonia for about 80 hours at 500° C (900° F). Nitriding has the advantage over carburizing in that the case remains hard when exposed to temperatures up to 500° C, whereas carburized cases begin to soften at about 200° C (400° F).

Nitriding procedure

Flame and induction hardening. In this method a part is made from a medium-carbon steel, with or without alloying elements. The part is quenched and tempered to produce the desired case properties. Then the portions of the surface that are to be hardened are quickly heated to above the Ac_1 temperature of the steel and quenched in oil or water. The heated region becomes martensitic, and the core retains its toughness. Heating is accomplished by numerous small gas flames or by high-frequency induction.

Other methods. As noted in the history section of this article, Hadfield manganese steels, which contain approximately 1.4 percent carbon and 14 percent manganese, are extremely tough and wear-resistant. These properties are obtained by heating the steel to between 1,000° and 1,060° C (1,800° to 1,940° F), followed by a water quench. This treatment preserves the austenitic structure down to room temperature. Abrasion of the surface and cold work causes the surface to transform to martensite. Thus, the hard surface is continually renewed as the surface wears away.

A number of other methods are also used to obtain a hard surface. One way is to apply a hard facing to a steel part by welding, brazing, or spraying a coating of hard material. Materials used for this are tungsten carbide and Stellite, tungsten, chromium, and carbon.

RESEARCH AND DEVELOPMENT

Steels face vigorous competition in performance and cost from many types of materials, ranging from other metals, such as aluminum and titanium, to plastics, glass, and concrete, and the industry finds that it is constantly necessary to improve processes in use for producing steels and to devise new steel compositions and ways to produce and process them. Much applied research and development work is going on throughout the world to meet these aims.

New methods of processing ores and producing agglomerated products for the blast furnace are being developed. Methods for producing and utilizing directly reduced ores are being tested. Experimentation on methods for continuous production of steel is in progress, and improved methods for operating and controlling the various processes in use for steelmaking are under development. Particular attention is being directed to developing methods for automatic control of basic-oxygen and electric-furnace steelmaking. Improved methods of vacuum process and refining steels have been developed recently and are being further improved. A great deal of effort is being devoted to describing and controlling the solidification of steels. Laboratories are working on new alloy compositions, particularly for the purpose of obtaining improved properties by new processing methods and with the use of lower levels of alloying elements or using less costly alloying elements. In recent years much attention has been devoted to developing new and improved coatings, such as chromium, aluminum, plastic, and thinner tin coatings for flat-rolled products. A strong trend toward the use of plates and sheets instead of rolled sections in the production of such items as structural members and building walls has required the development of improved methods of forming and welding.

Continuous production of steel

Currently, a great deal of money and effort are being directed to means of alleviating the problems of pollution caused in steelmaking. A major task is to avoid air and water pollution associated with conventional methods for coking coal. Modifications in the methods of charging and discharging the ovens are receiving much attention, and alternative methods for processing coke are being tested. The disposal of spent liquor from pickling of steels and pollution arising from sulfur in fuels are also receiving major attention.

Reducing pollution

Research of a more fundamental nature has been devoted to understanding the chemical processes of iron-ore reduction, the chemistry of the reactions in steelmaking, the details of how alloys of iron solidify, the chemical properties of steel, and how the details of the atomic structure of phases present in steels are related to the physical and mechanical properties of steels. Attention is being given to the reactions of steels with various environments and to the mechanisms by which steels corrode.

V. World steel production

During the early 18th century, Sweden was the chief producer of the relatively small volume of steel made in the world. With the development of mass-produced iron and steel, Great Britain became dominant. It has been estimated that the output of pig iron in Great Britain rose from 28,000 tons per year in 1720 to 75,000 tons in 1788 and to almost 3,000,000 tons by 1852 (unless otherwise noted, figures are in short tons). It is notable that in the early part of the 19th century much of this iron was exported: 20 percent in 1806 and almost 50 percent in 1856. Pig-iron production continued its rapid increase to 7,100,000 tons in 1875, when the United Kingdom produced almost half the world's output (see Table 1 for selected world figures). In that year, production in the United States was 2,200,000 tons, or 15 percent of world production. In 1880 British production reached 8,700,000 tons and then rose gradually to about 10,000,000 tons during the remainder of the century. During this time the British share of the world's production of pig iron declined from 42 percent in 1880 to approximately 22 percent as production of pig iron in the United States rose rapidly to over 15,000,000 tons in 1900, making that country's share of world production about 35 percent. Great Britain's annual output of pig iron was first surpassed by that of the United States in 1890, by Germany and France in 1921, when the coal supply was interrupted by a strike, by the Soviet Union in 1931, and by Japan in 1954.

British pig-iron production in the 19th century

Table 1: World Output of Pig Iron for Selected Years

year	metric tons per year (000,000)	year	metric tons per year (000,000)
1820	1.02	1925	76.9
1830	1.83	1929	99.9
1840	2.74	1930	80.5
1850	4.78	1935	74.2
1860	7.34	1940	104.1
1870	12.03	1943	118.8
1880	18.45	1945	80.6
1890	27.2	1950	136.3
1900	40.5	1955	215.8
1910	65.8	1960	264.5
1920	63.9	1965	326.2
		1970	407.5

Source: *U.S. Bureau of Mines Yearbook; Annual Statistical Reports*, American Iron and Steel Institute (pig iron and ferro-alloys).

Great Britain also was pre-eminent in the production of steel in the 19th century. The older methods for making iron and steel products with a great expenditure of labour rapidly gave way to the Bessemer and open-hearth processes after 1870. In 1875 Great Britain produced 790,000 tons of steel, or 40 percent of world production. In 1896 the output rose to about 4,400,000 tons, or approximately 22 percent of the world's total. U.S. production of steel and wrought iron had passed that of the British by 1880. British production continued to rise, however, to about 7,000,000 tons in 1900, 18,000,000 tons in 1950, 27,000,000 tons in 1960, reaching a high of 31,000,000 tons in 1970 but declining in 1971 to less than 27,000,000 tons.

Table 2 shows the production by processes in Great Britain from 1896 to 1906. During the interval, the increase in production was chiefly due to the increased output of open-hearth steel. In 1905 the output of Bessemer steels reached its peak for the 20th century, and, subsequently, the basic open hearth became the primary process for steel production in the United Kingdom.

From 1896 to 1970 the United States was the largest producer of steel in the world. In 1971, however, the Soviet Union surpassed the United States. The rise and decline of processes in use in the United States for the past 100 years illustrates in a general way the changes that have occurred throughout the world. The first production of acid Bessemer steel was in 1867, when 30,000 tons were made. In 1870, production of acid open-hearth steel was begun, and the output was 1,500 tons. In the same year, 42,000 tons of acid Bessemer steel were made, and the output by this process rose rapidly, reaching its all-time peak of 13,800,000 tons in 1906. In the interval from 1870 to 1906 the acid Bessemer was the primary steel-producing pro-

Table 2: Production of Steel and Wrought Iron in Great Britain
(in 000 metric tons)

year	open hearth			Bessemer			total steel	wrought iron
	acid	basic	total	acid	basic	total		
1896	2,180	174	2,354	1,380	465	1,845	4,209	1,234
1900	2,912	298	3,210	1,275	499	1,774	4,984	1,183
1905	3,092	808	3,900	1,418	588	2,006	5,906	954
1906	3,435	1,195	4,630	1,328	610	1,938	6,568	1,026

Source: J.C. Carr and W. Taplin, *History of the British Steel Industry* (1962).

cess in the country. Subsequently, it has declined in output, becoming of less relative importance each year until production by this method ceased in 1967. Initially, the use of the open-hearth process grew more slowly, because it was operated with acid hearth refractories and an acid slag and, as a result, was not competitive with the acid converter. Its growth quickened after 1884, when it was found that the process would operate satisfactorily with basic refractories. The basic furnace rapidly replaced the acid furnace for most purposes in the United States, because it could treat readily the moderate levels of phosphorus in pig iron made from the ores from the Great Lakes ranges. These ores were used extensively because they were plentiful, cheap, and of good quality. The basic open-hearth process surpassed the acid Bessemer in output in 1908 and was the primary means for producing steel in the United States until 1970, when it was surpassed in output by the basic oxygen process. The reasons for the long ascendancy of the basic open hearth has been its flexibility in producing virtually any type of carbon steel and a number of the low-alloy steels, its flexibility in utilizing scrap, hot metal, and cold iron in virtually any proportion, and its ability to make products of good quality. Between 1928 and 1961 the open-hearth process produced over 85 percent of the steel made in the United States each year.

Growth of oxygen steel

The use of the basic oxygen processes by the U.S. steel industry did not begin until 1955, six years after the basic oxygen furnace (or L–D furnace) had been developed in Austria in 1949. The growth in output of oxygen steel by the basic oxygen furnace has been dramatic in the subsequent 15 years. Production rose from 725,000 tons in 1955 to over 70,000,000 tons in 1971. The primary advantages of the basic oxygen furnace over the open hearth are its rapid operation, lower cost, and ease of control. It is best suited to the production of low-carbon steels, which are used in large quantities today. To meet the disadvantage of its inability to operate normally with more than 35 percent scrap in the charge, a number of steel plants have also installed basic electric furnaces for the production of carbon steels. With the two processes in effect, the inability of the oxygen furnace to melt scrap can be avoided.

High-quality and specialty steels were produced in the United States by the crucible process between 1870 and 1915. Growth was slow, and the peak in output of 146,000 tons was reached in 1907. Production held steady at approximately that level until 1920, but thereafter it declined rapidly to a few thousand tons a year. The commercial use of the electric-arc furnace for producing alloy and specialty steels began in 1906. Output by this process exceeded that by the crucible process in 1916 and increased rapidly to an average of approximately 650,000 tons per year, which was sustained from 1918 to 1938. The demands of World War II pushed production to about 4,500,000 tons in 1943, and the output declined only slightly in the immediate postwar years. The rapid growth in production between 1950 and 1970 resulted from the use of the process for melting carbon steels with scrap as the principal charge. The use of the electric furnace was expected to increase in the decade of the 1970s by increased production of metallized ore and increased recovery of iron and steel scrap from waste as a result of efforts to improve the environment.

In the year 1900 the basic Bessemer process produced 65 percent and the basic open hearth produced 31 percent of the total of 7,500,000 tons of steel made in Germany. In 1905 for France the corresponding figures are 60 percent and 34 percent, respectively, and the total production of steel that year was 2,500,000 tons. The balance of the output was by acid methods. With the increased importation of ores low in phosphorus from abroad in recent years, European steel plants have begun to use the basic oxygen furnace and the Kaldo process more extensively. The basic oxygen furnace with a two-slag operation is also being used to refine the high-phosphorus hot metal obtained from the local ores. Recently, the use of oxygen and steam, and oxygen and hydrocarbons in the basic Bessemer has received renewed interest. In the Soviet Union, the predominant steelmaking process continues to be the basic open hearth. In Japan, where steelmaking facilities were heavily damaged or destroyed in World War II, virtually the whole of the production of carbon steels is by the basic oxygen furnace. Electric furnaces are being used to produce alloy and stainless steels, but a limited tonnage of stainless steel is also produced in the basic oxygen process. Throughout the world the acid processes have given way almost completely to the basic processes, the basic open hearth, the basic Bessemer, the basic oxygen processes, and the basic electric-arc furnace. World production of steel by the basic oxygen furnace was estimated to be 280,000,000 tons in 1971, which was 44 percent of the total production of steel in that year.

World production of iron and steel has maintained an upward trend except for brief periods following World Wars I and II and in 1929 to 1932 during the Great Depression. The production of pig iron has been closely related to the production of steel, because it is used primarily as a raw material in steelmaking. In recent years the large industrial producers of cast-iron products have obtained the iron by melting steel scrap and then carburizing and alloying it. Discarded cast iron from the foundry is sent back to the melting furnace, but there is relatively little pig iron in the charge.

The leadership of the United States in the production of steel was increased after World War II, because the steel-producing facilities of all of the other major industrialized countries had been damaged to some degree in the war. In 1950 the United States produced 48 percent of the world total; in 1960, 27 percent; and, in 1970, 21 percent. The growth in output for each of the principal producing countries and the European Economic Community (Belgium, France, Italy, Luxembourg, The Netherlands, and West Germany) from 1945 to 1970 are shown in Figure 14. In that 25-year span the annual growth rate for the United States was approximately 2.3 percent, and after 1960 the rate was approximately 4 percent. It appears, however, that there has been no growth between 1966 and 1970. In those four years the production of steel in the Soviet Union grew at an annual rate of over 4 percent, that of the European Economic Community almost 7 percent, that of Japan 17 percent, and that of Great Britain slightly over 3 percent. The total world production of steel in 1970 was 634,000,000 tons, and its annual growth rate was approximately 6 percent. The economic conditions in 1971 reduced the rate of growth to a virtual standstill, and the uncertainty of the status of exports and imports of steel for many countries make it difficult to predict whether the growth rates of the 1950s and 1960s can be reached again, or sustained, worldwide.

The place of steel in international trade has changed since World War II. Just after the war, the United States was a major exporter of steel. With the rapid growth of steelmaking capacities in other highly industrialized countries, this has been reversed, and the difference between imports and exports of steel products for the United States has ranged from 6,500,000 to over 16,000,000 tons between 1965 and 1971. The flow of steel products between countries depends strongly on the economic climate and varies greatly from year to year. In the past, major steel-producing countries have always exported steel products to the less industrialized countries. To aid in reducing the drain on their currencies because of the importation of steel products, however, many small countries construct small steel plants using electric furnaces to melt scrap and metallized local ores.

BIBLIOGRAPHY. General information on the production and consumption of iron ores, iron, and steel may be found in publications of the United Nations, the International Iron and Steel Institute, and the United States Bureau of Mines. The best general source on the technology of iron and steelmaking in the United States is H.E. MCGANNON (ed.), *The Making, Shaping and Treating of Steel*, 8th ed. (1964). Excellent coverage of the properties of steels, methods of processing, metallography, and testing and inspection may be found in TAYLOR LYMAN (ed.), *Metals Handbook*, 8th ed. (1970). Treatment of the current technology is covered in the Swedish periodical *Scandinavian Journal of Metallurgy* (bimonthly); the *Journal* of the British Iron and Steel Institute

monthly); the *Journal* of the Japanese Iron and Steel Institute (monthly); *Tetsu to Hagane* (bimonthly); various publications of the American Mining, Metallurgical, and Petroleum Engineers—in particular, the annual proceedings of this society dealing with ironmaking, steelmaking, and mechanical working; and G.J. DERGE (ed.), *Basic Open Hearth Steelmaking*, 3rd ed. (1964); and C.E. SIMS (ed.), *Electric Furnace Steelmaking* (1962). Leading journals of the trade are *Iron Age, Iron and Coal Traders Review, Stahl und Eisen, Magazine 33, Stahl in English,* and *Iron and Steel Engineer*. A.K. OSBORNE, *Encyclopaedia of the Iron and Steel Industry* (1956), will be helpful to the student. Earlier publications that may also be of interest are L. AITCHISON and W.I. PUMPHREY, *Engineering Steels* (1953); H.P. TIEMANN, *Iron and Steel*, 3rd ed. (1933); R. FORSYTHE, *The Blast Furnace and the Manufacture of Pig Iron*, 3rd ed. (1922); and A.W. and H. BREARLEY, *Ingots and Ingot Moulds* (1918).

(J.F.E.)

Stefan Dušan

Stefan Dušan, Serbia's greatest ruler, who added Macedonia, Albania, and large parts of Greece to the Serbian empire in the 14th century, is remembered above all as a conqueror and a codifier of laws.

By courtesy of Jugoslavenski Leksikografski Zavod, Zagreb, Yugoslavia



Stefan Dušan, detail of a fresco in the Monastery Lesnovo, Yugoslavia, 1349.

Stefan Dušan was born in 1308, the son of Stefan Uroš III, who was the oldest son of the reigning king, Stefan Uroš II Milutin. While Dušan was still a boy, his father, who governed the maritime provinces of the Serbian state, rebelled against his own father. Milutin took him prisoner, blinded him in order to make him unfit to claim the throne, and about 1314 exiled him to Constantinople. With his father and mother, Stefan Dušan therefore spent several years in the Byzantine capital. Life in Constantinople as an exile, without any prospect of occupying the throne, had a lasting influence on the formation of his character. In addition to a basic education, he acquired a familiarity with the ways of government of the Byzantine world by which he was later to be guided. The exile in Constantinople ended with the reconciliation of Dušan's father and grandfather and the family's return to Serbia about 1320. After Milutin's death in 1321, Dušan's father greatly increased his chances in the contention for the crown by demonstrating that he was not blind, claiming a miraculous cure. With the support of a great majority of the nobility, he succeeded in defeating his enemies and was crowned king early in 1322. Dušan, still a youth, was crowned "young king," or heir apparent.

Background and early years

Too young to be able to pursue more active policies, Dušan as "young king" governed the maritime provinces of the state. He had to reconcile himself to the loss of Hum (Hercegovina), the most westerly region of Serbia, which the ruler of neighbouring Bosnia conquered in 1326. He gained valuable military experience, however, and the reputation of an able commander in the campaigns against the Bosnians; he particularly distinguished himself in the great battle against the Bulgarians of Velbužd in 1330. Although this victory freed Serbia of the great danger of an allied attack by both the Bulgarian and the Byzantine emperors, dissension soon arose between Dušan and his father. War broke out between them in the fall of 1330. Peace was concluded in the spring of 1331, but soon afterward Dušan rose again against his father and deposed him.

Dušan's reign began on a note of peace in September 1331. He subdued the sporadic revolts of the nobility, who had become more powerful during the period of civil wars, and strengthened his alliance with the new Bulgarian emperor, John Alexander, by marrying his sister Helen in 1332. Relations with Bulgaria remained untroubled to the end of Dušan's reign.

In 1334, however, he began his war of conquest against Byzantium. After taking the border fortresses, Dušan penetrated deep into Byzantine territory to the gates of Salonica, although he did not achieve lasting success in subduing the cities. He made peace with Emperor Andronicus III Palaeologus in August 1334, having extended the territories of his state close to what is today's Yugoslav-Greek border. The next year he secured his northern borders against the Hungarians. During the course of his entire reign, Dušan remained on the defensive in the north, although he succeeded in preventing Hungary from extending its boundaries south of the fortresses on the rivers Save and Danube.

With the death of Andronicus III, in 1341, the Byzantine Empire once more fell prey to family quarrels and civil war. Dušan, arriving before the gates of Salonica, received an unexpected ally in John Cantacuzenus, the late Emperor's general, who took up arms against the regents of the young successor, John V Palaeologus, and proclaimed himself emperor. Dušan aided the Byzantine pretender, but their alliance broke up in 1343, and they became bitter enemies. On his own, Dušan conquered Albania and a greater part of Macedonia in the same year. After conquering the stubbornly defended city of Seres in the fall of 1345, Dušan began to call himself "Emperor of the Serbs and Greeks." He was crowned emperor in April 1346. After subjugating other parts of the Byzantine Empire, Epirus and Thessaly, in 1348, he found himself master of a vast territory from the rivers Save and Danube to the Bay of Corinth and from the Adriatic and Ionian shores to the Aegean. Although Dušan was helped by the weakness of his adversaries, he would never have attained his great successes without the forces provided by the militant nobility of the old Serbian provinces and by his German mercenaries. Dušan regarded the subdued Byzantine lands, "Romania," as his own "imperial lands," whereas the old Serbian provinces were formally the "royal lands" of his son. In reality, he governed both territories himself, but his administrative reforms concerned mostly the Serbian provinces. On the eve of the coronation he raised the Serbian archbishopric to the status of a patriarchate, for in the Orthodox East an emperor was inconceivable without a patriarch. He introduced the Byzantine system of titles and ranks, and the imperial chancellery was organized on the Byzantine model, as was the uniform organization of local authorities. Dušan, moreover, considered the introduction of a law code part of his imperial duties, and in the diets of 1349 and 1354 he promulgated a code containing more than 200 statutes. The code covered mainly criminal law and the relations between the classes.

Emperor of the Serbs and Greeks

From the time he imposed himself as co-ruler of Constantinople with the young emperor John V Palaeologus, John VI Cantacuzenus had posed a serious threat to Dušan, especially after he acquired Turkish auxiliary forces, which had twice defeated the Serbs—in 1344 and in 1352. During Dušan's campaign against Bosnia in 1350, Cantacuzenus' supporters in the Greek cities sent their armies to attack him in the south, thus forcing him to return to Macedonia and to give up his plans for reconquering Hercegovina. Plans for the conquest of Con-

stantinople drew Dušan into close relations with Venice. The republic was sympathetic to him but remained indifferent toward his proposals for direct action against the Byzantine emperors. In his last years Dušan maintained cordial relations with the Pope in Avignon. Like some of his predecessors he showed a readiness to accept a union of the churches, especially since in 1350 the patriarch of Constantinople had proclaimed an anathema on both the Serbian emperor and the patriarch. Enmity between local Orthodox and Catholic clerics stood in his way, as did Hungary's unsuccessful attack on Serbia in 1354, which upset the work of the papal legation. This attack also frustrated Dušan's desire to be designated by the Pope as captain of a crusade against the Turks, who at this time had gained a foothold on European soil and had become a danger to the Christians of the Balkans. Dušan died suddenly on December 20, 1355. He was succeeded by his son Uroš, who was unable to preserve his father's large empire. The peripheral regions of the empire seceded under their powerful governors, and large areas fell to the Turks. Only 15 years after Dušan's death, the territory of Serbia was smaller than that from which he had set out on his conquests.

Assessment

Dušan's popularity was especially great during the era of national awakening and wars for liberation in the 19th century. "Dušan the Mighty" represented at that time a symbol of former glory and greatness, while the restoration of Dušan's empire was stressed as the ideal goal of contemporary Serbian politics. Modern historians stress his accomplishment in elevating the Serbian state to predominance in the Balkans and in introducing internal reforms, transplanted from Byzantine practices and institutions and aimed at centralizing and unifying the administrative system. Outlasting his other accomplishments was Dušan's law code, which remained in use until the end of the medieval Serbian state and later became a valuable source for the study of medieval society.

BIBLIOGRAPHY. There is no modern biography of Stefan Dušan. The chapter on Dušan in CONSTANTIN JIRECEK, *Geschichte der Serben I* (1911) is well documented. Based on Jireček are the following English accounts: H.W.V. TEMPERLEY, *History of Serbia* (1917); and WILLIAM MILLER, "The Medieval Serbian Empire," *Quarterly Review*, 226:488–507 (1916). In the short summary, "The Balkans, 1018–1499," in *The Cambridge Medieval History*, vol. 4, pt. 1, pp. 537–541 (1966), M. DINIC utilized results of recent Serbian scholarship. Dušan has been treated in G. OSTROGORSKY, *Geschichte des byzantinischen Staates*, 3rd ed. (1963; Eng. trans., *History of the Byzantine State*, rev. ed., pp. 499–533, 1969), with selected bibliography. Very useful also is G. SOULIS, "Tsar Stephen Dušan and Mount Athos," *Harvard Slavic Studies*, 2:125–139 (1954).

(S.M.Ci.)

Steichen, Edward

A pioneer of American photography for over 70 years, Edward Steichen occupied an eminent position at every stage of its development. He was born in Luxembourg on March 27, 1879; but in 1882 his parents moved to the United States and settled in Hancock, Michigan. Later the family moved to Milwaukee, where, at the age of 15, he was apprenticed to a commercial lithographer. He studied painting but became interested in photography, and by 1900 several of his pictures had been shown in Philadelphia and Chicago. Steichen's early photographs were influenced by his training as a painter. He frequently brushed on silver salts and other chemicals to eliminate or add images to negative or paper to achieve prints resembling soft, fuzzy mezzotints or loose wash drawings. These misty pictures were considered by most critics of the day to be the highest achievement of photographic art. Alfred Stieglitz, the best known American photographer of the time—although himself a purist who did not tamper with either his negatives or his prints—concurred. Stieglitz encouraged photographers to experiment, for he believed it justifiable to work on negative or print to obtain the desired image.

In 1902 Stieglitz invited Steichen to join him and 11 other photographers in founding the Photo-Secession, an organization dedicated to promoting photography as a fine art. In the course of time, the Photo-Secessionists did indeed succeed in transforming American photography from a craft that recorded and increasingly embellished to an art that freely experimented and explored. They showed that photographs of ordinary subjects and everyday scenes printed from the negative without manipulation can exert a far greater aesthetic appeal than artificial compositions printed by arcane methods.



Steichen, 1960.
Joanna T. Steichen

Member of Photo-Secession

In 1905 Photo-Secession opened, at 291 Fifth Avenue, New York, its Little Galleries (called "291" for short). It was designed by Steichen, and Alfred Stieglitz was its director. Steichen, who had become a friend of the sculptor Auguste Rodin in 1910 when he was studying painting in Paris, arranged exhibitions of the work of Rodin and that of other friends at the gallery, including paintings by Henri Matisse, Gordon Craig, Max Weher, Paul Cézanne, Pablo Picasso, and sculpture by Constantin Brancusi. Many of these avant-garde artists were shown for the first time in America at "291," starting eight years before the Armory Show of 1913, at which modern European art received its first full-scale display in the United States. The Photo-Secessionist venture was successful; for the first time in the United States, photographs by Photo-Secessionists, Steichen prominently among them, ranked with paintings and drawings as artistic works.

During World War I, Steichen was put in command of the photographic division of the United States Army's air force and retired as a lieutenant colonel in 1918. During the war, realism had been the only purpose of the pictures Steichen took. The experience led to a significant change in his mode of artistic expression—in fact, he made a bonfire of all his paintings. He ceased painting and renounced his impressionistic photography with its broad, simple effects. Instead, he concentrated on creating a precise, clean-cut, more realistic picture. In order to achieve maximum realism, he photographed a white cup and saucer against a black velvet background more than a thousand times in his search for the perfect rendering of the subtle gradations in white, black, and gray. From 1923 to 1938 he operated a commercial studio in New York City that specialized in portraits and advertising illustrations and soon became a successful enterprise. What he knew of art and design as a painter, combined with his masterful photographic technique, enabled him to produce memorable portraits of the literary and artistic personalities and the socially prominent who came to pose. In addition, he was in 1923 induced by Frank Crowninshield, the editor, to become chief photographer for the fashion magazines *Vogue* and *Vanity Fair*. Steichen's photographs of the 20s and 30s form a vital pictorial record. He created such striking images as the veiled Gloria Swanson, the hands-to-head portrait of Greta Garbo, and the smiling Charlie Chaplin.

Perfection of realistic techniques

At the outbreak of World War II, Steichen, rejected by the army because of age, was commissioned by the United States Navy to organize a department to photograph the war at sea. Soon after the Japanese attack on Pearl Harbor in 1941, his photographic exhibition entitled "Road to Victory" was presented in New York's Museum of Modern Art, followed three years later by a similar exhibit, "Power in the Pacific"; both proved to be powerful unifying forces for a nation at war.

After the war, from 1947 to 1962, Steichen was director of photography at the Museum of Modern Art in New York City. In 1955 he organized one of the most popular exhibitions of photographs ever presented, "The Family of Man." It was based on the concept of human solidarity that is the motif of his brother-in-law Carl Sandburg's biography of Abraham Lincoln. Together with his assistant Wayne Miller, Steichen selected 503 from the more than 2,000,000 prints submitted from all over the world. After its opening at the Museum of Modern Art in 1955, the exhibition toured most of the major museums in the world and was seen by more than 9,000,000 people.

"The Family of Man"

Steichen was 76 when "The Family of Man" first appeared, and in 1960 at the age of 80 he married for the third time—to Joanna Taub who was then in her late 20s. To him, "boredom and disinterest and lack of awareness are what characterize a superannuated mind, no matter what the chronological age may be. Happily, growth is not something for children only." He died March 25, 1973 in West Redding, Connecticut.

Early in his career, Steichen rejected the idea that there was one system of aesthetics in American photography. Such a notion, he felt, served only to confine and inhibit the creative photographer. As a consequence, he was always eager to serve as a promoter for American photography, arranging for exhibits, purchasing photographs, and generally acting as patron of younger photographers, regardless of the style of their work. According to Steichen, it is this variety—based on the unique vision of the individual photographer—that has made photography the vigorous art it is today.

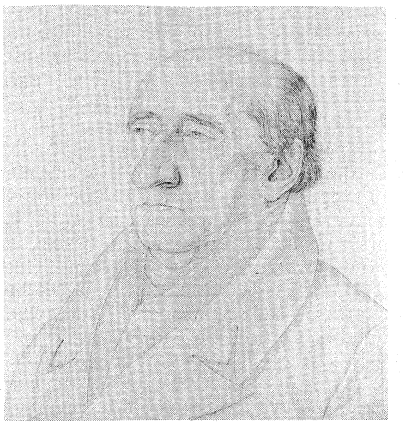
BIBLIOGRAPHY. EDWARD STEICHEN, *A Life in Photography* (1963), with biographical data by GRACE MAYER; PETER POLLACK, *Picture History of Photography*, rev. ed. (1970); BEAUMONT NEWHALL, *The History of Photography, from 1839 to the Present Day*, rev. ed. (1964), and ed. with NANCY NEWHALL, *Masters of Photography* (1958); *The Family of Man* (1955), a book from the exhibition of the same title for the Museum of Modern Art; *Camera Work*, a magazine of photography, with a Steichen number and special supplement (April 30–May 27, 1906), and a double number devoted to Steichen (April-July 1913). Steichen's photographs may be seen in issues of *Vogue* and *Vanity Fair* (1923–28).

(P.J.P.)

Stein, Karl, Freiherr vom

Heinrich Friedrich Karl, Freiherr vom und zum Stein, Prussian civil servant and minister, one of Prussia's greatest 19th-century statesmen, was the author of epochal reform legislation that liberated the peasants from hereditary servitude and granted self-government to the municipalities. As personal adviser to Tsar Alexander I, he urged that Russia continue the war against Napoleon (1813–14) until the French were defeated. A dedicated educator of his nation, he hoped that through active participation in local self-government the people would acquire a sense of political responsibility in place of their former submissiveness to the sovereign. His later years were occupied in promoting historical research in Germany.

Childhood and youth. Stein was born on October 26, 1757, in Nassau, Germany, into a family of the imperial nobility. His father, although a Protestant, was chamberlain to the Catholic elector and archbishop of Mainz. When not at court, the elder Stein was hunting or taking care of his estate. His mother, also a member of an old aristocratic family, was described by such a competent judge of character as Goethe as a woman of impressive personality: well-informed and original in her thinking, practical, pious, and cultured. Not only did she attend to her household and her large number of children, but she



Freiherr vom Stein, portrait by Friedrich Olivier, 1820.
By courtesy of the Stadelsches Kunstinstitut, Frankfurt/M.

was also deeply interested in the educational problems of the Enlightenment and had close contacts with the literary circles of her time.

In his autobiography (1823) Stein recounts that he grew up "under the influence of very worthy and devout parents" and that their example imbued his youthful mind with the ideals of piety, patriotism, family and professional honour, and the duty of devoting one's energies to the community's needs and of acquiring the necessary proficiency for such purposes by diligence and effort.

His parents wished him to become a judge at one of the imperial courts of the old German empire, and in 1773 he was accordingly sent to study law at the University of Göttingen, accompanied by a tutor, as was then customary. Stein studied not only law but statistics, economics, and history as well. In contrast to many of his peers "he tackled his work with truly ascetic diligence." He not only attended and reviewed lectures conscientiously, but also took private lessons in English and Italian, in mathematics and music, and in riding, dancing, and fencing. Reading widely and systematically, he was particularly interested in public law, German and English history, English literature, and economics. Austere moral principles guided his personal life and his evaluation of men and situations. He budgeted his time, avoided idleness and empty amusements, selected his friends "according to rigorous standards," seeking them only among "the better, more noble and able young people and avoiding superficial, vain, idle, and empty contemporaries."

Influence of August Rehberg

August Wilhelm Rehberg, whom he met in Göttingen, became such a friend and exercised a greater influence on Stein than any of his academic teachers. Highly cultured, Rehberg was a political thinker and writer of note, who advocated a liberal-conservative policy to preserve the old where it had proved itself and to make reforms where conditions demanded them. In his view, the state was founded in tradition—not on the "social contract" of the Enlightenment theoreticians of natural law—and in 1790 he became one of the most enthusiastic champions of the *Reflections on the Revolution in France* of the English statesman and philosopher Edmund Burke. It was in the constant exchange of ideas with Rehberg that Stein developed his own reform ideas in the quarter century between 1775 and 1800. He attempted to find a middle way between revolution and absolutism that could unite tradition and progress.

In 1777 Stein left the university and used the next three years to study the legal procedures of the imperial supreme courts in Wetzlar and in Vienna and to become familiar with the chief German principalities. In the winter of 1777–78 he stayed at the elector's court in Mainz; in the following summer he visited the capitals of some of the South German princes and attended the imperial diet at Regensburg. He visited Vienna and Austria during 1779, and in January 1780 he was presented to Frederick the Great in Berlin. In the course of his travels he decided against joining the imperial service and to enter the Prus-

sian civil administration instead. In 1780 through his friendship with Friedrich Anton von Heinitz, the Prussian minister of mines, he obtained a suitable post.

Career as Prussian civil servant. Stein began his career in the department of mines and factories, at first stationed for some years in Berlin, then for a long period in Westphalia. Travelling widely in the course of his work, he came to know all the Prussian territories as well as Saxony, Poland, Holland, and England. He was offered the ambassadorships at The Hague and at St. Petersburg, but life in the great world did not attract him, and he rightly regarded himself as a poor diplomat "because of my tendency to be independent, outspoken, and easily irritated." In his work as an administrator, on the other hand, he was content and highly successful: he improved the road network, made the rivers navigable, promoted textile production, and reformed the system of tax collection. When the ecclesiastical principalities were secularized after 1803, Stein was the chief official responsible in Westphalia for merging them into the Prussian state. In keeping with his political principles, he went about his task with such tact and respect that all those concerned were warmly grateful to him. Because of his many years of duty there, Westphalia became his second home.

Marriage and family life

In 1793 he married the countess Wilhelmine Wallmoden. She was the daughter of a Hanoverian general and a granddaughter of England's King George II, through one of his mistresses, the Countess of Yarmouth. On his engagement to Wilhelmine, Stein wrote enthusiastically:

> I should indeed be obtuse and insensitive if I failed to recognize the worth of a girl whose love is so pure . . . I hope that my harshness, hot temper, and impatience will be softened by the sight of this gentle creature and her good sense.

Nonetheless, during the first years of their marriage, Stein felt that his wife did not appreciate his way of life and his goals. As the years passed, however, he became increasingly respectful of his wife's character and abilities. Together they devoted themselves to the education of their two daughters. After the collapse of Prussia in 1806–07 and until the final defeat of Napoleon in 1814–15, Stein was separated from his family for long periods. During these difficult times, his wife took care of his estates, executed his commissions, and brought up the children. In her husband's words, she showed "serenity and nobility of mind." He outlived her by 12 years.

Achievements as minister and prime minister. In 1804 King Frederick William III of Prussia appointed Stein Prussian minister of economics and finance. At about that time he lost the sovereign rights over his hereditary lands in western Germany between the Rhine and the Lahn rivers as a result of the dissolution of the old German empire; his neighbour, the Duke of Nassau, acquired political sovereignty over them. In an angry letter to the Duke, Stein vainly protested this high-handed action.

Once installed as minister in Berlin, Stein gained an insight into the entire Prussian administrative system and realized that it stood in need of reform. His intention to modify the King's absolutist government by having him delegate clearly defined responsibilities to his ministers brought him into conflict with his monarch. Enraged by his "disrespectful, improper behaviour," Frederick William III dismissed Stein in January 1807 as an "intractable, defiant, obstinate, and disobedient servant of the state." Undoubtedly Stein possessed all these traits but viewed positively they spelled energy, civil courage, strength of character, and loyalty to principles. At the time of this clash, however, the King was fleeing from Berlin to East Prussia. War had broken out between France and Prussia, Napoleon had won the first battles, conquered Berlin, taken most of the fortifications and occupied the greater part of the kingdom. Just before the French arrived, Stein, with great presence of mind, had saved the considerable state funds for which he was responsible, by shipping them on river boats via Stettin to Königsberg, in East Prussia.

After his dismissal, Stein retired to his ancestral castle in Nassau in March 1807. Now aged 50, he occupied his enforced leisure by setting down his practical experiences and his political reflections in what was to become known as the famous Nassau Memorandum on the reform of the Prussian state. The program he envisaged was not to remain mere theory. Napoleon compelled the King of Prussia to dismiss his minister Karl von Hardenberg in the summer of 1807, and no successor seemed to be as suitable as Stein. Accordingly, he was appointed first minister in October, given wide powers, and entrusted with the conduct of all the affairs of the state. In this capacity he undertook the reform of the Prussian state. He had excellent men at his side as senior officials inspired by the ideas of the German philosopher Immanuel Kant, the Scottish economist Adam Smith, and the French Revolution. The King, too, was won over.

Nassau Memorandum on reform of the Prussian state

During his 14 months in office, Stein's achievements were outstanding. He emancipated the peasants from hereditary servitude and issued laws permitting both the unrestricted sale of land and the free choice of occupation—acts that represented a decisive step toward the goal of legal and personal equality for all citizens of the Prussian state. Stein's municipal ordinance granted the towns constitutions that enabled them to administer themselves. Beyond that he instituted reforms on all levels of administrative government. Suddenly, in the midst of these labours, one of Stein's letters containing imprudent remarks about Napoleon was intercepted by the French censor, and, yielding to French pressure, the King dismissed Stein.

Last years. Outlawed by Napoleon, Stein found refuge in the Austrian Empire, where his family was able to join him. He spent much of his time in the study of history and wrote a number of historical essays. He remained in close touch with his political friends in Prussia and Germany and as an ardent patriot was continuously concerned with means of freeing Germany from the foreign yoke.

Political adviser to the Tsar

In May 1812 Stein accepted Tsar Alexander I's offer to become his personal counsellor. He joined the Tsar in Vilna and thenceforth accompanied him first to Moscow and St. Petersburg and later, as the Russian army pursued Napoleon, to East Prussia, Germany, and France, and to the peace negotiations in Paris and Vienna. As Russian plenipotentiary he organized the anti-French uprising of the province of East Prussia (February 1813), took part in the Russo-Prussian treaty negotiations at Kalisz (March 1813), was appointed commissioner for the reconquered German territories, and was frequently consulted during the peace negotiations.

In 1815, after retiring from the Tsar's service and from public life, Stein spent most of his time on his estate in Cappenberg in Westphalia. He concentrated his energies on the study of history, his principal private interest. His library, numbering 7,000 volumes, was larger than that of any professional historian. In 1819, in his apartment in Frankfurt, he founded the Gesellschaft für ältere deutsche Geschichtskunde (Society for Earlier German History), which to this day has remained by far the most important organization for the publication of source material on German medieval history. Beyond that Stein hoped that historical research would "nurture our love for our common fatherland." He died at Cappenberg on June 29, 1831.

Almost all his learned contemporaries considered Stein a man of towering stature, whether they admired him or feared his indomitable energy and stormy temper. When, in 1815, the German poet Ernst Moritz Arndt saw Stein and Goethe walk through the Cathedral of Cologne, he called them the two greatest living Germans. Stein's work had a lasting influence on several generations. He introduced liberal and constitutional elements into the absolutist state and, by his example and influence, made participation in public life a moral postulate. He has been regarded as a political mentor of the German people well into the 20th century.

BIBLIOGRAPHY. GERHARD RITTER, *Stein: Eine politische Biographie*, 3rd ed. (1958), a basic work in regard to scientific research and interpretation of sources, but the evaluations are somewhat antiquated and overly nationalistic in character; KURT VON RAUMER, *Freiherr vom Stein,* ed. by

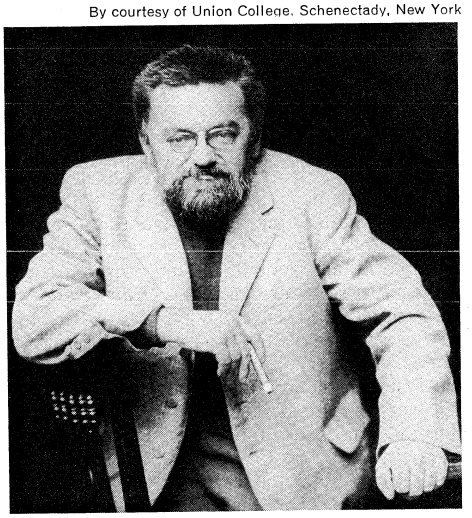
HEINZ GOLLWITZER and RUDOLF VIERHAUS (1961), essays based on expert knowledge that show the manysidedness of Stein's personality and provide an analysis of his political concepts; W.M. SIMON, "Variations in Nationalism During the Great Reform Period in Prussia," *American Historical Review*, 59:305–321 (1954), on Stein's observations on the awakening national consciousness in Germany. For works on Stein's personality, see *Die Autobiographie des Freiherrn vom Stein*, ed. by KURT VON RAUMER, 3rd ed. (1960), containing an important introduction; and ERNST M. ARNDT, *Meine Wanderungen und Wandelungen mit dem Reichsfreiherrn Heinrich Karl Friedrich vom Stein* (1858, ed. by WILHELM STEFFENS, 1957).

(E.W.Z.)

Steinmetz, Charles Proteus

One of the world's foremost electrical engineers, Charles Proteus Steinmetz developed fundamental ideas concerning alternating current systems and devised calculating methods that enabled engineers to design more efficient electrical machinery. His success in combining theory with practice made him immensely valuable to the electrical industry during its early period of growth. He led in transforming electrical engineering from a craft into a profession. Moreover, Steinmetz exemplified the engineer with a social consciousness; his political activity and efforts toward improving education, although not as well remembered as his work in electrical theory, had a widespread impact.

By courtesy of Union College, Schenectady, New York



Steinmetz.

He was born Karl August Rudolf Steinmetz on April 9, 1865, in Breslau, Germany (now Wrocław, Poland). At birth he was afflicted with a physical deformity, hunchback. As a youth he showed an unusual capability in mathematics, physics, and classical literature. On graduating from the *Gymnasium*, with honours, he went to the University of Breslau in 1882. Although a serious and hard-working student, he found time to join a student socialist club, which was banned by the government after becoming affiliated with the German Social Democrats. When some of his fellow party members were arrested, Steinmetz took over the editorship of the party newspaper, "The People's Voice." One of the articles he wrote was considered inflammatory; the police began a crackdown on the paper, and Steinmetz had to flee Breslau (1888). After a short stay in Zürich, Switzerland, he immigrated to the United States in 1889, travelling by steerage, the cheapest class at that time aboard ship. He carried with him a letter of introduction to Rudolf Eickemeyer, of Eickemeyer and Osterheld Manufacturing Company in New York, from the German editor of an electrical journal, in which he had published theoretical articles on the operation of a transformer (though until employed by Eickemeyer and Osterheld, he had never seen such a device). He soon joined the American Institute of Electrical Engineers and the New York (now American) Mathematical Society. At about the same time, Steinmetz Americanized his first name to Charles and substituted Proteus, a university nickname, for his two middle names.

Early socialist views

Under the tutelage of his employer, he became increasingly absorbed in the practical aspects of electrical engineering, and, in so doing, moved away from his theoretical interests in the mathematics of electricity. He established a small laboratory at the factory, where he did much of his scientific research. Eickemeyer, who, like Steinmetz, had been a social revolutionary in Germany, had more than 150 patents and was an important pioneer in the electrical industry. One of his inventions was a magnetic bridge, a device that Steinmetz used in his experiments on power losses in the magnetic materials used in electrical machinery and that led to his first important work, the law of hysteresis, dealing with the power loss that occurs in all electrical devices when magnetic action is converted to unusable heat. Until that time the performance of an electrically powered machine was known only after it was built. Once Steinmetz had found the law governing hysteresis loss, engineers could calculate and minimize these losses in the design of machines.

His classic work on hysteresis

In 1892 Steinmetz gave two papers before the American Institute of Electrical Engineers on this new law concerning hysteresis loss. His work was immediately recognized as a classic by the few who understood it, and the constant he calculated for this loss has remained a part of electrical engineering vocabulary. Thus, Steinmetz' reputation was assured at the age of 27.

His second contribution was a practical method for making calculations concerning alternating current circuits. This method was an example of using mathematical aids for engineering the design of machinery and power lines, so that the performance of the electrical system could be predicted in advance without the necessity of going through the expensive and uncertain process of building the system first and then testing it for its efficiency. Steinmetz' method of calculation was presented to an uncomprehending audience at the International Electrical Congress in 1893. His book *Theory and Calculation of Alternating Current Phenomena* in 1897 was read and understood by only a very few. The problem that Steinmetz faced was that electrical engineers were not taught enough mathematics to understand his new mathematical treatment of problems. To educate the electrical engineering profession, he published a book, *Engineering Mathematics*. Gradually, through his writing, lecturing, and teaching, his method of calculation was universally adopted.

In 1893 the newly formed General Electric Company purchased Eickemeyer's company, primarily for his patents, but Steinmetz was considered one of its major assets. At General Electric, Steinmetz gained an expanded opportunity for research and implementation of his ideas. He was assigned to the new calculating department, the first job of which was to work on the company's proposal for building the generators at the new Niagara Falls power station. In 1894 the General Electric Company transferred its operations to Schenectady, New York, and Steinmetz was made head of the calculating department. He at once began to indoctrinate the engineers with his method of calculating alternating current circuits and helped to establish an extension course.

He shared a house with other bachelors in Schenectady, where it was possible for him to indulge his interest in plants, such as orchids, and a menagerie of animals. One of Steinmetz' closest associates was Joseph LeRoy Hayden, a young engineer at General Electric, whom he legally adopted.

A conspicuous figure because of his deformity, his beard, and his ever-present cigar, Steinmetz won his fellow workers' esteem with his frankness and modesty. Despite his position as an executive of a giant business corporation, he expressed his political and social consciousness through his membership in the American Socialist Party. In 1911, when a Socialist was elected mayor of Schenectady, Steinmetz was appointed to the Board of Education, which then elected him president. Steinmetz ran for president of the Common Council of the city and

was elected in 1915 on the Socialist ticket. He was then successful in gaining funds for the school system. He established classes for subnormal, tubercular, and anemic children and for children with special difficulties.

Public life

Steinmetz ran on the Socialist and Farmer-Labor ticket in New York state in 1922 for the position of state engineer and surveyor. In his campaign he proposed a plan whereby the 9,000,000-horsepower potential at Niagara Falls be realized by diverting the water from the Falls. The plan would bring the state $2,000,000,000 per year. The Falls were to be released on Sundays and holidays for sightseers. Steinmetz promised a water power survey if he were elected. He criticized the use of 54,000,000 tons of coal burned in New York state per year and wanted to substitute water power in order to bring the state into the electrical age.

In his book *America and the New Epoch* (1916) Steinmetz gave a sweeping and incisive economic analysis of history from the Roman Empire to the present era. He saw the operation of natural law in the march of economic growth from individualism to cooperation. Competition, he wrote, should not be advanced since it was as out-of-date as feudalism. The trust-busting era was the result of public antagonism to self-centred and self-satisfied industrial corporations that had failed to represent accurately their social usefulness to the public. Steinmetz saw individualism as a remnant of the past and cooperation the wave of the future. In the midst of World War I, Steinmetz wrote that America must work out its own destiny as a cooperative democratic commonwealth, after ending its racial, industrial, and political strife. The end result he favoured would be an industrial reorganization by the federal government and the emergence of cooperative industrial corporations.

In a talk on "The Future of Electricity" in 1910, Steinmetz stated his belief that the electrical engineer should promote the use of electricity by consumers in order to reduce its cost. Eventually, he thought, the cost might be reduced to one or two cents per kilowatt-hour from the high (at that time) of 20 cents per kilowatt-hour. The era of coal would end, he declared, after the supply of coal was exhausted or sooner, if the American people awakened to the viciousness of poisoning the air by burning coal. He also argued that city sewage should be carried to the outlying countryside in order to restore the fertility of the soil and stop stream pollution. The electrical production of nitrogen fertilizers, he thought, was the only way of saving the farmlands.

Steinmetz' third major scientific achievement was in the study and theory of electrical transients—that is, changes in electrical circuits of very short duration. A prime example of this phenomenon, lightning, was interesting to him as a mathematical problem; his work included the development of protective devices for transmission lines. He designed a generator that produced a discharge of 10,000 amperes and over 100,000 volts, equivalent to a power of over 1,000,000 horsepower for one-one hundred thousandth of a second. The purpose of this project, on a smaller scale than an actual stroke of lightning, was to observe and measure its destructive effects in nature. This was his last major project at the General Electric Company, where he had become head of the engineering consulting department.

Steinmetz' honours included an honorary doctorate from Union College, Schenectady, where he had taught for over ten years, the presidency of the American Institute of Electrical Engineers (1901–02), the award of the Elliott Cresson gold medal by the Franklin Institute, and memberships in the American Academy of Arts and Sciences and the American Philosophical Society. He died in Schenectady on October 26, 1923.

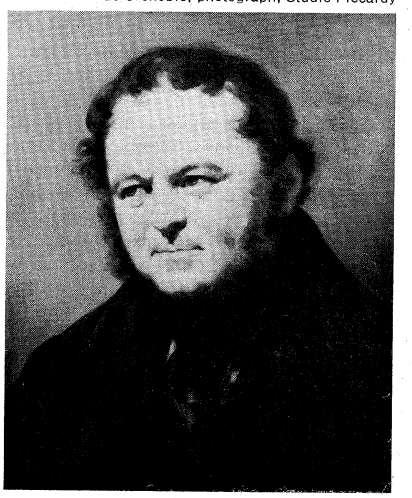
BIBLIOGRAPHY. JOHN WINTHROP HAMMOND, *Charles Proteus Steinmetz: A Biography* (1924), although not definitive, is the best available. See also ERNEST CALDECOTT and PHILIP L. ALGER (eds.), *Steinmetz the Philosopher* (1965), a compilation of Steinmetz' works with brief comments; SIGMUND A. LAVINE, *Steinmetz: Maker of Lightning* (1955); and JONATHAN N. LEONARD, *Loki: The Life of Charles Proteus Steinmetz* (1929), a romanticized account. Books by Steinmetz include: *America and the New Epoch* (1916); *Engineering Mathematics: A Series of Lectures Delivered at Union College*, 3rd ed. (1917); *Four Lectures on Relativity and Space* (1923); *The Future of Electricity* (1910); and *General Lectures on Electrical Engineering*, 5th ed., comp. and ed. by JOSEPH L. HAYDEN (1918).

(H.I.S.)

Stendhal

In describing the various aspects of the novelist Stendhal's character and vocation, his biographers have constantly used the word "failure": failure as a lover, failure in his military career, even failure in his vocation as a writer. Yet today almost every critic would acknowledge that he is one of the leading French writers of the 19th century, along with Balzac and Flaubert. Even more than they, he seems to speak directly to a young generation of readers, especially in France, Britain, and the United States. His writing reflects, to an unusual degree, his personality and thought, which combined countless paradoxical conflicts and desires. He claimed to write for the "happy few" who are not bound by convention, who do not find their happiness in subservience, who follow the lead of their senses and their instincts. Each hero in Stendhal's novels is presented as a man opposed to the social world in which he lives and serves. In his pursuit of happiness, Stendhal regarded the world as an enemy that must be fought in the full knowledge that it was such.

By courtesy of the Bibliotheque Municipale de Grenoble; photograph, Studio Piccardy



Stendhal, oil painting by Pierre-Joseph Dedreux-Dorcy (1789–1874). In the Bibliothèque Municipale de Grenoble.

Early life. "Stendhal" was the chief pseudonym of Marie-Henri Beyle, who was born in Grenoble on January 23, 1783. His father, Chérubin Beyle, was a barrister in the Grenoble parliament, and his mother, Henriette (*née* Gagnon), was the daughter of an eminent Grenoble physician. (He liked to think that, as a Gagnon, he was descended from an Italian family who had moved to Avignon with the papacy during the 14th century.)

The boy worshipped his mother, who died when he was seven; for a long time afterward, he continued to speak and write of her beauty, sensitivity, and intelligence. After his mother's death, the boy quickly grew to hate his father's ideas, manners, and love of money. He spent as much time as possible with his maternal grandfather, Henri Gagnon, whom he loved and respected and called his "real father" and who encouraged his interest in literature and the arts. At home, under the oppressive treatment (which he doubtless exaggerated) of his father, of Séraphie, his mother's sister, and of the Abbé Raillane, a private tutor, Beyle developed a nature individualistic, disrespectful, and impulsive. These three "tyrants" represented for him the bourgeoisie of Grenoble that he came to loathe. Almost until adolescence he was overly protected and isolated from most normal encounters. Extrovert activity was replaced by daydreaming and a taste for meditation that lasted throughout his life.

In his autobiography, *Vie de Henri Brulard*, Stendhal later explained his reaction against his family and education as *espagnolisme*, which he defines as a feeling of disgust for what is commonplace, a scorn for money and commerce, and a strong attraction for a willed forcefulness in man, for passionate ebullience. At the age of 17 he went to Paris, ostensibly to take the examination to enter the prestigious École Polytechnique but in reality to escape the suffocation of Grenoble. His knowledge of the world was limited to what he had read in novels (favourites included the Abbé Prevost's *Manon Lescaut* and Jean-Jacques Rousseau's *La Nouvelle Héloïse*). After only a few weeks' miserable existence, he fell ill with a stomach complaint, but a cousin of Dr. Gagnon, Noël Daru, gave the youth a room in his house. He was a stern gentleman and never fully approved of the young provincial, but he did find him a modest secretarial position with the war office.

Move to Paris

Professional career and writings. Three months later, in May 1800, Henri Beyle crossed the Alps to join the reserve army of Napoleon, who was fighting his second Italian campaign, and reached Milan the day after Napoleon's decisive victory at Marengo. Milan was a place of enchantment for him. He wrote of the magnificent music at La Scala opera house and praised the beauty of several Milanese ladies, especially that of Angela Pietragrua, who had been an intimate friend of the French painter Antoine-Jean Gros. Beyle fell in love with her. (Eleven years later he became her lover. His first sexual encounter, which took place soon after his arrival in Milan, left him with a venereal disease that he had to treat intermittently for the rest of his life.) As Beyle languished over his love for Angela—while at the same time attending performances at La Scala and gatherings at the casino—he developed a sensitivity to amorous excitement that was to be characteristic of his life. Now, at the beginning of his military career, he was also concerned with Italian lessons, fencing lessons, horseback riding, plays and operas, and infatuations with various Milanese ladies. He spent 18 months of campaign fighting without ever really being a soldier. By the end of 1801, he had made up his mind to resign from the army and return to Paris.

Literature was by this time, at the age of 20, his principal preoccupation, and drama was the aspect of literature that attracted him most. In Paris he announced that he intended to become Molière's successor. But to study he needed books, and to appear in society he needed clothes and spending money. He was often depressed by his Grenoble accent, his far from handsome physical appearance, his sense of inferiority, and his unsuccessful career as a seducer, as well as his inability to write verse drama. And yet the painful experiences he endured in Paris all helped to transform the bourgeois Henri Beyle from Grenoble into Stendhal the novelist.

He became infatuated with an actress, Mélanie Guilbert (called Louason), and in 1805 followed her to Marseille, where for a few months he became a wholesale grocer. He then returned to the war office, was present at Napoleon's entrance into Berlin in October 1806, and was stationed for two years in Brunswick (German, Braunschweig), Germany. There he continued his reading, begun earlier, of 18th-century philosophers—in particular of Destutt de Tracy and Claude-Adrien Helvétius—and cultivated a taste for Mozart. By 1809 he had returned to the active army and, as a member of the quartermaster's staff, took part in Napoleon's Russian campaign of 1812. He saw not the battles but the aftermath of the battles, the scenes of horror of the battlefield, which he described in some of his most brilliant letters. After the fall of the empire, seeing no prospect of employment under the new regime, he departed for Italy.

Military service

Milan became more than ever his favourite city. His first book, written under the pseudonym L.-A.-C. Bombet, *Vies de Haydn, de Mozart et de Métastase* (1814), raised the question of what kind of writer he should be: musicologist, art critic, or literary critic. Beyle was to use some 170 pseudonyms, but it was for the publication of *Rome, Naples et Florence en 1817* (1817) that he first used the name "M. de Stendhal," which he borrowed from the name of a small Prussian town, birthplace of J.J. Winckelmann, a famous German art critic.

When he returned to Paris, in 1821, he was 38 years old. He led there a limited social life and during the next nine years published six books: *De l'amour* (1822), "Racine et Shakespeare" (two pamphlets, 1823 and 1825), *Vie de Rossini* (1823), *Armance* (his first novel, 1827), *Promenades dans Rome* (1829), and *Le Rouge et le noir* (1830), which was to become his most famous novel.

After the July Revolution of 1830, he was named French consul at Trieste but then was sent, in April 1831, to Civitavecchia, close to Rome. He was bored by the pettiness connected with his profession and by the dullness of Civitavecchia. Nevertheless, there was plenty of opportunity to write, and during the last ten years of his life he gradually learned to love Rome as well as Milan.

For so long as he was a civil servant, Stendhal refrained from publishing. But he wrote extensively: his *Souvenirs d'égotisme* (1892), the novel *Lucien Leuwen* (1894), which he never completed, and his autobiography, *Vie de Henri Brulard* (1890), which he took only as far as his adolescence. All aspects of Rome fascinated him: the social habits, the arts, Roman chronicles and memoirs. But Stendhal had little interest in the Catholic Church and no interest in spiritual problems or mysticism. For him the Roman Catholic Church was simply an extraordinary empire, an unusually competent political system. As Stendhal grew older, he maintained a certain degree of curiosity about life, even if the desire to live diminished markedly. He dreamed of love even more than in his earlier years and indulged in memories of affairs that had been almost successful. The ghosts of innumerable feminine figures obsessed him at the time when, in the space of 52 days, he wrote *La Chartreuse de Parme* (1839) on one of his frequent visits to Paris because of poor health.

La Chartreuse de Parme

Stendhal pursued the image of happiness throughout his life, without ever trying to understand why he failed the reality. Toward the end of his life he was consoled by the conviction that posterity would honour him. Just before his death, he had the joy of reading a glowing article, written by Honoré de Balzac, on *La Chartreuse de Parme*. This praise was the first indication of what "the happy few" would eventually discover in the novels of Stendhal.

Stendhal suffered an attack of—presumably—apoplexy in March 1841. Afterward, he suffered intermittently from aphasia, an illness that resulted in spells during which he lost the power of speech. Another official leave of absence was granted, and he set out for France on October 11, 1841. Five months later Stendhal collapsed under a massive stroke and died on March 23 in his hotel room at the age of 59.

His cousin Romain Colomb and his sister Pauline were in Paris when he died. A monument with the inscription that he himself had written was placed at the grave in the cemetery at Montmartre: *Arrigo Beyle, Milanese*. Only three people, including the novelist and historian Prosper Mérimée, followed his coffin to its grave.

Emotional life. The circumstances of Stendhal's death might easily be interpreted as symbolic of the life story of a man who cherished the traditional recipes for happiness in the family and society but who never followed them—or was unable to follow them—himself. He never had a permanent address or a permanent profession; he had no home, no children, and, one might almost say, no mistress; after 1808 and the marriage of his sister Pauline (who was his confidante and to whom he wrote many letters from Marseille and from Germany), he literally had no family. Yet his nature craved intimacy. More than most men, he strove to maintain friendships. But according to all the available documents it seems that his friends showed little real sympathy for him and had no comprehension of his nature.

The fundamental life of Henri Beyle—the biography of his thoughts and fantasies and sentiments—was, therefore, lived in the novels he signed with the name "Stendhal." These books narrate basically the same story of the same young man. They are Henri Beyle's fantasies, lived out in a variety of settings and telling the story of a young

Stendhal's heroes

man who refuses to grow into adulthood. In each book a young man's charm affects the people around him: either it attracts them or else it breeds envy in them. Even though his heroes (Julien in *Le Rouge et le noir*, for example; Fabrice in *La Chartreuse de Parme*; or Octave in *Armance*) appear in different societies and become members of different communities within them, their desires, their sensitivities, and their needs are identical. They are what Stendhal himself was in his imagination, and so the novels are autobiographies in retrospect: in them he relived his own emotional life.

"Beylisme." Stendhal was so struck by his strange compulsion to write these novels, by the pleasure and the therapy they provided, that he had to invent a name for this mystery. He used his own name and not his pseudonym: *Beylisme*. The word designates both a method of writing fiction and a search for happiness.

The disciple of *Beylisme* is one of the "happy few" described above. There was nothing mystical about happiness, in Stendhal's conception of it. For him, the pursuit of happiness was comparable to a mathematical demonstration: its logic lies in following the dictates of egoism, self-interest, which Stendhal saw as the sole motivation of all man's actions. Its attainment in a complex world—which witnessed the fall of Napoleon, the restoration of the Bourbons, and the Revolution of 1830—meant calculating moves, estimating chances, practicing flattery. In many of his letters to Pauline, Beyle acknowledged a great change in himself as he began to understand the reasons for some of his defects in life and saw the ruses he would have to adopt and cultivate in order to overcome them. The Stendhal of the 1820s was very different from the young man who had spent most of his time during the early years of the century alone in his room. In *De l'amour* he analyzed his sentiments not only about love but also about himself as he was in 1821 and 1822. By that time he had begun to resemble the kind of polished gentleman he was later to portray in M. Leuwen (*Lucien Leuwen*) and Count Mosca (*La Chartreuse de Parme*). After 1830 and the publication of *Le Rouge et le noir*, he was still searching for happiness and a "method" that would permit this happiness to be attained. The philosophy from which he derived a method was that of the Idéologues, a group of 18th-century thinkers who believed in the perfectibility of the human race and whose doctrines were explained by Destutt de Tracy. *Beylisme* makes out of man's actions a fully conscious means of enjoying life to the full, materialistically and sensually.

Stendhal's epicureanism

The ways in which Stendhal trained himself to enjoy the arts, the beauty of painting and of music, reveal the epicurean side of his nature. His dreams of an existence with enough income to assure himself of the usual comforts and more—a carriage, a box at the opera, a house where he could adequately receive his friends—were the dreams of an epicurean. But the novelist's heroes were not so socially and aesthetically limited in their aims. Julien committed a crime in the name of passion, and Fabrice had in him a boldness and spirit that were not the traits of gentlemen in society. The contradictions between the epicureanism of Stendhal's own nature and the bold heroism of his fictional heroes are fused and harmonized in *Beylisme*.

Emotionally, Stendhal separated himself from his family, and intellectually he separated himself from the bourgeoisie. What Baudelaire was to call *dandysme* corresponded closely to Stendhal's attitude. It was a form of isolation on many levels: artistic, social, intellectual. The perpetual wanderings of Stendhal—from country to country, from one hotel to another—and his perpetual use of pseudonyms testify to a quest or a pilgrimage of a psychological if not a mystical nature. Subconsciously, the happiness he sought from life, which always eluded him, he found in the writing of his books and in particular of his two masterpieces, *Le Rouge et le noir* and *La Chartreuse de Parme*. These are not only novels but treatises on the ways of the world and mythical projections of the way in which Henri Beyle had learned to live, both materialistically and imaginatively, with the drama of his many selves.

MAJOR WORKS

NOVELS: *Armance, ou quelques scènes d'un salon de Paris en 1827* (1827; trans. by C.K. Scott-Moncrieff, 1928); *Le Rouge et le noir* (1830; *The Red and the Black*, 1926; *Scarlet and Black*, trans. by C.K. Scott-Moncrieff, 1927); *La Chartreuse de Parme* (1839; *The Charterhouse of Parma*, trans. by C.K. Scott-Moncrieff, 1925); *Lamiel* (unfinished, published posthumously 1889; trans. by T.W. Earp, 1952); *Lucien Leuwen* (unfinished; first published in 1855 in *Nouvelles inédites* under the title *Le Chasseur vert;* published posthumously 1894).

SHORT STORIES: *Chroniques italiennes* (1839; enlarged edition, *Nouvelles inédites*, 1855; *The Abbess of Castro, and Other Tales*, trans. by C.K. Scott-Moncrieff, 1926); *Romans et nouvelles* (1854).

AUTOBIOGRAPHY: *Vie de Henri Brulard* (1890; trans. by C.A. Phillips, 1925).

OTHER WORKS: *Vies de Haydn, de Mozart et de Métastase* (1814); *Histoire de la peinture en Italie* (1817); *Rome, Naples et Florence en 1817* (1817); *De l'amour* (1822; *On Love*, trans. by V. Holland, 1928); "Racine et Shakespeare" (two pamphlets, 1823 and 1825; published in book form as *Racine et Shakespeare, études sur le romantisme*, 1854); *Vie de Rossini* (1823); *Promenades dans Rome* (1829); *Mémoires d'un touriste* (1838); *Voyage dans le midi de la France* (1854); *Souvenirs d'égotisme* (1892; *Memoirs of an Egotist*, trans. by T.W. Earp, 1949); *Journal*, 5 vol. (1923–34; originally published in 1888 as *Journal de Stendhal*); *Vie de Napoléon* (1876, fragments; 1929, complete edition).

BIBLIOGRAPHY. V. DEL LITTO, *Bibliographie stendhalienne: 1944–46* (1948), is the fullest bibliography to date, although since 1948 there have been many studies and monographs and articles of value. The principal centre of research is the Musée Stendhal, Grenoble, which holds Stendahl's manuscripts and papers. Among individual volumes of biography and criticism, the following are useful: VICTOR BROMBERT (ed.), *Stendhal: A Collection of Critical Essays* (1962), containing studies by Erich Auerbach, Jean Prévost, Judd Hubert, Léon Blum, Jean Starobinski, Jean-Pierre Richard, and Simone de Beauvoir that are among the most penetrating essays on the work of the novelist; WALLACE FOWLIE, *Stendhal* (1969), the most recent of the introductions to the life and writings of Stendhal—the bibliography of works in English and French completes to some extent the del Litto bibliography; MATTHEW JOSEPHSON, *Stendhal: Or the Pursuit of Happiness* (1946), a moving study of the life of Stendhal and the way in which the novels reflect the life; HARRY LEVIN, *The Gates of Horn: A Study of Five French Realists* (1963), an exemplary critical estimate (section 3) that stresses *Beylisme;* and MARTIN TURNELL, *The Novel in France* (1951), a study of six aspects of Stendhal, who seems to be for this English critic the best of the French novelists.

(W.F.)

Stereotyped Response

A stereotyped response may be defined as an unlearned behavioral reaction of an organism to some environmental stimulus. It is an adaptive mechanism and may be expressed in a variety of ways. All living organisms exhibit one or more types of stereotyped response.

The capacity for unlearned behaviour is genetically determined in much the same sense as are the position, size, shape, and function of organs. Like structural features, stereotyped responses are the result of a continuing process of evolutionary modification and refinement. Those actions that most successfully aid the plant or animal in its basic drives (*e.g.*, reproduction, search for nourishment, escape from predators) are the ones most likely to be retained in succeeding generations. As environmental conditions change, inherently determined responses also become modified in order to ensure continuation of the species. Because of the fundamental differences that widely separate multicellular plants from multicellular animals, few other generalizations that apply equally to behaviour in the two groups can be made.

Problems of identification and analysis

Particular stereotyped responses in animals do not readily lend themselves to identification in highly evolved forms because learned behaviour patterns obscure the underlying unlearned behaviour; in addition, stereotyped responses provide the building blocks of instinctive behaviour, the complexity of which may obscure the integral parts (see BEHAVIOUR, ANIMAL; INSTINCT). In plants and lower animals, in which learned behaviour is absent or nil, the analysis of behavioral mechanisms is limited

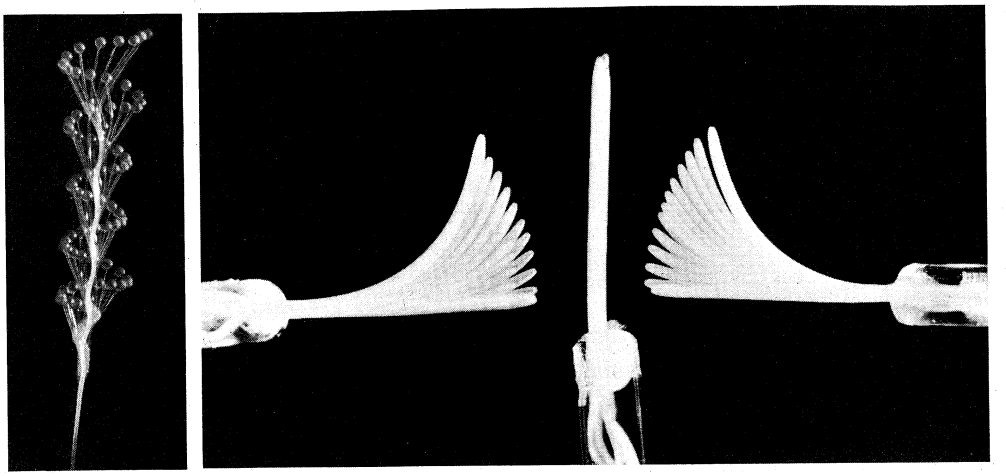


Figure 1: *Tropism.*
(Left) Phototropism shown in the multiple exposure of a fungus fruiting body. Exposures taken at a rate of one every four minutes show the sporangiophore oscillating between two light sources. (Right) Geotropism shown in the multiple exposure of oat coleoptiles. Gravitational redistribution of hormones causes the two seedlings set horizontally to grow upright like the central vertical seedling.
By courtesy of (left) D. Dennison, Dartmouth College, Hanover, N.H., (right) *Life* © (1963) Time Inc.

by the fact that many fundamental cell processes—*e.g.*, the mechanism of plant-growth hormone translocation in geotropism, which is the response to gravity in plants—are not well understood.

Animal behaviour, as a branch of psychology, represents a confluence between the disciplines of ethology and comparative psychology. Most of the pioneer work in stereotyped responses of animals was done by ethologists. During the first half of the 20th century, when much of the groundwork in experimental psychology was laid, ethologists (who were for the most part European) concerned themselves with behaviour in insects, fishes, and birds and were particularly interested in the evolution of instinct. The comparative psychologists during this formative period were mostly Americans. They studied primarily behaviour in common laboratory animals such as guinea pigs, mice, rats, and monkeys, and their interest tended to focus on environmental influences on behaviour as opposed to genetic influences. Since the 1950s, psychologists in general have recognized that both environmental and genetic factors play essential roles in any biological phenomenon. As a consequence of the separate development of ethology and comparative psychology, however, some difficulties have arisen in the use of terminology. The German-American biologist Jacques Loeb applied the term tropism to all oriented movements of organisms, and he proposed that all behaviour is composed of tropisms. Subsequently, to avoid confusion, the terms taxes (singular: taxis) and kineses were introduced by other investigators to refer to animal responses other than those of sedentary, plantlike forms. The terms also have been applied to certain plant movements.

PLANT MOVEMENTS

Although a variety of discrete stereotyped response movements occur in plants, particularly in higher forms such as flowering plants, these autonomous movements usually occur too slowly to permit detection by casual observation. That movements of plants or plant organs actually take place can be strikingly demonstrated by time-lapse photography, in which single photographs are taken at regular intervals as brief as seconds or as long as several days or more. The photographs are then compared or viewed in rapid sequence as a motion picture.

Growth movements

Growth movements, which result from cell enlargement or from an increase in the number of cells, or from both, are of three types: tropic (sometimes called tropistic) movements, nastic movements, and nutations. Tropic movements are responses to stimuli that act with greater intensity from one direction than from another. These include phototropism, response to light; geotropism, response to gravity; chemotropism, response to particular substances; hydrotropism, response to water; thigmotropism, response to mechanical stimulation; traumatotropism, response to wound lesion; and galvanotropism (or electrotropism), response to electrical stimulus.

Autonomous movements in living plants are generally classified as growth movements and turgor movements. So-called hydration movements also occur. These, however, are not necessarily associated with physiological activity and take place in dead as well as in living tissues.

Tropic movements. Most tropic movements in plants are orthotropic; *i.e.*, they are directed toward the source of the stimulus. Many instances are known, however, of movements that are oblique to the direction of stimulus (plagiotropic) or that are at right angles to it (diatropic). The oblique position of many lateral shoots of a growing plant probably results from an equilibrium between negative geotropism (movement away from gravity) and an epinasty (a more rapid growth on the upper side of an organ than on the lower side); the equilibrium is promoted by a chemical factor originating in the terminal bud of the plant. If the bud is removed, at least one of several lateral shoots curves upward. Diatropism may be observed in the growth movement of certain rhizomes (rootlike underground stems), which normally grow horizontally.

Phototropism. Phototropism occurs generally in all organs of the growing plant that are exposed to light, and it is perhaps the most easily observed of the tropic responses. The leaves of vines such as the English ivy (*Hedera helix*), which grow against a flat, vertical surface and thus receive light mostly from one direction, develop in a mosaic-like pattern, with little overlapping of the leaf blades. Similar, though less apparent, leaf mosaics occur in many other plants (*e.g.*, maple) that bear relatively large numbers of leaves. Some herbaceous, or nonwoody, plants (*e.g.*, *Silphium*, *Lactuca*) orient their blade surfaces to face east and west. At midday the flat surfaces are parallel to the direct sunlight. Because they orient in this way, such plants are commonly known as compass plants. The leaves of the mallow *Malva neglecta* turn slowly from sunrise until sunset, so that the broad surface of the leaf is constantly at a right angle to the sun.

The flowers of many plants of the family Compositae also turn with the sun, facing it from sunrise to sunset. Certain grasses (*e.g.*, Bermuda grass) are apparently negatively phototropic at high intensities of light—*i.e.*, the plant surfaces are oriented away from the light source. They lie flat under strong light and stand upright when the light is dim.

Roots normally grow in total darkness; however, when exposed to light, many roots exhibit a negatively phototropic response in that they curve away from the light.

Instances of positively phototropic roots—*i.e.*, with curvature toward the light—are rare.

Growth-rate differences

The orientation of leaves and stems toward a light source is caused by growth-rate differences between different areas of tissue and by differences in the turgor (*i.e.*, rigidity and tension resulting from the pressure of fluid contents) in cells of different tissues. The curvature of a growing leaf or stem toward a source of light is caused by the uneven growth rate between different areas of tissues. Cells on the darker side grow faster than those on the lighter side, resulting in curvature toward the light source. It may be demonstrated experimentally that such differences result from differences in the amount of auxin, a growth-stimulating hormone, present in the cells. The rapidly growing cells contain greater amounts of auxin. It has been shown that in tissue in which illumination is unilateral, auxin migrates from cells on the lighter side to cells on the darker side. It has also been postulated that auxin is destroyed by sunlight. Some investigators have shown that sunlight has an effect on plant growth independent of its influence on the distribution of auxin; the nature of this effect, however, is not yet fully understood.

The English biologist Charles Darwin demonstrated many years ago that, although bending occurs in the basal portions of a grass coleoptile (a protective sheath around the base of the terminal bud), if the tip is shaded with a cap, curvature of the unilaterally illuminated coleoptile may be prevented. Detipped coleoptiles likewise evince little or no phototropic reaction. Bending of the basal portions of the coleoptile is, therefore, subject to a controlling influence of the tip. Subsequent to Darwin's investigation, it was shown that the auxin in the tip of an oat coleoptile migrates laterally across the tip from the lighter to the shadier side. Since downward translocation of the auxin is almost strictly longitudinal, the cells on the shaded side of the coleoptile receive more auxin than those on the lighted side. The cells on the shaded side therefore elongate more than those on the illuminated side, and the coleoptile bends toward the light.

Geotropism. Typically, a plant root grows downward (toward gravity), and the stem grows upward. These are examples of positive and negative geotropism, respectively. In a horizontally oriented oat coleoptile, auxin migrates from the upper to the lower side of the tip and thence longitudinally in the usual manner. The greater concentration of auxin on the lower side than on the upper side induces greater elongation of cells in the former, and the coleoptile bends upward. In a horizontally oriented root tip, auxin also migrates from the upper to the lower side. Elongation of root cells is favoured by much lower concentrations of auxins than those that stimulate elongation of coleoptile cells, and it is retarded at concentrations higher than necessary to stimulate elongation in coleoptile cells. A relatively high auxin concentration on the lower side of the root, therefore, results in a smaller elongation of the cells on the lower as compared with the upper side; the result is a downward bending of the root.

A plant placed horizontally on a horizontally spinning wheel develops no stem or root curvature, for the force that tends to drive the plant outward (centrifugal force) serves to counteract the geotropic response. Primary, or main, roots from which the tops have been removed do not demonstrate a geotropic response. If tops are grafted upon such roots, however, the normal geotropic response is elicited.

Thigmotropism. The growth movement of a plant organ in response to contact with a solid object is known as thigmotropism. Such movements are observable in petioles (leaf stalks), stems, and especially in tendrils, which are modified stems, leaves, or leaflets. Because of inequality in growth rates in different parts of a tendril, it grows in a spiral. If a tendril encounters a solid object, the cells in immediate contact with the object shorten somewhat; those cells on the opposite side elongate. The result is that the tendril bends or creeps close to the object, thus obtaining support for the growing plant. In some species the thigmotropic reaction occurs in less than one minute, suggesting that the speed of the response results from changes in turgor rather than from a growth process. Although contact of a tendril with hard surfaces that are extremely smooth (*e.g.*, drops of mercury) does not elicit a thigmotropic response, the delicate movement of a light thread against a tendril elicits a positive response (see below *Nutation*).

Spiral growth of tendrils

Chemotropism. Unequal growth movements in plants in response to certain substances are known as chemotropisms. Seedlings of the European dodder, *Cuscata europaea,* which is parasitic on a variety of plants, respond to suitable host plants by curving toward them from as far away as ten centimetres (four inches). Cells on the side away from the stimulus grow more rapidly than do those on the near side, thus causing the seedling to "reach" toward the host plant.

Hydrotropism. The response of plant roots to unequal distribution of water is known as hydrotropism, a phenomenon that may be regarded essentially as a specific chemotropic response. Roots do not grow into soil with a moisture content so low as to induce wilting; however, they ordinarily grow into soil with a moisture content high enough to prevent wilting. A more rapid rate of growth of cells on the side of the root away from the moist areas of the soil results in curvature of the root toward the moist area.

Traumatotropism. Damage to plant tissue frequently results in the curvature of the affected organ as a result of unequal growth—a phenomenon known as traumatotropism. Injury to plant tissues probably results in changes in the distribution of auxin or interferes with the movement of food and other substances into growing tissues.

Nastic movements. Nastic movements are those in which the direction of movement is limited by the structural symmetry of the plant organ. In general, nastic movements are evoked by factors such as diffuse light (photonasty) or temperature (thermonasty), which influence the organ equally from all directions. Photonastic movement, which causes an organ to alter its shape or position when taken from the darkness and placed in diffuse light, is thus distinguished from phototropic movement, which occurs when the plant is subjected to unilateral illumination. Nastic movement occurs primarily in leaves, the structure of which largely or entirely restricts movement to specific directions.

Epinasty and hyponasty

The more rapid growth of the upper side of a plant organ is called epinasty; the more rapid growth of its lower side, hyponasty. Opening of flowers is caused by epinastic movements of the sepals or petals; closing of flowers, such as occurs in some species at night or at low temperatures, results from hyponastic movements of the floral parts.

The growing leaves of some plants (*e.g.*, the common jewelweed, *Impatiens*) droop at night, but during the day they are more or less horizontal. The growing leaves of pigweeds (*Amaranthus*) are horizontal during the day but turn upward into a vertical position at night. The movements of both *Impatiens* and *Amaranthus* are caused by differences in the growth rate between the upper and lower sides of the leaves. The movements cease after full growth is reached.

Certain flowers, such as those of *Oxalis* and the evening primrose, *Oenothera,* also demonstrate photonastic movements. Oxalis flowers are open during the day but close at night. Those of *Oenothera* open at night but are closed during the day. The flower petals of some plant species respond by movement when there are changes in temperature. A temperature increase causes the flowers of *Crocus* to open as the result of accelerated growth of the inner petal surfaces. A temperature decrease results in closure of the flower because the growth rate of cells on the outer surface of the petals exceeds that of cells on the inner surface. The daily opening or closing of floral parts and the response of leaves to environmental changes that occur with each daily cycle are sometimes called nyctinasty, or "sleep movements."

Nutation. As the growing tip of any plant organ (*e.g.*, stem, tendril, root, flower stalk, sporangiophores [spore-bearing stalks] of some fungi) elongates, it describes an

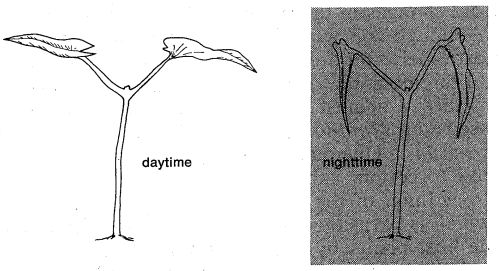


Figure 2: "Sleep movements" in bean seedlings (see text).
From F.A. Brown Jr., J.W. Hastings, and J.D. Palmer, *The Biological Clock* (1970); Academic Press, Inc.

irregular, spiral course. This results from inequality of growth rates in tissues at the periphery of the growing tip and is known as nutation. It occurs most strikingly in twining plants. In these, nutation may be caused by a thigmotropic response and, in part, by the effects of gravity and other factors. Stem tips are ordinarily slender, elongated, and lack leaves. The tip usually droops into a more or less horizontal position. More rapid growth usually occurs on the outermost and lower sides of the stem, resulting in an upward swinging motion of the growing tip. This extension results in a twisting of the stem as a consequence of the unequal rates of growth; thus, different sides of the stem are successively positioned on the lower side as the spiralling growth continues.

Turgor movements. Turgor movements may be initiated by a variety of stimuli—electrical shock, insufficient water supply, jarring physical contact, injury, exposure to certain gases, and changes from light to darkness or from darkness to light. Many turgor movements are caused by pressure changes in special cells or organs. In certain grasses, for example, rolling and folding movements of the leaves result from turgor changes in cells on the upper surface at the base of two depressions that lie parallel to the principal veins of the leaves. When in a state of high turgor, these cells maintain the leaf in a flat, expanded condition. As turgor decreases, the pressure of cells on the bottom of the leaf forces it to curl. In the sensitive plant *Mimosa pudica*, turgor movements commence within 0.075 second after the plant is stimulated. When the plant is touched—for example, at the axis of one of the secondary petioles—the leaflets move upward on each side until they meet. The stimulus may then be communicated to the leaflets of the other secondary petioles, which close (the petioles also converging), and then to the main petiole, which sinks rapidly downward toward the stem, the bending taking place at the pulvinus (swollen base of the leafstalk). The entire reaction may be completed in scarcely more than one second. Recovery of turgor in *M. pudica* occurs in from eight to 20 minutes.

Movements in the sensitive plant

Turgor movements are also responsible for the opening and closing of stomates, which are minute pores in the epidermis (a thin surface layer of cells) of plants—particularly on the lower surfaces of leaves—and through which gaseous exchange occurs between the plant and the atmosphere. The aperture size of the stomate is determined by the turgor of two kidney-shaped guard cells that lie facing each other on either side of the stomate. When the guard cells are turgid, the opening between them widens, thus increasing the quantity of gas exchange. When turgidity decreases, the aperture becomes smaller.

Guard-cell turgor is influenced by light, temperature, and by the internal water condition of the plant. The amount of light absorbed by the cells appears to be inversely related to the state of turgor. All wavelengths of the visible spectrum seem to be effective, although light at the red end of the spectrum appears to be weaker. With the cessation of light, the stomates usually close gradually. In general, the greater the quantity of light absorbed during the day, the more slowly the guard cells lose their turgor.

Each plant species has a particular temperature range at which the guard cells are least turgid; that is to say, at which the stomatal opening is smallest. In cotton and tobacco plants, for example, guard-cell turgor is greatest at 25°–30° C (77°–86° F). At both higher and lower temperatures, the size of the stomatal opening decreases. The shrinkage of total water volume in a plant usually results in a water loss in individual guard cells as the result of osmotic (*i.e.*, through the cell wall) transfer and in redistribution of water among the various tissues and organs. The resultant reduction in turgor of the guard cells causes partial-to-complete closure of the stomatal opening.

Hydration movements. Autonomous movements other than growth movements and turgor movements may occur in dead as well as in living tissues and organs of plants and are caused by changes in water content (hydration or dehydration) of cell walls. They include the splitting of seed pods and the rapid movements of sporangia (spore-bearing organs) in ferns. Such movements do not result from physiological changes.

From P. Marler and W.J. Hamilton, *Mechanisms of Animal Behavior* (1966); John Wiley & Sons, Inc.

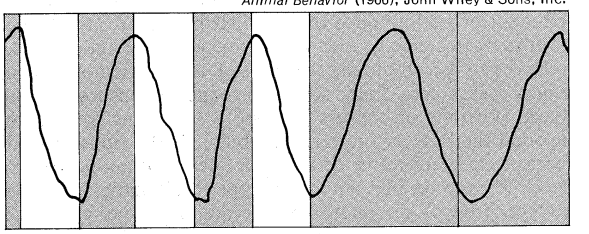


Figure 3: Circadian and tidal rhythms in *Phaseolus multiflorus* under normal day and night and continuous darkness.

Circadian rhythm in plants. Although experiments have clearly demonstrated the causal relationship between various stimuli and plant responses, it has also been shown that certain plant movements commonly associated with particular stimuli will occur in their absence.

In the laboratory, plants that move predictably according to a 24-hour cycle continue to move according to that cycle even after all extrinsic factors (*e.g.*, sunlight, temperature) known to change in a daily cyclical fashion have been stabilized or eliminated. The reason for this circadian, or daily, rhythm is not yet known (see PERIODICITY, BIOLOGICAL).

Movements in insectivorous plants. The most rapid autonomous movements in plants are probably those used in the capture of insects or other small animals by insectivorous plants. In the Venus's-flytrap (*Dionaea muscipula*), the left and right halves of the leaf blade snap together upon stimulation of hairs on the upper surface; the snapping action occurs in less than half a second, fast enough to trap even the most alert insect. The movements of the leaf probably result from a release of tautness, or tension, produced by differences in the growth rate between the upper and lower surfaces of the leaf; the mechanism by which the tension is released, however, is not yet understood. Other types of insect-eating plants respond to such stimulation even more rapidly than does the Venus's-flytrap.

Venus's-flytrap

ANIMAL MOVEMENTS

Stereotyped response in animals may be classified as unorganized or poorly organized response, reflex movements of part of an organism, reflex-like activity of an entire organism, and instinct.

Unorganized or poorly organized responses are given by early embryos or by animals (such as sponges) that lack nervous systems.

Reflex. Reflexes proper, or reflex-arc movements, include responses such as the immediate withdrawal of the hand on touching a hot surface. The basic components of the reflex arc are the receptor, or sensory-nerve cell, which senses the stimulus, and the affector, the nerve cell that directly activates the muscle. These are a theoretical minimum rather than an observed functional arrangement of cells in the body of an animal (see INSTINCT).

Reflex-like activities. Reflex-like activities of entire organisms may be unoriented or oriented. Unoriented re-

Ortho-kinesis and klinokinesis

sponses include kineses—undirected speeding or slowing of the rate of locomotion or frequency of change from rest to movement (orthokinesis) or of frequency or amount of turning of the whole animal (klinokinesis), the speed of frequency depending on the intensity of stimulation. Examples of orthokinesis are seen in lampreys, which are more active in high intensities of light, and in cockroaches, which are more active in low intensities; flatworms and many fly larvae, among other invertebrates, show orthokinesis. Klinokinesis is well demonstrated by the movements of the wood louse (*Porcellio scaber*). When wood lice are placed in dry air, they crawl about actively but without direction until they become gradually dehydrated. When the lice are placed in humid air, they move at first, but any activity they exhibited soon ceases and they become quiet. Wood lice placed in a container with dry air at one end and humid air at the other gradually congregate at the humid end. This transfer is achieved through apparently random rather than directed movements.

Oriented reflex activities of entire organisms include tropisms, taxes, and orientations at an angle. Tropisms in animals are those directed growth–curvature movements of sessile (*i.e.*, sedentary) forms that lead to equal intensities of stimulation of symmetrically placed body parts. These movements are demonstrated by hydroid animals such as *Eudendrium*.

Taxes. Taxes may be described as oriented locomotory reactions of motile organisms. They exist in purest form as oriented, forced movements; that is, as reflex actions of entire organisms. When exposed to a single source of stimulation, the body is oriented in line with the source. Movement toward the source is said to be positive; that away from it is negative.

Klinotaxis is the achievement of orientation by alternate lateral movements of part or all of a body; there appears to occur a comparison of intensities of stimulation between one position and another and a "choice" between

From J.P. Scott, *Animal Behavior*, © 1958 by the University of Chicago; all rights reserved

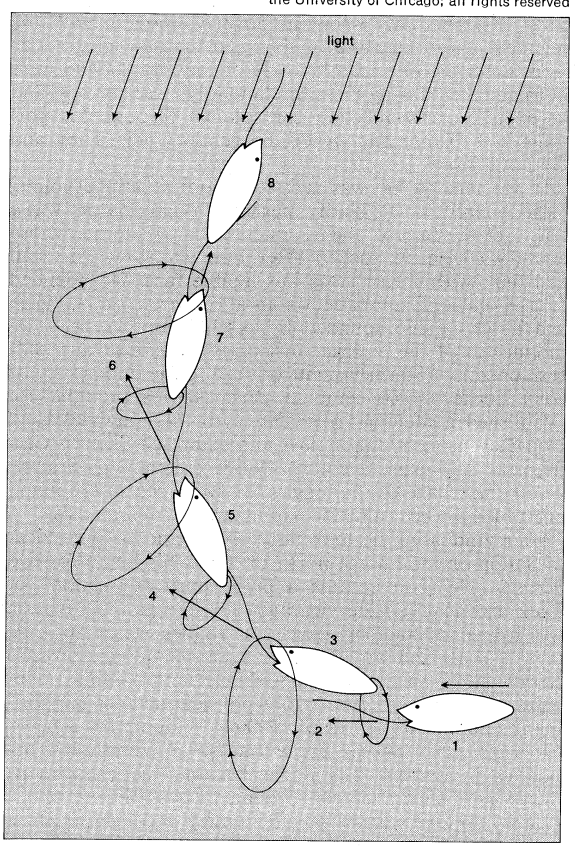


Figure 4: Klinotaxis in *Euglena*. As the animal swims in a spiral, the single eyespot at its front end swings in a circle. *Euglena* gradually orients itself so that the eyespot continues to receive even illumination.

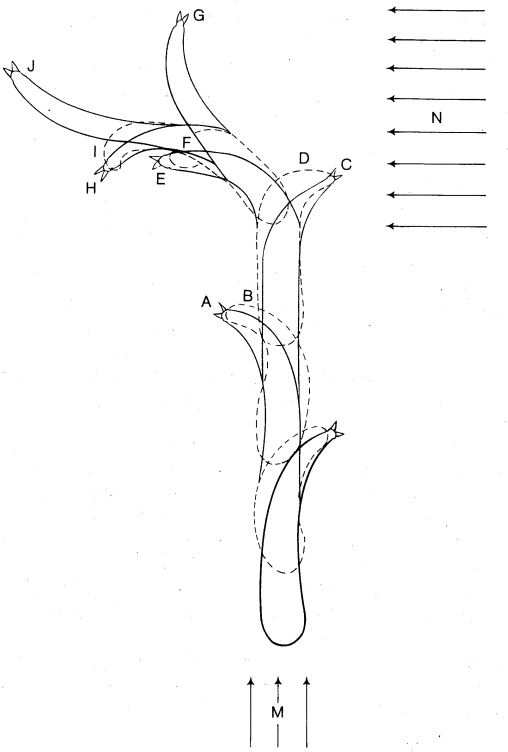


Figure 5: Klinotaxis in blow fly maggot. The maggot expands to A, contracts to B, expands to C, etc., with the light M on. At D, M is put out and N put on. The lateral movement to E is somewhat larger than A; the next one, G, increases the light falling on the receptors and is corrected by a swing over to H.
From N. Tinbergen, *The Study of Instinct;* the Clarendon Press, Oxford

them. Klinotaxis is shown by animals with a single intensity receptor such as the protozoan *Euglena*, earthworms, and fly larvae. For several days before going into the pupal (or resting) state, the blow fly maggot tends to move away from a light source. As it crawls, it swings its head alternately left and right. Presumably a light receptor is located on the maggot's head, and differences in intensity between successive light stimuli as it moves its head determine the direction in which the maggot travels. This type of response is given more commonly to chemical stimuli than to light.

Blow fly maggot response to light

In tropotaxis, attainment of orientation is direct, resulting from turning toward the less stimulated (negative) or more stimulated (positive) side as simultaneous, automatic comparisons of intensities on two sides of the body are made. No deviations (trial movements) are required. Tropotaxis is shown by animals with paired intensity receptors. If exposed to stimulation from two sources, orientation is to some intermediate point and is determined by the relative intensity of the sources. If one receptor is effectively covered, the animal moves in spirals (circus movement). Tropotaxis is shown by many arthropods, especially insects.

In telotaxis, known only for responses to light, attainment of orientation is direct and without trial movements. When between lights from two sources, the animal orients to one light, rather than to some intermediate point. The animal switches orientation from one source to the other at unpredictable intervals and consequently follows a zigzag course. The response is given to the source as though it were a goal. Bilateral balance is not necessary, and circus movements, if they occur, demonstrate that the animal is reacting tropotactically rather than telotactically. Honeybees (*Apis*) and hermit crabs (*Eupagurus*), among others, show telotaxis (see BEHAVIOUR, ANIMAL).

Orientations at an angle (transverse orientations) may or may not be accompanied by locomotion. They include the light-compass reaction (menotaxis) and dorsal (or ventral) transverse reaction. Menotaxis is shown by for-

aging insects such as ants and bees that return to a fixed nest. It has been demonstrated experimentally by covering for 2½ hours an ant returning to its nest. After being uncovered, the ant proceeds not toward the nest but at the same angle to the sun that it had been moving at the time it was hidden from the light.

In another demonstration of menotaxis, the sea slug *Elysia viridis* has been shown to move at angles of from 45° to 135° in relation to a steady source of light. No satisfactory explanation for this type of response in the sea slug is known.

Dorsal transverse action

Dorsal (or ventral) transverse reaction is demonstrated when the impact of the stimulus is kept at right angles to both longitudinal and transverse axes of the body. Locomotion need not occur. This reaction is given to light by various aquatic crustaceans—*Argulus*, the fish louse, and *Artemia*, the brine shrimp—and is given to gravity by crayfish.

Fixed-action patterns and instinct. Fixed-action patterns, commonly called FAPS, are generally sequences of uniform stereotyped responses that occur with a high degree of predictability throughout a species. Territorial behaviour in the fiddler crab (*Uca*), nest building in the three-spined stickleback fish (*Gasterosteus aculeatus*), and courtship in the green heron (*Butorides virescens*) are examples of such stereotyped response. Some fixed-action patterns, such as birdsong, although highly uniform throughout a species, often depend for normal development upon an early opportunity for a young bird to hear other birds and thus to imitate them (see BEHAVIOUR, ANIMAL).

The concept of instinct has come to refer to complex unlearned behaviour that is recognizable and predictable in at least one sex of a species (see INSTINCT).

BIBLIOGRAPHY

General Works: V.G. DETHIER and ELIOT STELLAR, *Animal Behavior: Its Evolutionary and Neurological Basis*, 2nd ed. (1964); R.A. HINDE, *Animal Behaviour: A Synthesis of Ethology and Comparative Psychology* (1966); P.R. MARLER and W.J. HAMILTON, *Methods of Animal Behavior* (1966); W.H. THORPE, *Learning and Instinct in Animals* (1963).

Specific animal groups: E.A. ARMSTRONG, *Bird Display and Behaviour: An Introduction to the Study of Bird Psychology* (1965); J.D. CARTHY, *An Introduction to the Behaviour of Invertebrates* (1958), *The Behaviour of Arthropods* (1965); H.E. EVANS, *The Comparative Ethology and Evolution of the Sand Wasps* (1966); M.W. FOX, *Canine Behavior* (1965); N.L. MUNN, *Handbook of Psychological Research on the Rat: An Introduction to Animal Psychology* (1950).

Orientation: J.D. CARTHY, *Animal Navigation: How Animals Find Their Way About* (1956); G.S. FRAENKEL and D.L. GANN, *The Orientation of Animals: Kinesis, Taxes, and Compass Reactions* (1940, reprinted 1961); J. LOEB, *Forced Movements: Tropisms and Animal Conduct* (1918).

Ontogeny of behaviour: E.L. BLISS (ed.), *Roots of Behavior: Genetics, Instinct and Socialization in Animal Behavior* (1962).

Evolution and behaviour: KONRAD LORENZ, *Evolution and Modification of Behavior* (1965); ANNE ROE and G.G. SIMPSON (eds.), *Behavior and Evolution* (1958).

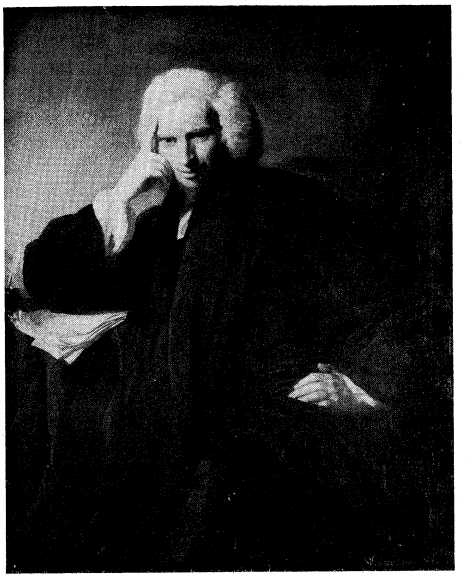
Circadian rhythm and behaviour: F.A. BROWN, JR., *Biological Clocks* (1962); F.A. BROWN, JR., J.W. HASTINGS, and J.D. PALMER, *The Biological Clock: Two Views* (1970); B.M. SWEENEY, *Rhythmic Phenomena in Plants* (1969).

Responses in plants: N.G. BALL, "Plant Tropisms," in *Vistas in Botany*, vol. 3 (1963); M.H.M. GOLDSMITH, "The Transport of Auxin," *A. Rev. Pl. Physiol.*, 19:347–360 (1968); WALTER STILES and E.C. COCKING, *An Introduction to the Principles of Plant Physiology*, 3rd ed. (1969).

Sterne, Laurence

An Anglican clergyman and humorist, Laurence Sterne was a great novelist and literary innovator although he produced only two novels, *Tristram Shandy* and *A Sentimental Journey*. His unconventional, nearly plotless books can be bitingly satirical, though usually the humour is tolerant and sympathetic. Thomas Jefferson called Sterne's work "the best course of morality that ever was written"; but the sexual innuendoes shocked many readers, and Sterne's own conduct with women dampened public enthusiasm. Today Sterne is recognized as the forerunner of modern psychological fiction.

Chichester Photographic Service



Sterne, oil painting by Sir Joshua Reynolds, 1760. In a private collection.

Family and education

His father, Roger, though grandson of an archbishop of York, was an infantry officer of the lowest rank who fought in many battles during the War of the Spanish Succession (1701–14). In Flanders, Roger married the widow of an officer, Agnes, who was probably the daughter of a poor sutler; certainly her class was much below Roger's. When England withdrew from the war, Roger's regiment returned to Ireland, and there Laurence was born at Clonmel on November 24, 1713. Most of his early childhood was spent in poverty, moving with the regiment about England or living in the barracks of Ireland. Later Sterne spoke lovingly of the infants who were born and died during their "melancholy" wanderings, and he expressed his affection for soldiers through his portraits in *Tristram Shandy* of the gentle Uncle Toby and Corporal Trim.

At age ten, Sterne was sent to school at Hipperholme, close to Halifax. His kindly master, Nathan Sharpe, was a friend of the family, and his uncle Richard Sterne, whose estate was nearby, looked after him. He grew very tall and thin, with a long nose but likable face. He attended Jesus College, Cambridge, on an allowance from a cousin and a scholarship founded by Archbishop Sterne, once the master of the college. Instruction was dull and old-fashioned, but Sterne was introduced to the philosophy of John Locke, which came to permeate his thought. He established a lifelong friendship with the convivial John Hall-Stevenson, a minor poet and squire of Skelton Castle, near Saltburn. At college Sterne suffered his first severe hemorrhage of the lungs—"it happened in the night, and I bled the bed full." He had incurable tuberculosis.

Career as a clergyman and political journalist

After graduating he took holy orders and became vicar of Sutton-on-the-Forest, north of York, his home for two decades. He soon became a prebendary (or canon) of York and acquired the vicarage of Stillington. At first he was helped by another uncle, Dr. Jaques Sterne, precentor of York and archdeacon of Cleveland, a powerful clergyman but a mean-tempered man and a rabid politician. In 1741–42 Sterne wrote political articles supporting the administration of Sir Robert Walpole for a newspaper founded by his uncle but soon withdrew from politics in disgust. His uncle became his archenemy, thwarting his advancement whenever possible.

He fell in love with Elizabeth Lumley, a cousin to Elizabeth Montagu, the bluestocking, and married her in 1741. Elizabeth soon showed a quarrelsome disposition, and Sterne grew restless. After their arguments, he would spend nights of debauchery in York. The neighbours gos-

siped. His mother, abetted by his Uncle Jaques, made life miserable, demanding that he assign to her his wife's small fortune. Sterne refused. Once Jaques had her lodged in a prison to embarrass his nephew.

As a clergyman Sterne worked hard but erratically. One Sunday, when his dog sprang some quail, he went home for his gun, leaving the congregation in the lurch. Yet he was kind to his poor parishioners, and gave them careful religious instruction. In two ecclesiastical courts he served as commissary (judge), and his frequent sermons at York Minster were very popular, until tuberculosis injured his voice in 1762. Externally, his life was typical of the moderately successful clergy. He farmed and served as a justice of the peace. For diversion he read, painted, and "fiddled." An amusing companion, he socialized with the gentry and sometimes went to Hall-Stevenson's "Crazy Castle" to meet the Demoniacs, a club of sportsmen and amateur writers. But Elizabeth, who had several stillborn children, was unhappy. Only one child lived, Lydia.

In 1759, to support his dean in a church squabble, Sterne wrote *A Political Romance* (later called *The History of a Good Warm Watch Coat*), a Swiftian satire upon dignitaries of the spiritual courts. At the demands of embarrassed churchmen, the book was burned. Thus, Sterne lost his chances for clerical advancement but discovered his real talents. Turning over his parishes to a curate, he began *Tristram Shandy*. An initial version, probably sharply satirical, was rejected by Robert Dodsley, the London printer, just when Sterne's personal life was upset. His mother and uncle both died. His wife had a nervous breakdown and threatened suicide. Sterne continued his comic novel, but every sentence, he said, was "written under the greatest heaviness of heart." In this mood, he softened the satire and told about Tristram's opinions, his eccentric family, and ill-fated childhood with a sympathetic humour, sometimes hilarious, sometimes sweetly melancholic—a comedy skirting tragedy.

Success of *Tristram Shandy*

Sterne himself published volumes 1 and 2 at York late in 1759, but he sent half of the imprint to Dodsley to sell in London. While awaiting the outcome, he carried on a gay flirtation with Catherine Fourmantel, a concert singer. By March, when he went to London, *Tristram Shandy* was the rage, and he was famous. Dodsley's brother, James (the new proprietor), brought out a second edition. A volume of sermons followed. The witty, naughty "Tristram Shandy," as Sterne was called, or "Parson Yorick"—after his self-portrait in the book—was the most sought-after man in town. Although the timing was coincidental, Lord Fauconberg, a Yorkshire neighbour, presented him with a third parish, Coxwold.

Sterne returned north joyfully to settle at Coxwold in his beloved "Shandy Hall," a charming, odd, old house, now a museum. He began to write at Shandy Hall during the summers, going to London in the winter to publish what he had written. James Dodsley brought out two more volumes of *Tristram Shandy*; thereafter, Sterne became his own publisher. In London he enjoyed the company of many great people, but his nights were sometimes wild. In 1762, after almost dying, he fled the damp air of England into France, a journey he described as Tristram's flight from death. His illness took away his voice, which he never fully recovered. Nevertheless, this and a later trip abroad gave him much material for his later *Sentimental Journey*. Elizabeth, now recovered, followed to France, where she settled permanently. Sterne returned to England alone.

In 1767, after publishing the final volume of *Tristram Shandy*, he met Elizabeth Draper, the most important love of his life other than his wife. "Eliza," unhappily married to an official of the East India Company, was half Sterne's age, but they were drawn together by similar temperaments and interests. They carried on an open, sentimental flirtation until Daniel Draper commanded his wife to return to Bombay. After her departure, Sterne, though very weak, began the so-called "Journal to Eliza," a lachrymose narrative of heartache that he never attempted to publish. That autumn he finished *A Sentimental Journey*. He published it and collapsed.

On March 18, 1768, lying in his London lodgings, he put up his arm as though to ward off a blow, saying, "Now it is come," and died. Soon after burial at London, Sterne's body was stolen by grave robbers, taken to Cambridge, and used for an anatomy lecture. Someone recognized the body, and it was quietly returned to the grave. The story, only whispered at the time, was confirmed in 1969: the old graveyard being sold for apartment buildings, the remains were exhumed. They now rest in the churchyard at Coxwold, close to Shandy Hall.

MAJOR WORKS

The Life and Opinions of Tristram Shandy, Gentleman, 9 vol. (1759–67); *A Sentimental Journal Through France and Italy*, 2 vol. (1768); *The Sermons of Mr. Yorick*, 4 vol. (1760–66); *Sermons by the Late Rev. Mr. Sterne*, 3 vol. (1769).

BIBLIOGRAPHY. WILBUR L. CROSS, *Life and Times of Laurence Sterne*, 3rd ed. rev. (1929), the standard biography; *Letters of Laurence Sterne*, ed. by L.P. CURTIS (1935), letters and memoirs, fully annotated; *A Sentimental Journey*, ed. by GARDNER D. STOUT, JR. (1967), a scholarly edition including biographical information; LODWICK HARTLEY, *Laurence Sterne in the Twentieth Century* (1966), a bibliography of modern studies.

(A.H.Ca.)

Steroids

Steroids are a class of naturally occurring organic substances and their derivatives that are of great importance in biology, medicine, and chemistry. The steroid group includes all sex hormones, adrenal cortical hormones, bile acids, and sterols of vertebrates, the molting hormones of insects, and many other physiologically active substances of animals and plants. Among the synthetic steroids of therapeutic value are a large number of anti-inflammatory agents, anabolic (growth-stimulating) agents, and oral contraceptives. Different categories of steroids are frequently distinguished from each other by names that relate to their biological source—*e.g.*, phytosterols (found in plants), adrenal steroids, bile acids—or to some important physiological function—*e.g.*, gestogens (promoting gestation), androgens (favouring development of masculine characteristics), cardiotonic steroids (facilitating proper heart function). Because steroids of similar function and origin tend to fall into families of closely related structure, biological categorization underlies an internationally accepted systematic chemical nomenclature.

HISTORICAL BACKGROUND

Studies of steroids commenced in the early 19th century with investigations of the unsaponifiable (*i.e.*, remaining undissolved after heating with excess of alkali) material, largely cholesterol, of animal fat and gallstones and of acids obtainabe from bile. This early work, with which many of the noted chemists of the time were associated, led to the isolation of cholesterol and some bile acids in reasonable purity and established some significant features of their chemistry. Insight into the complex polycyclic steroid structure, however, came only after the turn of the 20th century, following the consolidation of chemical theory and the development of chemical techniques by which such molecules could be broken down step by step. Arduous studies, notably by the research groups of the German chemists Adolf Windaus and Heinrich Wieland, ultimately established the structures of cholesterol; of the related sterols, stigmasterol and ergosterol; and of the bile acids. Investigation of ergosterol was stimulated by the realization that it can be converted into vitamin D. Only in the final stages of this work (1932) was the arrangement of the component rings of the nucleus clarified by results obtained by pyrolytic (heat-induced bond-breaking) dehydrogenation and X-ray crystallography.

The foundations of steroid chemistry

With the foundations of steroid chemistry firmly laid, the next decade saw the elucidation of the structures of most of the physiologically potent steroid hormones of the gonads and the adrenal cortex. Added impetus was given to steroid research when the Americans Philip S.

Hench and Edward C. Kendall announced in 1949 that the thitherto intractable symptoms of rheumatoid arthritis were dramatically alleviated by the adrenal hormone cortisone. New routes of synthesis of steroids were developed, and many novel analogues were therapeutically tested in a variety of disease states. From these beginnings there has developed a flourishing steroid pharmaceutical industry and, with it, a vastly expanded fundamental knowledge of steroid reactions that has influenced many other areas of chemistry (see, for example, CONFORMATIONAL ANALYSIS, PRINCIPLES OF).

Knowledge of the biochemistry of steroids has grown at a comparable rate, assisted by the use of radioisotopes and new analytical techniques. The metabolic pathways (sequences of chemical transformations in the body), both of synthesis and of decomposition, have become known in considerable detail for most steroids present in mammals, and much current research relates to control of these pathways and to the mechanisms by which steroid hormones exert their effects. The hormonal role of steroids in lower organisms is also of growing interest.

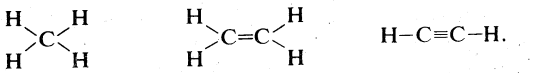
GENERAL CHEMISTRY

Organic compounds. Steroids are like many other classes of natural products in that small modifications in their molecular structures can produce remarkable differences in their biological activities. Appreciation of the relationships between structure and activity requires some familiarity with the general principles of organic chemistry; that topic is treated in detail in the article CHEMICAL COMPOUNDS, ORGANIC, and only a glimpse of it will be presented here.

Organic chemistry, the study of compounds of carbon, traditionally has been one of the major subdivisions of the science of chemistry. The reason for setting the compounds of this one element apart from those of all the 100-odd others is that the carbon atom can form stable bonds not only with other kinds of atoms but also with other carbon atoms. This property multiplies the possibilities of compound formation because it allows the existence of molecules in which any number of carbon atoms are linked together in chains, rings, or three-dimensional networks to which atoms of other elements can be attached. The most common elements other than carbon found in organic compounds are hydrogen, oxygen, and nitrogen.

Carbon bonds

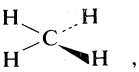
The bonds that hold carbon atoms in combination with other atoms are almost always of the type described as covalent and result from sharing of valence electrons between each pair of bonded atoms, each atom contributing one, two, or three electrons to form a single, double, or triple bond. (The other types of chemical bonds, electrovalent and metallic, are less important in organic chemistry.) Every carbon atom has four electrons that participate in formation of covalent bonds with four, three, or two other atoms, as in the compounds methane, ethylene, and acetylene. The composition of the molecules is indicated by molecular formulas: the methane (CH_4) molecule contains one atom of carbon and four of hydrogen; ethylene (C_2H_4) contains two atoms of carbon and four of hydrogen; and acetylene (C_2H_2) contains two of carbon and two of hydrogen. The arrangements of the bonds in these molecules are depicted by structural formulas in which each pair of shared electrons is represented by a line and each atom by its chemical symbol, as C for carbon and H for hydrogen:



methane ethylene acetylene

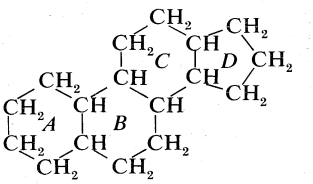
Additional important features of covalent bonds are that they are specifically directed in space and are effective only if the bonded atoms remain at characteristic distances from one another. The angles between the bonds in the acetylene molecule are 180°, so that the four atoms lie in a straight line; the angles in ethylene are close to 120°, and all six atoms lie in a plane; the angles in methane are close to 109½°, and the four hydrogen atoms lie at the corners of an imaginary tetrahedron with the carbon atom at the centre. Distortions of the characteristic bond angles or distances decrease the stability of the molecule. Methane is an example of a saturated compound, so called because the carbon atom in its molecule is bonded to the largest possible number—four—of other atoms, in contrast to ethylene and acetylene, which are unsaturated compounds.

The structural formulas shown for ethylene and acetylene adequately represent the spatial arrangements of the atoms, but a nonplanar molecule such as that of methane is difficult to illustrate by a two-dimensional drawing. When as in discussing the steroids, it is important to emphasize the orientation, or configuration, of the bonds about an atom, arbitrary conventions are used. For example in



representing methane, the full lines stand for bonds lying in the plane of the paper, the wedge stands for a bond projecting up from the page toward the viewer, and the dotted line stands for a bond projecting beneath the page away from the viewer.

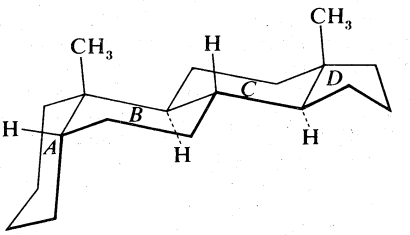
Steroid numbering system and nomenclature. All steroids are related to a characteristic molecular structure composed of 17 carbon atoms—arranged in four rings conventionally denoted by the letters *A*, *B*, *C*, and *D*—bonded to 28 hydrogen atoms. This parent structure, indicated in Diagram 1,



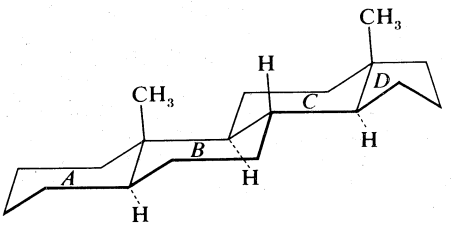
(1)

named gonane and often referred to as the steroid nucleus, may be modified in a practically unlimited number of ways by removal, replacement, or addition of a few atoms at a time; hundreds of steroids have been isolated from plants and animals, and thousands more have been prepared by chemical treatment of natural steroids or by synthesis from simpler compounds.

The steroid nucleus is a three-dimensional structure, and atoms or groups are attached to it by spatially directed bonds. Although many stereoisomers (compounds that are alike bond for bond but different in three-dimensional arrangement of atoms) of this nucleus are possible (and may be synthesized), the saturated nuclear structures of most classes of natural steroids are alike, except at the

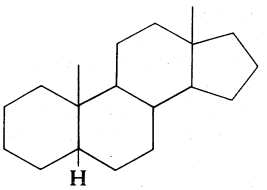


(2)

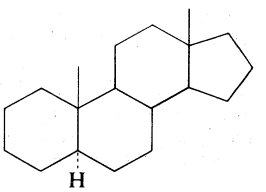


(3)

junction of rings *A* and *B*. Simplified three-dimensional diagrams may be used to illustrate stereochemical details. For example, Diagrams 2 and 3 show the alternative forms of the skeletal structure, androstane, common to a number of natural and synthetic steroids, in which the *A/B* ring fusions are called *cis-* and *trans-*, respectively. Note that at the common side of rings *A* and *B* in Diagram 2, bonds to the methyl group, CH_3, and to the hydrogen atom, H, both project upward from the general plane defined by the rest of the molecule, whereas in diagram 3, CH_3 projects up, and H projects down. Usually, however, steroid structures are represented as plane projection diagrams such as 4 and 5, which correspond to 2 and 3, respectively.



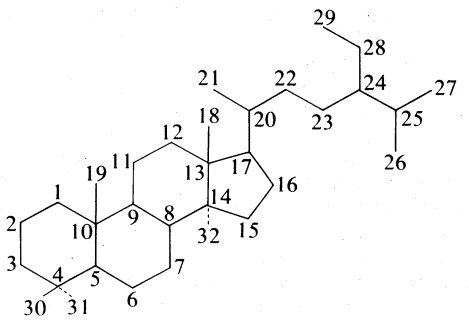
(4)



(5)

The stereochemistry of rings *A* and *B* must be specified by showing the orientation of the hydrogen atom attached at C-5 (that is, carbon atom number five; steroid numbering is explained below) as either above the plane of the diagram (designated β) or below it (α). The α-, β-symbolism is used similarly to indicate the orientation of any substituent group that is attached to a saturated (fully substituted) carbon of the steroid ring system. Groups attached to unsaturated carbons lie in the same plane as the adjacent carbons of the ring system (as in ethylene), and no orientation need be specified. When the orientation of a substituent is unknown, it is assigned the symbol ξ- (xi). Bonding of β-attached substituents is shown diagrammatically as in 4 by a full line; that of α-substituents by a broken line, as in 5, and that of ξ-substituents by a wavy line.

Each carbon atom of a steroid molecule is numbered, and the number is reserved to a particular position in the hypothetical parent skeletal structure (6) whether this position is occupied by a carbon atom or not.



(6)

Principal structural types

Steroids are named by modification of the names of skeletal root structures according to systematic rules agreed upon by the International Union of Pure and Applied Chemistry (IUPAC). By attaching prefixes and suffixes to the name of the appropriate root structure, the character of substituent groups or other structural modification is indicated. The prefixes and suffixes include numbers, called locants, indicative of the position in the carbon skeleton at which the modification occurs, and, where necessary, the orientation of a substituent is shown as α- or β-. The carbon atom at position 3, for example, is referred to as C-3; a hydroxyl group attached to C-3 is referred to as a 3-OH group or, more specifically, as a 3α-OH or 3β-OH group. In addition to differences in details of the steroid nucleus, the various classes of steroids are distinguished by variations in the size and structure of an atomic group (the side chain) attached at position 17. The derivations of the names of the more common root structures from those of naturally occurring compounds or classes of compounds of which they are most typical are indicated in Table 1.

Table 1: Names of Fundamental Structures Most Often Used in Steroid Nomenclature

	carbon atoms present (as numbered in Diagram 6)	naturally occurring general classes	examples shown in the text
Gonane	1–17	none	gonane* (1)
Estrane	1–18	estrogens	estradiol (17 f)
Androstane	1–19	androgens	androstane* (2–5); testosterone (17 c); androstanedione (17 c)
Pregnane	1–21	gestogens and adrenal steroids	progesterone (17 a); cortisol (17 b); aldosterone (17 d); batrachotoxin (14)
Cholane	1–24	bile acids	cholic acid (10 a); sodium glycocholate (10 b)
Cholestane	1–27	sterols	cholesterol (16 d); scymnol (11 a)
Ergostane	1–28	sterols	ergosterol (8); cyasterone (22)
Stigmastane	1–29	sterols	stigmasterol (7); antheridiol (12)
Lanostane	1–27, 30–32	trimethyl sterols	lanosterol (16 c); 22,25-oxido-holothurinogenin (13)
Cardanolide	1–23	cardiac glycosides	digitoxigenin (23)
Bufanolide	1–24	toad poisons	bufotoxin (24)
Spirostan	1–27	sapogenins	dioscin (25)

*Gonane and androstane themselves do not occur in nature.

For unambiguous use of the names given in Table 1, the orientation (α or β) of hydrogen at C-5 must be specified. If no other modification is indicated, the nucleus is assumed to be as shown in 2 and 3, except in the cardanolides and bufanolides: compounds of these types characteristically possess the 5β,14β configurations which, however, are specified.

For brevity in discussion and in trivial nomenclature, a number of prefixes are often attached, with locants, to the names of steroids to indicate specific modifications of the structure. In addition to the usual chemical notations for substituent groups replacing hydrogen atoms (*e.g.*, methyl-, chloro-, hydroxy-, oxo-), the following prefixes are commonly used: dehydro- (lacking two hydrogen atoms from adjacent positions); dihydro- (possessing two additional hydrogen atoms in adjacent positions); deoxy- (hydroxyl group replaced by a hydrogen atom); epi- (differing in configuration of a carbon atom bonded to two other carbon atoms); iso- (differing in configuration of a carbon atom bonded to three other carbon atoms); nor- (lacking one carbon atom); homo- (possessing one additional carbon atom); cyclo- (with a bond between two carbons that are normally not united); and seco- (with a carbon–carbon bond of the nucleus broken).

Depending upon the number and character of their functional groups, steroid molecules may show diverse reactivities. Moreover, the reactivity of a functional group varies according to its location within the molecule (*e.g.*, esters are formed readily by 3-OH groups but only with difficulty by the 11β-OH group). An important property of steroids is polarity; *i.e.*, their solubility in oxygen-containing solvents (water, alcohols) rather than hydrocarbon solvents (hexane, benzene). Hydroxyl, ketonic, or ionizable (*i.e.*, capable of dissociating to form electrically charged particles) groups in a steroid molecule increase its polarity to an extent that is strongly influenced by the spatial arrangement of the atoms within the molecule.

Polarity

Methods of isolation. Procedures for isolation of steroids differ according to the chemical nature of the steroids and the scale and purpose of the isolation. Steroids are isolated from natural sources by extraction with organic solvents, in which they usually dissolve more readily than in the aqueous tissue fluids. The source material often is treated initially with an alcoholic solvent, which dehydrates it, denatures (renders insoluble) proteins associated with the steroids, and dissolves many steroids. Saponification (treatment with aqueous alkali) either of whole tissues or of substances extracted from them by

alcohol splits the molecules of sterol esters, triglycerides, and other fatty esters and permits the extraction of the sterols by means of water-immiscible solvents, such as hexane or ether, with considerable purification. Intact sterol esters or hormonal steroids and their metabolites (compounds produced by biological decomposition) that are sensitive to strong acids or alkalies, however, require essentially neutral conditions for isolation, and, although some procedures for analysis of urinary steroids employ acid treatment, milder hydrolysis, as by enzymes, is preferred. The acidity of some steroids allows them to be held in alkaline solution while nonacidic impurities are extracted with organic solvents.

Commercially, abundant steroids usually are purified by repeated crystallization from solvents. Small-scale laboratory isolations for investigative or assay purposes usually exploit differing polarities of the steroid and of its impurities, which may be separated by partition between solvents differing in polarity or by chromatography (see below). Occasionally, special reagents may selectively precipitate or otherwise sequester the desired steroid. A classical example is the precipitation of 3β-hydroxy sterols such as cholesterol by the natural steroid derivative digitonin. New steroids of great physiological interest often are isolated from tissue only with extreme difficulty because they are usually trace constituents. For example, 500 kilograms (1,100 pounds) of silkworm pupae yielded 25 milligrams of pure molting hormone, the steroid ecdysone (*i.e.*, 20×10^6-fold purification). In such cases each isolation step is followed by an assay for the relevant physiological activity to ensure that the desired material is being purified. The percentage recovery of known steroid hormones during their assay in small biological samples usually is assessed by adding a trace of the same steroid in radioactive form to the initial sample, followed by radioassay (analysis based on radioactivity) after purification is complete. The efficiency of recovery of the radioactive steroid is assumed to be the same as that of the natural substance.

Determination of structure and methods of analysis. The systematic, stepwise breakdown by chemical methods of the steroid ring systems, used in early investigations of structure, is now mainly of historical interest. The small number of different nuclear structures found in steroids often has permitted establishment of the structure of a new steroid by conversion to related compounds of known structure. Structure elucidation in the steroid field, as in all areas of organic chemistry, now depends heavily upon physical methods, particularly nuclear magnetic resonance, infrared spectroscopy, optical rotatory dispersion, mass spectrometry, and X-ray crystallography. Data obtained by these methods (which are described in CHEMICAL ANALYSIS) reinforce and often replace the classical criteria of characterization of steroids: melting point, optical rotation, elemental analysis, and ultraviolet absorption at a fixed wavelength.

Chromatography

Chromatography (*q.v.*) is a crucial technique in steroid chemistry. The behaviour of a steroid in selected chromatographic systems often identifies it with a high degree of probability. The identification may be made virtually certain by the conversion of the material to derivatives that in turn are examined chromatographically. Abundant data for the behaviour of steroids in paper chromatography, thin-layer chromatography, and gas-liquid chromatography show that individual features of molecular structure determine the chromatographic properties of steroids in a predictable manner. Because gas–liquid chromatography is the most precisely controllable technique, it provides the most valuable data for structural identification. The gas–liquid chromatograph linked directly to the mass spectrometer permits characteristic mass-spectral fragmentation patterns and critical gas–liquid chromatographic data to be obtained simultaneously, using a sample containing less than a microgram of a steroid. This powerful technique is of growing importance in the structural analysis of steroids in extracts of such body fluids as blood and urine.

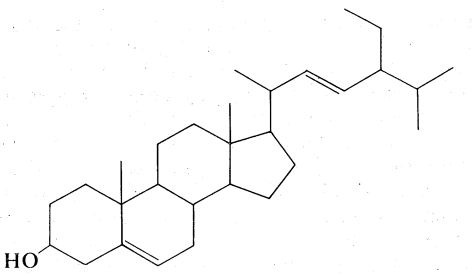
Synthesis of steroids. *Total synthesis.* The total synthesis of any steroid (*i.e.*, synthesis from simple organic chemicals by reactions that follow known pathways) is formally the final proof of its chemical structure. To reproduce all stereochemical details of the naturally occurring material, however, is an exacting intellectual and technical challenge that has been met with extraordinary ingenuity and originality by several chemists.

In most total syntheses of steroids, a monocyclic starting material such as a quinone provides one ring upon which the other rings of the nucleus are elaborated, step by step, by condensation reactions with smaller molecules to give the desired stereochemistry in successive ring fusions. Each new ring closure must also provide functional groups that can be used in building up the next ring. In a quite different approach, stereochemical control of ring fusions is achieved by using the fact that, under acidic conditions, open-chain molecules containing suitably located double bonds cyclize to multi-ring structures that have the necessary stereochemistry and that can be relatively easily converted to steroids. From its analogy with the cyclization of squalene 2,3-oxide to lanosterol in the biosynthesis of cholesterol (see below), this method is said to involve "biogenetic-type cyclization."

Partial synthesis. Although total synthesis of steroids has proved commercially feasible, it is often more practical to prepare them by partial synthesis; that is, by modification of other steroids that are naturally abundant. To be useful as a starting material for partial synthesis, the naturally occurring steroid must possess a molecular structure that can be easily converted to that of the desired product. For the synthesis of cortisol, cortisone, and their analogues, which carry an oxygen function at C-11, a pre-existing oxygen function at this position or at the adjacent C-12 is highly desirable. Indeed, prior to the advent of methods for microbiological oxidation, this was a crucial requirement, since the introduction of any functional group at C-11 of most steroids was extremely difficult.

Commercial synthesis

In the early commercial synthesis of androgenic steroids, cholesterol was the main starting material. Cholic acid and deoxycholic acid, inexpensive by-products from slaughterhouses, were starting materials for production of cortisone. The bulk of steroid drugs now is manufactured from the abundant steroids of plant origin, notably the sapogenins. Diosgenin, obtainable from several varieties of *Dioscorea*, a Mexican yam, is used in the commercial manufacture of progesterone, which can be converted to androgenic and estrogenic hormones and to the more complex adrenal steroid hormones, such as cortisone and cortisol. A most important advance in this field was the discovery that micro-organisms such as *Rhizopus nigricans* introduce hydroxyl groups into a variety of steroids at C-11 and elsewhere: they are used in the commercial synthesis of a large number of steroid hormone analogues. A sapogenin, hecogenin, obtainable in quantity from the waste of sisal plants, is used for synthesis of cortisol. Stigmasterol, which is readily obtainable from soybean oil, can be transformed easily to progesterone and to other hormones, and commercial processes based on this sterol have been developed.



stigmasterol (7)

NATURAL DISTRIBUTION AND FUNCTIONS

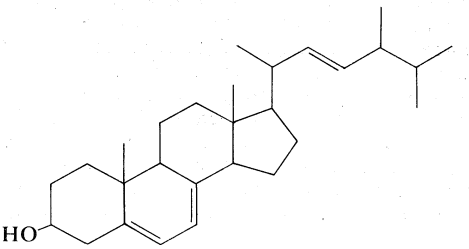
Sterols and bile acids. The most generally abundant steroids are sterols, which occur in all tissues of animals, green plants, fungi, and yeasts. Evidence for the presence of steroids in bacteria and in primitive blue-green algae is

Table 2: Some Important Sterols and Their Sources

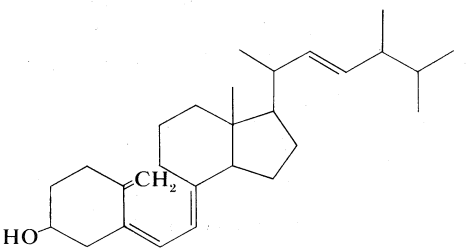
common name	systematic name	occurrence
Cholesterol	5-cholesten-3β-ol	principal sterol of most animals and all vertebrate tissues
Coprostanol	5β-cholestan-3α-ol	feces of vertebrates
Cholestanol	5α-cholestan-3β-ol	minor vertebrate sterol: guinea pig and rabbit adrenal
Lathosterol	5α-cholest-7-en-3β-ol	vertebrate skin, intestine
7-Dehydrocholesterol	5,7-cholestadien-3β-ol	mammalian skin, intestine
Desmosterol	5,24-cholestadien-3β-ol	chick embryo, barnacle (*Balanus glandula*)
Zymosterol	5α-cholesta-8,24-dien-3β-ol	minor sterol of yeasts
Ergosterol	5,7,22-ergostatrien-3β-ol	principal sterol of yeasts, ergot (*Claviceps purpurea*), and other fungi
Stigmasterol	5,22-stigmastadien-3β-ol	most green plants, soy beans
Sitosterol	5-stigmasten-3β-ol	most green plants, wheat germ
Fucosterol	5,24(28)-stigmastadien-3β-ol	principal sterol of marine brown algae (*Fucus* species)
Lanosterol	8,24-lanostadien-3β-ol	skin, sheep wool fat, yeasts
Lophenol	4α-methyl-5α-cholest-7-en-3β-ol	skin, intestine, feces, cactus (*Lophocerus schotti*)
Cycloartenol	9,19-cyclo-24-lanosten-3β-ol	generally minor sterol of green plants (*Artocarpus* species)

conflicting. The major sterols (Table 2) of most tissues are accompanied by traces of their precursors—lanosterol in animals and cycloartenol in plants—and of intermediates between these compounds and their major sterol products. In mammalian skin, one precursor of cholesterol, 7-dehydrocholesterol, is converted by solar ultraviolet light to cholecalciferol, vitamin D_3, which controls calcification of bone by regulating intestinal absorption of calcium. The disease rickets, which results from lack of exposure to sunlight, can be treated by administration of the vitamin or of the corresponding derivative of ergosterol, ergocalciferol (vitamin D_2).

Rickets and atherosclerosis



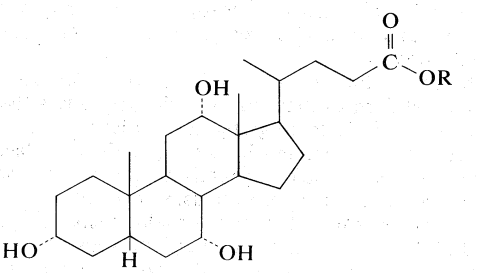
ergosterol (8)



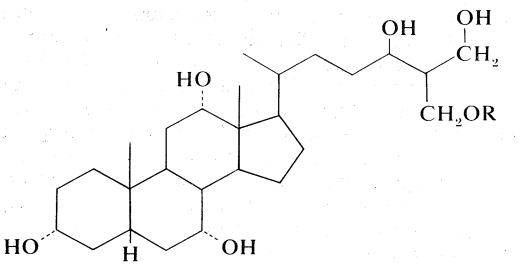
ergocalciferol (9)

Sterols are present in tissues in both the non-esterified (free) form and as esters of aliphatic fatty acids. In the disease atherosclerosis, fatty materials containing cholesterol form deposits (plaques), especially in the walls of the major blood vessels, and vascular function may be fatally impaired. The disease has many contributory factors but typically is associated with elevated concentrations of cholesterol in the blood plasma. One aim of medical treatment is to lower the plasma cholesterol level.

Free sterols appear to stabilize the structures of cellular and intracellular membranes. Because the sheath of nerve fibres is a deposit of many layers of the membranes of neighbouring cells, mature mammalian nerve tissue (*e.g.*, beef brain) is the richest source of cholesterol. Cholesterol also is converted in animals to steroids that have a variety of essential functions, and in plants to steroids the functions of which are less clearly understood. The bile acids (cholanoic acids, earlier called cholanic acids) of higher vertebrates form conjugates with the amino acids taurine and glycine, and the bile alcohols (cholane derivatives) of lower animals form esters with sulfuric acid (sulfates; Table 3). These conjugates and sulfates enter the intestine as sodium salts and assist in the emulsification and absorption of dietary fat, processes that may be



cholic acid (10a; R = H) and sodium glycocholate (10b; R = $NHCH_2COO^-Na^+$)



scymnol (11a; R = H) and sodium scymnol sulfate (11b; R = $SO_3^-Na^+$)

Table 3: Evolutionary Relationships of Bile Alcohols and Bile Acids

species	bile alcohol or bile acid (trivial name)
Elasmobranch fish	
Myxine glutinosa (hagfish)	5α-cholestane-3β,7α,16α,26-tetrol (myxinol)
Mustelus manazo (dogfish)	5β-cholestane-3α,7α,12α,24,26,27-hexol (scymnol)
Teleostian fish	
Latimeria chalumnae (coelacanth)	5α-cholestane-3β,7α,12α,26,27-pentol (latimerol)
Cyprinus carpio (carp)	5α-cholestane-3α,7α,12α,26,27-pentol (5α-cyprinol)
Amphibians	
Rana temporaria (frog)	5α,27-norcholestane-3α,7α,12α,24,26-pentol (5α-ranol)
Bufo vulgaris japonica (toad)	5β-cholestane-3α,7α,12α,25,26-pentol (5β-bufol)
Reptiles	
Alligator mississippiensis (alligator)	3α,7α,12α-trihydroxy-5β-cholestan-27-oic acid (trihydroxycoprostanoic acid)
Bitis arietans (puff adder)	3α,12α,23-trihydroxy-5α-cholanoic acid (bitocholic acid)
Bird	
Anser domesticus (goose)	3α,7α-dihydroxy-5β-cholanoic acid (chenodeoxycholic acid)
Mammals	
Man and most (but not all) species	3α,7α,12α-trihydroxy-5β-cholanoic acid (cholic acid) 3α,12α-dihydroxy-5β-cholanoic acid (deoxycholic acid) 3α-hydroxy-5β-cholanoic acid (lithocholic acid)

impaired when bile acid secretion is reduced, as in some liver diseases and in obstructive jaundice. The mixture of bile acids found in feces reflects the actions of intestinal micro-organisms upon the primary bile acid secretory products (*e.g.*, deoxycholic acid arises by bacterial transformation of cholic acid).

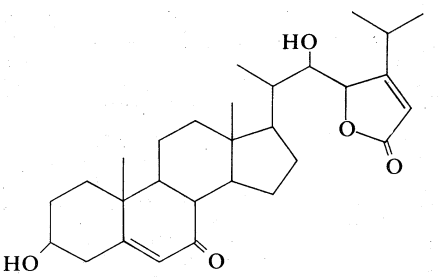
Sex hormones. Steroids that have a phenolic ring *A* (that is, those in which ring *A* is aromatic and bears a hydroxyl group) are ubiquitous products of the ovary of vertebrate animals. These are the estrogens, of which estradiol is the most potent. They maintain the female reproductive tissues in a fully functional condition, promote the estrous state of preparedness for mating, and stimulate development of the mammary glands and of other feminine characteristics. Estrogenic steroids have been isolated from urines of pregnant female mammals of many species including man, from placental and adrenal tissues, and, unexpectedly, from the testes and urines of stallions.

Estrogens and androgens

The corpus luteum, a modification of vertebrate ovarian tissue that forms following ovulation, produces the gestogens, which include progesterone and its derivatives. Gestogens also are secreted by the adrenals and placenta in some species. These compounds, in combination with the estrogens, regulate the metabolism of the uterus to permit implantation and subsequent development of the fertilized ovum in mammals. In birds, estrogens and gestogens stimulate the development of the oviduct and its secretion of albumin. Estrogens and gestogens suppress ovulation (release of the mature egg cell from the ovary); this fact is the basis of action of steroid anti-fertility drugs (see below). Estrogens and gestogens occur in primitive invertebrates, but their functions in those animals are obscure.

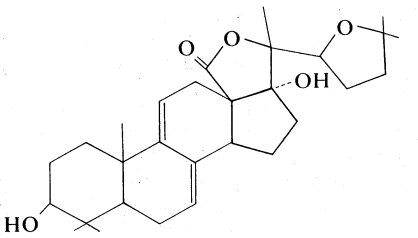
In male vertebrates, the androgens—steroids secreted by the testes—maintain spermatogenesis and the tissues of the reproductive tract. They promote male sexual behaviour and aggressiveness, muscular development, and, in man, the growth of facial and body hair and deepening of the voice. Testosterone and androstenedione are the principal androgens of the testes. Testosterone is more potent than androstenedione, but in the sexual tissues it appears to be converted to 5α-dihydrotestosterone, an even more potent androgen.

Adrenal hormones. The adrenal cortex of vertebrates synthesizes oxygenated progesterone derivatives. These compounds are hormones that are vital to survival and are classified according to their biological activity. The glucocorticoids promote deposition of glycogen in the liver and the breakdown of body proteins. Mineralocorticoids stimulate retention of sodium in the extracellular body fluids. Cortisol is the principal glucocorticoid in many species including man; in most rodents, this role is filled by corticosterone. The most potent mineralocorticoid of all species is aldosterone. Aldosterone has about 20 percent of the glucocorticoid activity of cortisol, which, conversely, has about 0.1 percent of the mineralocorticoid activity of aldosterone. Either steroid can maintain life in an animal from which the adrenal glands have been removed. The secretion of glucocorticoids is exquisitely responsive to injury and fear in animals and is primarily responsible for metabolic adaptation to stressful conditions. Failure of the adrenal cortex in man gives rise to Addison's disease, a formerly fatal condition that can now be successfully treated with synthetic adrenal steroids.



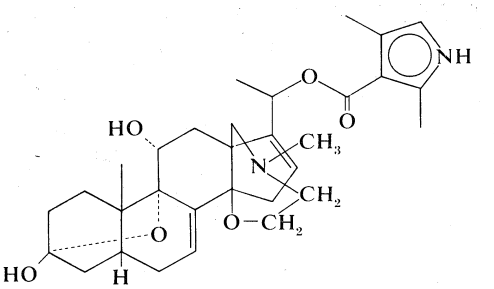
antheridiol (12)

Steroids of lower organisms. An area of increasing interest is the role of steroids in the reproduction, development, and self-defense of lower organisms. In an aquatic fungus, *Achlya bisexualis*, the steroid antheridiol (12) of the female stimulates male gamete formation. Insects and crustaceans produce the ecdysones, steroid hormones that promote molting and the development of adult characteristics (Table 4). Many plants, especially ferns and conifers, contain steroids that may protect them against some predatory insects, although this function is not established. Progesterone, 11-deoxycorticosterone, and related steroids with no known endocrine function in insects are released into the water by several species of water beetles to repel predatory fish, and the sea cucumbers (Holothuroideae) produce the holothurinogenins, a group of lanosterol derivatives toxic to nervous tissue.

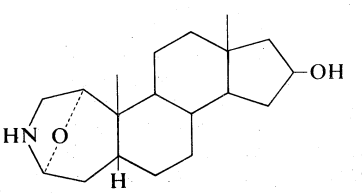


22,25-oxidoholothurinogenin (13)

Cardanolide and bufanolide derivatives, found in many plants and in the skin of toads, cause vomiting, visual disturbances, and slowing of the heart in vertebrates and are strong deterrents to predators. Birds and other predators avoid certain grasshoppers and butterflies that store cardenolides of the plants upon which they feed. The skin of the Colombian frog, *Phyllobates aurotaenia*, produces a deadly alkaloid, batrachotoxin, which is used by tribal peoples as an arrow poison. The skin of salamanders secretes a comparably poisonous alkaloid—samandarin.



batrachotoxin (14)



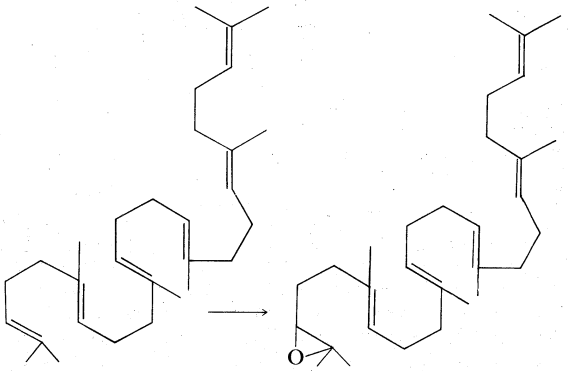
samandarin (15)

Many steroid alkaloids occur in plants, but their functions, like those of the steroid saponins, are unknown. It is possible that the taste of many of these compounds deters grazing animals or that they attract certain insect species to the plant.

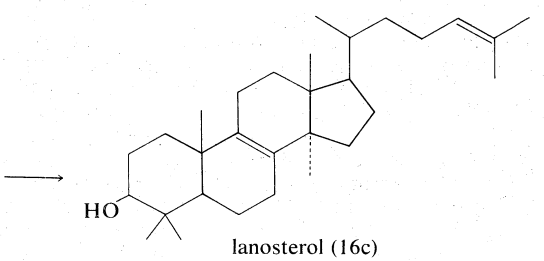
BIOSYNTHESIS AND METABOLISM OF STEROIDS

Cholesterol. Steroids are probably synthesized in all vertebrates and in many invertebrates by the same pathway, which includes cholesterol. Biosynthesis (formation in intact organisms) of cholesterol is especially vigorous in the liver of vertebrates but also occurs in the intestine, gonads, skin, and immature brain. It is barely detectable in the adult brain. The insects, certain mollusks, annelids, and some protozoa do not synthesize cholesterol but must obtain it, or a related sterol, in their diet.

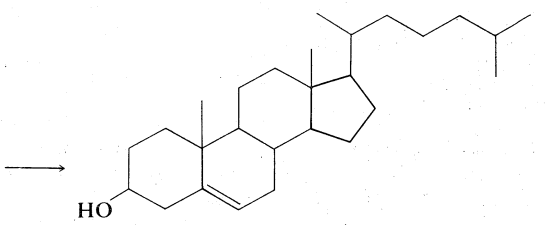
Cholesterol and other steroids are biosynthesized by extension of the enzyme pathway by which terpenoids are synthesized (see ISOPRENOIDS).



squalene (16a) squalene 2,3-oxide (16b)



lanosterol (16c)



cholesterol (16d)

biogenesis of cholesterol

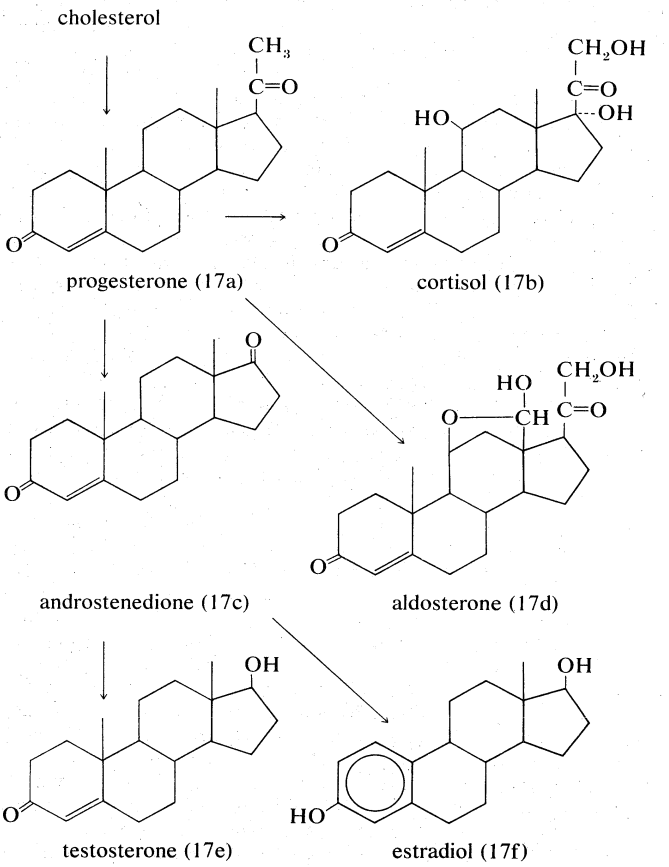
Squalene and lanosterol

Acetate fragments derived from common nutrient materials are converted into mevalonic acid, from which the terpenoid hydrocarbon squalene is formed. One end of the squalene molecule is then oxidized, giving squalene 2,3-oxide, which, by an intramolecular reaction (cyclization) and structural rearrangement, yields the steroid lanosterol. This enzyme-controlled reaction may be initiated by introduction of a positive charge into the oxide ring because it is remarkably similar to the nonenzymic, acid-catalyzed cyclizations of certain unsaturated hydrocarbons similar in structure to squalene. Cholesterol is formed from lanosterol by further structural changes.

The principal forms in which cholesterol is excreted by vertebrates are the bile acids, which are synthesized in the liver. Their formation involves specific modifications of the steroid nucleus and formation of a carboxylic acid group that is linked to the amino acids taurine or glycine to give the forms in which the bile acids are secreted into the bile. The biosynthesis of cholesterol is influenced by feedback mechanisms that suppress the formation of mevalonic acid and, consequently, of cholesterol when levels of cholesterol and bile acids in the tissues are elevated.

Steroid hormones. In vertebrates, cholesterol is the central precursor of all steroid hormones secreted by the testes of the male, the ovaries of the female, and the adrenals of both sexes. These tissues share an embryonic tissue of origin and, in consequence, many enzymes for the transformation of cholesterol. A major (though not exclusive) common pathway involves conversion to progesterone. Progesterone is secreted by the corpus luteum of the ovary, but in the adrenal cortex it is further metabolized to steroid hormones (corticosteroids) such as cortisol and aldosterone. In both ovary and testis, progesterone is transformed further to the androgenic steroid androstenedione, which, together with its derivative testosterone, is secreted by the testis. In the ovary, androstenedione is modified to the estrogen estradiol. Normally, each organ secretes its own characteristic pattern of hormones, but in some disease states (*e.g.*, genetic defects and some tumours of these endocrine glands) these patterns may be profoundly distorted.

Progesterone



biogenetic relationships of principal vertebrate steroid hormones

Many tissues, but mainly the liver, metabolize the steroid hormones to physiologically inactive products that are voided mainly in the urine, though some are also eliminated via the bile and, ultimately, the feces. Diagnosis of endocrine abnormalities may be assisted by analysis of urinary steroids. Urinary 17-ketosteroids (androstane derivatives with a $>C=O$ function at C-17) arise principally through oxidation of adrenal steroid hormones in the liver and thus gauge secretion by the adrenal rather than by the testis. In pregnancy, the urinary excretion of pregnanediol, the principal metabolite of progesterone, measures placental progesterone output. Its decline before term may forewarn of abortion, which may be averted by administration of progestational hormones.

Steroid metabolism in plants. The early steps in the biosynthesis of steroids of both plants and animals are the same, except that in plants lanosterol is replaced by the related compound cycloartenol, which contains a three-membered ring (C-9, C-10, C-19) in lieu of the nuclear double bond of lanosterol. The side chains of the phytosterols, such as stigmasterol, and of the sterol ergosterol of yeasts and fungi contain extra carbon atoms that are incorporated in reactions involving *S*-adenosylmethionine, which donates methyl (CH_3) groups in numerous biological processes. Although most plant tissues contain only traces of cholesterol, this sterol is the biogenetic precursor of such important plant steroids as the sapo-

genins, cardiac glycosides, and steroid alkaloids. Because pregnane derivatives are intermediates in some of these transformations, plants and animals appear to have important features of steroid metabolism in common.

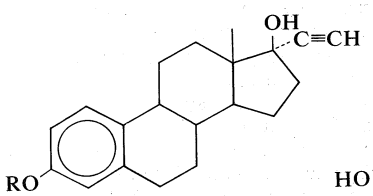
STRUCTURAL RELATIONSHIPS OF THE PRINCIPAL CATEGORIES OF STEROIDS

Sterols. A large group, the sterols, comprises the common 3-monohydroxy steroids of the cholestane, ergostane, and stigmastane series and their "methyl sterol" biogenetic precursors: lanosterol, cycloartenol, and certain derivatives of these sterols, such as lophenol (Table 2). Most sterols have a 3β-hydroxyl group, and many (though not the 4-methyl sterols) have a double bond between carbon atoms 5 and 6. Various sterols have double bonds at other positions in the nucleus corresponding to stages in the biosynthesis of cholesterol and other steroids that resemble it in the structure of ring *A*. Animal sterols, especially in embryonic tissues and skin, and phytosterols (*e.g.*, stigmasterol) also may have a double bond in the side chain. Sterols of feces (*e.g.*, coprostanol) have a 3α-hydroxyl group and *cis*- (5β) linked rings *A* and *B;* they are formed by metabolism of other sterols by intestinal bacteria. Certain sterols are transformed to calciferols (D vitamins) by ultraviolet light; this process occurs naturally in the skin and is used commercially in the manufacture of vitamin D_2 (ergocalciferol) from ergosterol and of vitamin D_3 (cholecalciferol) from synthetic 7-dehydrocholesterol.

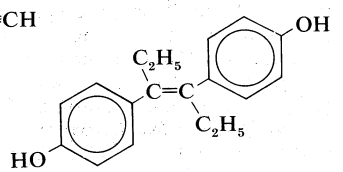
Bile acids and alcohols. The molecular structures of metabolites of cholesterol form an evolutionary series from the bile alcohols, such as myxinol and scymnol of the elasmobranch fishes and the related alcohols of some bony fishes and frogs, through the 5β-cholestanoic acids of crocodiles and alligators, to the 5β-cholanoic acids of the birds and mammals. This evolutionary trend is illustrated by the selected compounds in Table 3. They are not exclusively confined to the species indicated; for example, chenodeoxycholic acid is a major bile acid in man and many other mammals, and cholic acid is found in many nonmammalian species, together with primitive bile acids or alcohols that are not found in the mammals.

Estrogens. The estrogens of the ovary of vertebrates are steroids that are abundant in the urines of pregnant mares and of stallions. The most potent natural estrogen is estradiol; the less potent estrogens—estrone, estriol, and other oxygenated phenolic steroids—are metabolites of estradiol. Some species, notably the Equidae, secrete the less active estrogen equilenin. Estrone, synthesized from diosgenin, has been used as a starting material for synthesis of androgenic and progestational steroids lacking a C-19 methyl group (19-nor steroids). Synthetic estrogens, as estranol or mestranol, commonly used in oral contraceptives and for other therapeutic purposes, have acetylenic (containing triple bonds between carbon atoms) substituents. Non-steroidal synthetic estrogens—*e.g.*, diethylstilbestrol and related compounds—are used clinically and also in animal husbandry to promote fattening of livestock and poultry and to improve the quality of their meat.

Estradiol



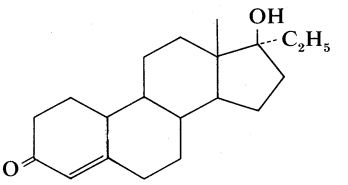
estranol (18a; R = H) and mestranol (18b; R = CH_3)



diethylstilbestrol (19)

Progestational steroids (gestogens, progestins). Three naturally occurring steroids of the corpus luteum and placenta have progestational action; these are progesterone and two of its metabolites. All possess an unsaturated ketonic structure in ring *A*. Pregnanediol, the main metabolite of progesterone, lacks both this structural feature and progestational activity.

Synthetic progestational steroids that are used in oral contraceptives and for other therapeutic purposes (see below) are derivatives of progesterone or of 19-nortestosterone. Among the latter is norethandrolone.



norethandrolone (20)

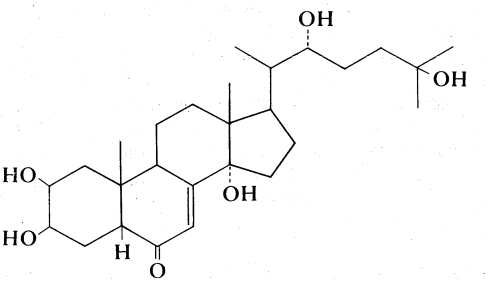
Androgens. Testosterone and androstenedione are the major testicular androgens. Several other less active androgens occur naturally. Major metabolites of testosterone are androsterone and etiocholanolone. The latter compound is androgenically inactive, but it is a pyrogen (*e.g.*, a fever-producing agent) that has been associated clinically with some febrile conditions. Various carboxylic acid esters and a number of synthetic analogues of testosterone are used therapeutically (see below).

Adrenal cortical hormones. The steroids of the adrenal cortex are progesterone derivatives that bear hydroxyl groups at positions 11, 17α, or 21. The potent mineralocortoid aldosterone carries an aldehyde function in place of the more usual C-18 methyl group. Glucocorticoid potency is higher in the trihydroxy derivative cortisol of man, monkeys, and dogs than in the dihydroxy steroid corticosterone of rats and mice. Every adrenal steroid hormone has a ketone group at C-3 and a double bond between C-4 and C-5. In the liver their physiological activities are lost when ring *A* is reduced, and they are partially inactivated when the 11β-hydroxyl is oxidized to a ketone group (as in cortisone and 11-dehydrocorticosterone). The adrenal corticoids are among the most chemically reactive of the steroid hormones; they are sensitive to strong acids and alkalies and to elevated temperatures.

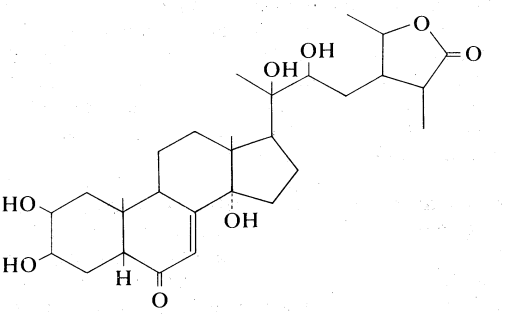
In many synthetic analogues of cortisol and cortisone, physiological activity is modified. Derivatives with enhanced anti-inflammatory activity are especially valuable therapeutically and are discussed below.

Ecdysones. The molting hormones (zooecdysones) of insects and crustaceans are generally derivatives of cholestane (Table 4). All possess a ketone group at position 6, a double bond between positions 7 and 8, and 2β, 3β, and 14α-hydroxyl groups. The side chain is hydroxylat-

Molting hormones



α-ecdysone (21)



cyasterone (22), a phytoecdysone

Table 4: Representative Ecdysones of Arthropods and Plants

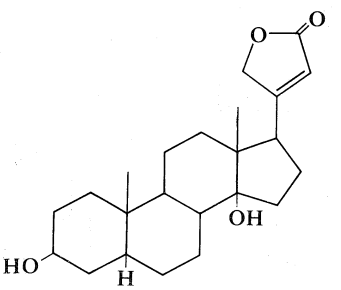
compounds	variant of ecdysone structure	occurrence
C_{27} steroids		
Ecdysone (α-ecdysone)		silkworm, *Bombyx mori*; oak silk moth, *Antheraea pernyi*, and other insects; also the ferns *Pteridium aquilinum* and *Polypodium vulgare* and *Osmunda* species
2β,3β,14α,22,25-pentahydroxy -5β-cholest-7-en-6-one		
Ecdysterone, crustecdysone (β-ecdysone)	20α-hydroxy-	*Bombyx mori*, *Antheraea pernyi* and other insects; crab, *Callinectes sapidus* and crayfish, *Jasus lalandei*; the conifer, *Podocarpus elatus*, the yew, *Taxus baccata* and several ferns
2-Deoxycrustecdysone	2-deoxy-20α-hydroxy-	*Jasus lalandei*
Callinecdysone A	20, 26-dihydroxy-, 25-deoxy-	*Callinectes sapidus* (pre-molting)
C_{28} steroids		
Callinecdysone B	20-hydroxy-, 24-methyl-	*Callinectes sapidus* (post-molting)
Makisterone B	20, 26-dihydroxy-, 24-methyl-25-deoxy-	*Podocarpus macrophyllus*
C_{29} steroids		
Makisterone C	20-hydroxy-, 24-ethyl-	*Podocarpus macrophyllus*
Cyasterone	20, 22-dihydroxystigmasteno-26, 28-lactone side chain	*Cyathula capitata* and several *Ajuga* species

ed at C-22 and variously at C-20, C-25, and C-26. Some of these compounds occur in plants, many of which also contain potent ecdysone analogues (phytoecdysones) with ergostane and stigmastane side chains.

The ecdysones show cross-species activity—that is, the ecdysone of one species induces molting in another species. The preponderance of ecdysones related to cholestane even in phytophagous insects probably reflects their capacity to convert dietary phytosterols to cholestane derivatives.

Cardiac glycosides and aglycones. Many species of plants contain toxic (specifically, heart-arresting) steroids of the cardanolide type as glycosides (compounds that contain structural groups derived from sugars) of up to four sugar residues, which may include glucose, rhamnose, and ten other sugars peculiar to this group of natural products. Typically, these compounds (Table 5) are 5β-steroids and have 3β- and 14β-hydroxyl groups, but hydroxyl groups may occur in many other positions. In all cases, the aglycone (the steroid that results when the sugar groups are removed) is less active than its glycosides, but generally, activity declines with increasing numbers of sugar residues after the first. The structures of the sugars have important but not predictable effects on activity.

The most important cardiac glycosides, medicinally, are those occurring in foxglove (*Digitalis*): digitoxin, gitoxin, and digoxin. Each of these contains a specific aglycone (*e.g.*, digitoxigenin is the aglycone of digitoxin)



digitoxigenin (23)

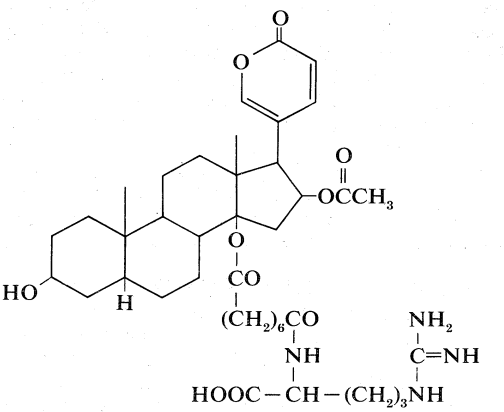
Table 5: Some Cardiac Glycosides and Their Cardenolide Aglycones

glycoside	aglycone	systematic name of aglycone
Digitoxin	digitoxigenin	3β,14β-dihydroxy-5β-card-20(22)-enolide
Gitoxin	gitoxigenin	3β,14β,16β-trihydroxy-5β-card-20(22)-enolide
Digoxin	digoxigenin	3β,12β,14β-trihydroxy-5β-card-20(22)-enolide
Periplocin	periplogenin	3β,5β,14β-trihydroxy-5β-card-20(22)-enolide
Sarmentocymarin	sarmentogenin	3β,11α,14β-trihydroxy-5β-card-20(22)-enolide
k-Strophanthin	strophanthidin	3β,5β,14β-trihydroxy-19-oxo-5β-card-20(22)-enolide
Ouabain	ouabagenin	1β,3β,5β,11α,14β,19-hexahydroxycard-20(22)-enolide

linked to three molecules of the sugar digitoxose and is derived from a more complex glycoside (digilanides A, B, and C, respectively) from which glucose and acetic acid are removed during the isolation procedures.

The squill, or sea onion, *Scilla maritima*, a seashore plant, contains several toxic glycosides of which the aglycones are bufadienolides more typical of the toad poisons than of plant products. (In a bufadienolide, two double bonds are present in the bufanolide side chain.) A white form of the squill produces the glycoside scillaren A, which contains the aglycone scillarenin, while a red form produces scilliroside, which is specifically toxic to rodents and has long been used as a rat poison. The contribution of the side chain to cardiac activity differs little between the bufanolides and the cardanolides.

Toad poisons. Defensive venoms secreted by skin glands (principally the parotid glands) of the toad owe their high toxicity to bufadienolides that occur both free (bufogenins) and combined (bufotoxins). These compounds have digitalis-like properties and have been used medicinally in a traditional Chinese preparation, Ch'an Su. The best characterized is bufotoxin, from the European toad *Bufo vulgaris* and the Oriental toad *Bufo gargarizans*, of which the bufogenin is bufotalin, a close structural relative of gitoxin.



bufotoxin (24)

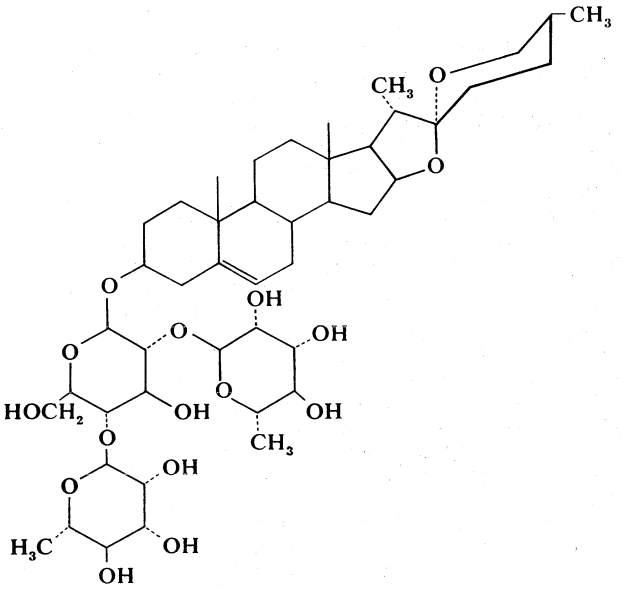
Sapogenins and saponins. Sapogenins are steroids of the spirostan type that occur widely and in great variety in plants. They are linked to sugars as glycosides, usually through a 3β-hydroxyl group. The glycosides are saponins, so called because they form soapy solutions and have other surface active (*e.g.*, hemolytic) properties. Since saponins are difficult to purify, the complete structures of only a few are known. Among these is dioscin (from certain yams, *Dioscorea* species, and yuccas, *Trillium* species); the steroid portion of this saponin is diosgenin.

The nature and number of sugar residues per molecule are known for many saponins. These include the common sugars glucose, xylose, galactose, rhamnose, and arabi-

Table 6: Relative Oral Androgenic and Anabolic Activities of Testosterone and Some Synthetic Analogs

nonproprietary name (trade name)	systematic name	relative potency	
		androgenic*	anabolic†
Testosterone	17β-hydroxy-4-androsten-3-one	100	100
Methyltestosterone (Metandren, Oreton-M)	17β-hydroxy-17-methyl-4-androsten-3-one	200	200
Methandrostenolone (Dianabol)	17β-hydroxy-17-methyl-1, 4-androstadien-3-one	120	400
Norethandrolone (Nilevar)	17α-ethyl-17-hydroxy-4-estren-3-one	40	400
Stanozolol (Winstrol)	17β-hydroxy-17-methyl-5α-androstano(3, 2-*c*)pyrazole	60	600
Oxymetholone (Adroyd, Anadrol)	2-hydroxymethylene-17β-hydroxy-17-methyl-5α-androstan-3-one	100	1,000
Fluoxymesterone (Halotestin, Ultandren)	9α-fluoro-11β, 17-dihydroxy-17α-methyl-4-androsten-3-one	1,500	3,500

*Seminal vesicle weight response. †*Levator ani* weight response.



dioscin (25), a saponin of diosgenin

nose. In most cases, however, only the structure of the sapogenin, which can be released from the saponin by acid hydrolysis, is known with certainty. Linkage of rings *A* and *B* may be *cis*- (5β) or *trans*- (5α) or may involve unsaturation at C-5. A hydroxyl group is nearly always present at position 3, and hydroxyl or ketonic groups may be present at positions 1, 2, 4, 5, 6, 11, 12, or 15. Many pairs of natural sapogenins differ only in configuration at C-25. Their structural features and abundance make diosgenin and hecogenin useful as starting materials for steroid hormone manufacture.

PHARMACOLOGICAL ACTIONS OF STEROIDS

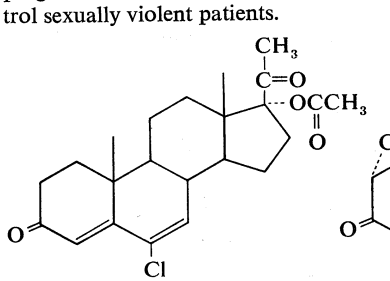
Side effects

Aside from their principal physiological effects, all steroid hormones have generalized influences upon metabolic systems throughout the body. These are sometimes seen as powerful pharmacological side effects when, either during hormone therapy or through some endocrine abnormality, the body is exposed to excessive amounts of a naturally occurring steroid hormone. In some synthetic analogues of the natural hormones, a desired activity is accentuated while others are minimized. Furthermore, just as naturally occurring steroid hormones of differing biological activity (estrogens, androgens, glucocorticoids, mineralocorticoids) often act antagonistically, the many steroid analogues include a number of inhibitors of the natural hormones. Details of the effects of steroids upon living organisms, including pharmacological applications, are more fully covered in other articles (see HORMONE; ENDOCRINE SYSTEM, HUMAN; ENDOCRINE SYSTEM DISEASES AND DISORDERS).

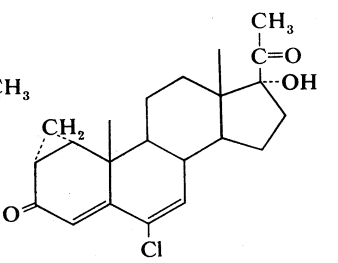
Androgens and anabolic steroids. Androgens secreted or administered in abnormally large amounts can cause development of male characteristics in the female and precocious sexual development in the male. Conversely, hypogonadism of the male (inadequate testicular function) leads to retarded sexual development and retention of feminine bodily characteristics (eunuchoidism), which can sometimes be remedied by administration of androgenic steroids. Several esters of testosterone are commonly used by injection for this purpose. Many orally active analogues of testosterone are also available (Table 6) in which activity is greatly enhanced, and often the ratio of androgenic activity to anabolic activity is shifted markedly in favour of the latter. This ratio primarily determines the therapeutic value of these compounds as anabolic agents. They are used together with growth hormone to promote growth in children in whom growth is retarded. They are also used to promote physical recovery from debilitating diseases. Their alleged use by some athletic competitors in sports has been decried.

The psychic effects of androgens have been exploited in their use to correct frigidity in women and lack of self-esteem in adolescents. Androgens are also used in the treatment of estrogen-dependent tumours, such as breast cancer in women.

Anti-androgens. The estrogens and a number of the synthetic gestogens, such as medroxyprogesterone acetate and chlormadinone acetate (Table 7), have anti-androgenic properties that are the basis for their use against benign or malignant hyperplasia of androgen-dependent tissues such as the prostate. Other anti-androgens are A-nortestosterone and A-norprogesterone and their derivatives. The anti-androgens cyproterone and medroxyprogesterone acetate have been used successfully to control sexually violent patients.



chlormadinone acetate (26)



cyproterone (27)

Synthetic estrogens and gestogens: steroid contraceptives. The most effective currently available method of contraception uses combinations of synthetic estrogens and gestogens that prevent ovulation and render the uterine environment inimical to conception and to the development of the ovum. Unlike the natural hormones, these synthetic steroids are highly active when taken orally, as tablets, and as contraceptives they are popularly referred to as the Pill. Ovulation is suppressed because release of the pituitary hormone that stimulates it is inhibited, primarily by the estrogenic component. The gestogens are also anti-ovulatory, but their principal function in oral contraceptives is to control the uterine bleeding (menstruation) that follows withdrawal of the estrogen. The tablets are taken according to schedules

The Pill

Table 7: Oral Gestogens and Contraceptive Preparations

nonproprietary name (trade name)	systematic name	proprietary name of contraceptive*
17α-Hydroxyprogesterone caproate (Delalutin)	17α-hydroxy-4-pregnene-3, 20-dione caproate	
Medroxyprogesterone acetate (Provera)	17α-hydroxy-6α-methyl-4-pregnene-3, 20-dione acetate	Provest (E)
Megestrol acetate	17α-hydroxy-6-methyl-4, 6-pregnadiene-3, 20-dione acetate	Volidan (E)
Chlormadinone acetate	6-chloro-17α-hydroxy-4, 6-pregnadiene-3, 20-dione acetate	C-Quens (M)†
Dimethisterone	6α, 21-dimethyl-17α-ethynyl-4-androsten-3-one	Oracon (E)†
		Ovin (E)
Norethindrone, norethisterone (Norlutin)	17α-ethynyl-17-hydroxy-4-estren-3-one	Ortho-Novum (M)‡
		Norinyl (M)
		Norquen (M)†
Norethynodrel	17α-ethynyl-17-hydroxy-5(10)-estren-3-one	Enovid, Enovid-E (M)
Lynestrenol	17α-ethynyl-4-estren-17-ol	Lyndiol (M)
Norethindrone acetate	17α-ethynyl-17-hydroxy-4-estren-3-one acetate	Anovlar 21, Norlestrin (E)
Norgestrel	18-methyl-17α-ethynyl-17-hydroxy-4-estren-3-one	Ovral (E)
	17α-ethynyl-17-hydroxy-4, 9(10), 11(12)-estratrien-3-one	Planor (M)
	3-cyclopentoxy-17α-ethynyl-3, 5-estradien-17-ol acetate	Riglovis (E)
Ethynodiol diacetate	17α-ethynyl-4-estrene-3β, 17β-diol diacetate	Ovulen (M)

*The estrogen used is denoted by (E), estranol or (M), mestranol. †Sequential. ‡Sequential or combination.

that simulate a 28-day menstrual cycle. Most frequently, a combination of the two types of steroid is taken daily during 21 or 22 days and is then interrupted for the remainder of the 28-day period. During the nonmedication days, the menstrual flow occurs. An alternative sequential schedule aims to approximate more closely to the endocrine events of the normal menstrual cycle by the use of tablets of estrogen alone for 14–15 days followed by tablets of combined estrogen and gestogen for 5–7 days prior to interruption to permit menstruation. The combination method is somewhat more reliable, probably because the gestogens in the combination regime influence the uterine environment to prevent implantation of the ovum in those few cycles in which ovulation is not inhibited. Other regimes of administration in the investigative stage in the early 1970s are a "once-a-month pill," utilizing the long-acting estrogen quinestrol, and a postcoital ("morning-after") pill.

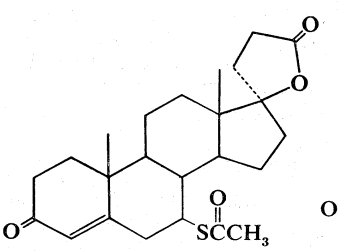
The contraceptive steroids, which were brought into popular use in the early 1960s, have proved remarkably safe. Their discontinuance is usually followed by a prompt return of fertility. Some individuals suffer unpleasant symptoms (nausea, water retention, tiredness) in the use of certain preparations, but a variety of alternative types have become available (Table 7), among which a satisfactory one usually may be found. In oral contraceptives available in the early 1970s, only two estrogens were used: estranol, which has about the same potency as estradiol, and mestranol, which is less potent. The main differences between preparations are the character of the gestogens and the quantities and ratios of the steroids that they contain.

Other uses of estrogens and gestogens

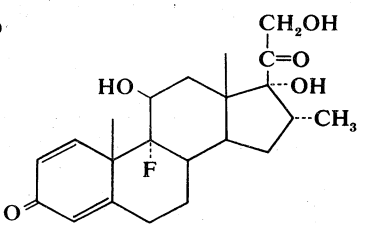
Besides their use in birth control, the contraceptive steroid preparations and synthetic gestogens are used to correct irregularities of the menstrual cycle and its frequent physical and psychological symptoms. The gestogens are used to maintain pregnancy in cases of threatened abortion, some as anti-androgens, and others as anti-estrogens. Contraceptive steroid preparations are also used for control of breeding cycles and synchronization of estrus in farm animals. The synthetic estrogens are used to alleviate the unpleasant physical and psychic symptoms of the menopause in women and for the treatment of some cancers of the prostate in men and of the breast in women. They are also used in some cases to suppress the plasma cholesterol concentrations in sufferers from advanced atherosclerosis, a disease that is markedly more prevalent in men and postmenopausal women than in younger women with intact ovaries.

Anti-estrogens. Androgenic steroids are anti-estrogenic and are often used to suppress the growth of estrogen-dependent cancers. Methyltestosterone and fluoxymesterone (Table 6), and norethindrone (norethisterone) and norgestrel (Table 7) are anti-estrogens. Since estrogens block the release of the pituitary hormone responsible for ovulation, a potent anti-estrogen can stimulate ovulation by inhibiting this action of estrogens.

Adrenal hormones. The anti-inflammatory and glucocorticoid activities of cortisol are enhanced, in some cases with relative reduction of its mineralocorticoid activity, by various structural modifications (Table 8). A 9α-fluorine atom enhances the glucocorticoid activity of cortisol about 10-fold, but its salt-retaining activity about 50-fold. On the other hand, unsaturation at C-1 increases glucocorticoid, but not mineralocorticoid, activity, and 6α-fluorine or methyl and 16-methyl or hydroxyl groups (and especially 16α,17α-acetonides—*i.e.*, compounds formed from 16α,17α-dihydroxy compounds and acetone) enhance anti-inflammatory activity while reducing salt activity. These groupings, therefore, appear in various combinations in anti-inflammatory steroids, many of which, however, lack the salt-retaining activity necessary for total adrenal replacement therapy. Cortisol analogues, such as dexamethasone, are used to treat many inflammatory and rheumatic diseases, to suppress the immune response in allergies and in organ transplantation, and to delay the progress of leukemia. They are also widely used for the treatment of local inflammatory reactions. A synthetic steroid of a quite different type, spironolactone (Aldactone A),



spironolactone (28)



dexamethasone (29)

is used as an antagonist to the action of aldosterone in certain cases of hypertension.

Cardiotonic steroids. Preparations in which cardiotonic steroids of both vegetable and animal origin are the active principles have been used as emetics, diuretics, and arrow poisons since ancient times. The use of digitalis, ouabain, and strophanthin glycosides to slow the rate and strengthen the contractility of the failing heart is one of the most important methods of treatment of this condition. Of these agents, the digitalis glycosides are the most widely used. The therapeutic effects of these agents are related to their influence upon muscle cells, but knowledge of their mechanism of action is incomplete and their use in medicine remains essentially empirical.

Conclusion: The general biological significance of steroids. That such diverse physiological functions and effects should be shown by steroids, all of which are synthesized by essentially the same central biosynthetic pathway, is a remarkable example of biological economy. Most of these functions, especially those of a hormonal type, involve the transmission of biologically essential in-

Table 8: Approximate Relative Anti-inflammatory Potency of Some Analogs of Cortisol

nonproprietary name	modification of cortisol structure	clinically estimated dose (mg) equivalent to 20 mg. cortisol
Cortisone	11-dehydro-	25
Prednisone	1, 11-dehydro-	5
Prednisolone	1-dehydro-	5
Methylprednisolone	1-dehydro-, 6α-methyl-	4
Triamcinolone	1-dehydro-, 9α-fluoro-, 16α-hydroxy-	4
Fluprednisolone	1-dehydro-, 6α-fluoro-	2
Paramethasone	1-dehydro-, 6α-fluoro-, 16α methyl-	2
Triamcinolone acetonide	1-dehydro-, 9α-fluoro-, 16α-hydroxy-, 16, 17-acetonide	1
Dexamethasone	1-dehydro-, 9α-fluoro-, 16α-methyl-	0.75
Betamethasone	1-dehydro-, 9α-fluoro-, 16β-methyl-	0.6
Fluocinolone acetonide	1-dehydro-, 6α, 9α-difluoro-, 16α-hydroxy-, 16, 17-acetonide	0.5
Fluandrenolone acetonide	6α-fluoro-, 16α-hydroxy-, 16, 17-acetonide	0.4

formation. The specific information content of the steroid resides in the character and arrangement of its substituent groups and in other subtle structural modifications. The mechanism of action of no steroid hormone at the molecular level is as yet known, nor is it known why the steroid nucleus is so highly appropriate as a carrier of informational signals in systems of biological control. But that it is, is beyond doubt, and it seems reasonable to expect that many further regulatory functions of steroids, especially in lower life forms, remain to be discovered.

BIBLIOGRAPHY. The chemistry of steroids is comprehensively treated in L.F. and M. FIESER, *Steroids* (1959); more recent information appears in several chapters in S. COFFEY (ed.), *Rodd's Chemistry of Carbon Compounds*, 2nd ed., vol. 2, *Steroids*, 2 pt. (1970–71); for biochemical aspects, see M. FLORKIN and E.H. STOTZ (eds.), *Sterols, Bile Acids and Steroids* (1963); or D.M. GREENBERG, *Metabolic Pathways*, vol. 2 (1968); for comprehensive coverage of the biology and chemistry of the Pill, see D. LEDNICER (ed.), *Contraception: The Chemical Control of Fertility* (1969).

(R.B.C.)

Stevenson, Robert Louis

Though he was best known in his lifetime, and is still popular, as an author of adventure stories and as a personal essayist in the Romantic tradition, Robert Louis Stevenson is a writer of greater variety and subtlety than this would suggest, and his mature fiction includes work of great originality and power.

Early life. Born at Edinburgh on November 13, 1850, Stevenson was the only son of Thomas Stevenson, a prosperous civil engineer, and his wife, Margaret Isabella Balfour. His poor health made regular schooling difficult, but he attended Edinburgh Academy and other schools before, at 17, entering Edinburgh University, where he was expected to prepare himself for the family profession of lighthouse engineering. But Stevenson had no desire to be an engineer, and he eventually agreed with his father, as a compromise, to prepare instead for the Scottish bar.

He had shown a desire to write early in life, and once in his teens he had deliberately set out to learn the writer's craft by imitating a great variety of models in prose and verse. His youthful enthusiasm for the Covenanters (*i.e.*, those Scotsmen who banded together to defend their version of Presbyterianism in the 17th century) led to his writing *The Pentland Rising*, his first printed work. During his years at the university he rebelled against his parents' religion and set himself up as a liberal Bohemian who abhorred the alleged cruelties and hypocrisies of bourgeois respectability.

Friendship with Fanny Sitwell

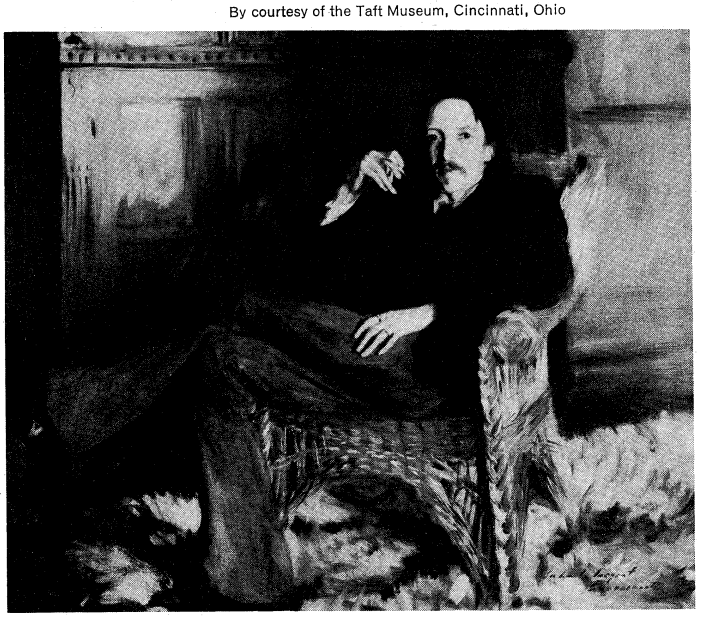
In 1873, in the midst of painful differences with his father, he visited a married cousin in Suffolk, where he met Sidney Colvin, the English scholar, who became a lifelong friend, and Mrs. Fanny Sitwell (who later married Colvin). Mrs. Sitwell, an older woman of charm and talent, drew the young man out and won his confidence. Soon Stevenson was deeply in love, and on his return to Edinburgh he wrote her a series of letters in which he played the part first of lover, then of worshipper, then of son. One of the several names by which Stevenson addressed her in these letters was "Claire," a fact that many years after his death was to give rise to the erroneous notion that Stevenson had had an affair with a humbly born Edinburgh girl of that name. Eventually the passion turned into a lasting friendship.

Later in 1873 Stevenson suffered severe respiratory illness and was sent to the French Riviera, where Colvin later joined him. He returned home the following spring. In July 1875 he was called to the Scottish bar, but he never practiced. Stevenson was frequently abroad, most often in France. Two of his journeys produced *An Inland Voyage* (1878) and *Travels with a Donkey in the Cévennes* (1879). His career as a writer developed slowly. His essay "Roads" appeared in the *Portfolio* in 1873, and in 1874 "Ordered South" appeared in *Macmillan's Magazine*, a review of Lord Lytton's *Fables in Song* appeared in the *Fortnightly*, and his first contribution (on Victor Hugo) appeared in *The Cornhill Magazine*, then edited by Leslie Stephen, a critic and biographer. It was these early essays, carefully wrought, quizzically meditative in tone, and unusual in sensibility, that first drew attention to Stevenson as a writer.

Stephen brought Stevenson into contact with Edmund Gosse, the poet and critic, who become a good friend. Later, when in Edinburgh, Stephen introduced Stevenson to the writer W.E. Henley. The two became warm friends, and were to remain so until 1888, when a letter from Henley to Stevenson containing a deliberately implied accusation of dishonesty against the latter's wife precipitated a quarrel that Henley, jealous and embittered, perpetuated after his friend's death in a venomous review of a biography of Stevenson.

In 1876 Stevenson met Fanny Vandegrift Osbourne, an American lady separated from her husband, and the two fell in love. Stevenson's parents' horror at their son's involvement with a married woman subsided somewhat when Mrs. Osbourne returned to California in 1878, but it revived with greater force when Stevenson decided to join

By courtesy of the Taft Museum, Cincinnati, Ohio



Robert Louis Stevenson, portrait by John Singer Sargent, 1887. In the Taft Museum, Cincinnati, Ohio.

her in August 1879. Stevenson reached California ill and penniless (the record of his arduous journey appeared later in *The Amateur Emigrant*, 1895, and *Across the Plains*, 1892). His adventures, which included coming very near death and eking out a precarious living in Monterey and San Francisco, culminated in marriage to Fanny Osbourne (who was by then divorced from her first husband) early in 1880. About the same time a telegram from his relenting father offered much-needed financial support, and after a honeymoon by an abandoned silver mine (recorded in *The Silverado Squatters*, 1883) the couple sailed for Scotland to achieve reconcilation with the Thomas Stevensons.

Marriage

Romantic novels. Soon after his return, Stevenson, accompanied by his wife and his stepson, Lloyd Osbourne, went, on medical advice (he had tuberculosis), to Davos, Switzerland. The family left there in April 1881 and spent the summer in Pitlochry and then in Braemar, Scotland. There, in spite of bouts of illness, Stevenson embarked on *Treasure Island* (begun as a game with Lloyd), which started as a serial in *Young Folks*, under the title *The Sea-Cook*, in October 1881. Stevenson finished the story in Davos, to which he had returned in the autumn, and then started on *Prince Otto* (1885), a more complex but less successful work. *Treasure Island* is an adventure presented with consummate skill, with atmosphere, character, and action superbly geared to one another. The book is at once a gripping adventure tale for boys and a wry comment on the ambiguity of human motives.

In 1881 Stevenson published *Virginibus Puerisque*, his first collection of essays, most of which had appeared in *The Cornhill*. The winter of 1881 he spent at a chalet in Davos. In April 1882, he left Davos; but a stay in the Scottish Highlands, while it resulted in two of his finest short stories, "Thrawn Janet" and "The Merry Men," produced lung hemorrhages, and in September he went to the south of France. There the Stevensons finally settled at a house in Hyères, where, in spite of intermittent illness, Stevenson was happy and worked well. He revised *Prince Otto*, worked on *A Child's Garden of Verses* (first called *Penny Whistles*), and began *The Black Arrow: A Tale of the Two Roses* (1888), a historical adventure tale deliberately written in anachronistic language. The threat of a cholera epidemic drove the Stevensons from Hyères back to England. They lived at Bournemouth from September 1884 until July 1887, but his frequent bouts of dangerous illness proved conclusively that the British climate, even in the south of England, was not for him. The Bournemouth years were fruitful, however. There he got to know and love Henry James. There he revised *A Child's Garden* (first published in 1885) and wrote "Markheim," *Kidnapped* (1886), and *The Strange Case of Dr. Jekyll and Mr. Hyde* (1886). The poems in *A Child's Garden* represent with extraordinary fidelity an adult's recapturing of the emotions and sensations of childhood; there is nothing quite like them in English literature. In *Kidnapped* the fruit of his researches into 18th-century Scottish history and of his feeling for Scottish landscape, history, character, and local atmosphere mutually illuminate one another. But it was *Dr. Jekyll* —both moral allegory and thriller—that established his reputation with the ordinary reader.

American sojourn

In August 1887, still in search of health, Stevenson set out for America with his wife, mother, and stepson. On arriving in New York, he found himself famous, with editors and publishers offering lucrative contracts. He stayed for a while in the Adirondack Mountains, where he wrote essays for *Scribner's* and began *The Master of Ballantrae* (1889). This novel, another exploration of moral ambiguities, contains some of his most impressive writing, although marred by its contrived conclusion.

Life in the South Seas. In June 1888 Stevenson, accompanied by his family, sailed from San Francisco in the schooner yacht "Casco," which he had chartered, on what was intended to be an excursion for health and pleasure. In fact, he was to spend the rest of his life in the South Seas. They went first to the Marquesas Islands, then to Fakarava Atoll, then to Tahiti, then to Honolulu, where they stayed nearly six months, leaving in June 1889 for the Gilbert Islands, and then to Samoa, where he spent six weeks.

Journalist in the South Pacific

During his months of wandering around the South Sea islands, Stevenson made intensive efforts to understand the local scene and the inhabitants. As a result, his writings on the South Seas (*In the South Seas*, 1896; *A Footnote to History*, 1892) are admirably pungent and perceptive. He was writing first-rate journalism, deepened by the awareness of landscape and atmosphere, such as that so notably rendered in his description of the first landfall at Nuku Hiva in the Marquesas.

In October 1890 he returned to Samoa from a voyage to Sydney and established himself and his family in patriarchal status at Vailima, his house in Samoa. The climate suited him; he led an industrious and active life; and when he died suddenly on December 3, 1894, it was of a cerebral hemorrhage, not of the long-feared tuberculosis. His work during those years was moving toward a new maturity. While *Catriona* (U.S. title, *David Balfour*, 1893) marked no advance in technique or imaginative scope on *Kidnapped*, to which it is a sequel, *The Ebb-Tide* (1894), a grim and powerful tale written in a dispassionate style (it was a complete reworking of a first draft by Lloyd Osbourne), showed that Stevenson had reached an important transition in his literary career. The next phase was demonstrated triumphantly in *Weir of Hermiston* (1896), the unfinished masterpiece on which he was working on the day of his death. "The Beach of Falesá" (first published 1892; included in *Island Night's Entertainments*, 1893), a story with a finely wrought tragic texture, as well as the first part of *The Master of Ballantrae*, pointed in this direction, but neither approaches *Weir*. Stevenson achieved in this work a remarkable richness of tragic texture in a style stripped of all superfluities. The dialogue contains some of the best Scots prose in modern literature. Fragment though it is, *Weir of Hermiston* stands as a great work and Stevenson's masterpiece.

Assessment. Stevenson was an indefatigable letter writer, and his letters (edited by Sidney Colvin in 1899) provide a lively and enchanting picture of the man and his life. But Colvin omitted many of the most interesting letters and compressed and dovetailed others, with the result that many important facts about Stevenson's emotional life remained unknown until the true text of all the letters was available. Colvin presented Stevenson's letters to Mrs. Sitwell to what is now the National Library of Scotland with the proviso that they were not to be opened until 1949; the revealing and often fascinating letters to Charles Baxter, a friend, are deposited in the Yale University Library. Stevenson's biography suffered from his being early canonized; later writers built up a counterpicture of an immoral swaggerer restrained into reluctant respectability by a jealous wife. Access to the crucial letters yielded a picture of a Stevenson who was neither the "seraph in chocolate" against whom Henley protested nor a low-living rake nor an optimistic escapist nor the happy invalid but a sensitive and intelligent writer who had no illusions about life and wryly made the best of a world to which he did not profess to have the key.

Literary reputation

His literary reputation has also fluctuated. The reaction against him set in soon after his death: he was considered a mannered and imitative essayist or only a writer of children's books. But eventually the pendulum began to swing the other way, and by the 1950s his reputation was established among the more discerning as a writer of originality and power; whose essays at their best are cogent and perceptive renderings of aspects of the human condition; whose novels are either brilliant adventure stories with subtle moral overtones or original and impressive presentations of human action in terms of history and topography as well as psychology; whose short stories produce some new and effective permutations in the relation between romance and irony or manage to combine horror and suspense with moral diagnosis; whose poems, though not showing the highest poetic genius, are often skillful, occasionally (in his use of Scots, for example) interesting and original, and sometimes (in *A Child's Garden*) valuable for their exhibition of a special kind of sensibility.

MAJOR WORKS

TALES AND NOVELS: *New Arabian Nights*, 2 vol. (1882; the stories in vol. 1 appeared as *Latter-Day Arabian Nights*, 1878; those in vol. 2 had appeared in *The Cornhill* and other magazines); *Treasure Island* (1883; serialized in a slightly different form in *Young Folks*, 1881–82); *More New Arabian Nights*, with Fanny Van de Grift Stevenson (1885); *Prince Otto* (1885); *Strange Case of Dr. Jekyll and Mr. Hyde* (1886); *Kidnapped* (1886; in *Young Folks*, 1886); *The Merry Men and Other Tales and Fables* (1887), including "Thrawn Janet" and "Olalla"; *The Black Arrow: A Tale of the Two Roses* (1888; in *Young Folks*, 1883); *The Master of Ballantrae* (1889; serialized in *Scribner's Magazine*, 1888–89); *Catriona*, (1893), a sequel to *Kidnapped; The Ebb-Tide*, with Lloyd Osbourne (1894; serialized in *To-Day*, 1893–94); *Weir of Hermiston* (unfinished 1896); *St. Ives*, completed by Sir Arthur Quiller-Couch, pseudonym "Q" (1897).

ESSAYS AND MISCELLANEOUS: *An Inland Voyage* (1878); *Travels with a Donkey in the Cévennes* (1879); *Virginibus Puerisque* (1881), collected essays, mainly from *The Cornhill Magazine; Familiar Studies of Men and Books* (1882); *Memories and Portraits* (1887), 16 essays; *Across the Plains* (1892), 12 essays; *Vailima Letters*, to Sidney Colvin (1895); *From Scotland to Silverado*, ed. by J.D. Hart (1966), brings together all Stevenson's previously published and unpublished writings about his trip to California in 1879–80.

POETRY: *A Child's Garden of Verses* (1885); *Underwoods* (1887), 38 poems in English, 16 in Scots; *Ballads* (1890); *Songs of Travel and Other Verses* (1896). *Collected Poems*, ed. by J. Adam Smith, 2nd ed. (1971), is the standard edition of Stevenson's poetry.

BIBLIOGRAPHY. There is a bibliography of works by and about Stevenson in *The New Cambridge Bibliography of English Literature*, vol. 3 (1969). See also WILLIAM F. PRIDEAUX, *A Bibliography of the Works of Robert Louis Stevenson*, rev. by F.V. LIVINGSTON (1917); and GEORGE L. MCKAY (comp.), *A Stevenson Library . . . , Formed by Edwin J. Beinecke*, 6 vol. (1951–64). Of the many collected editions of Stevenson's works the following are notable: Pentland edition, ed. by EDMUND GOSSE, 20 vol. (1906–07); Swanston edition, 25 vol. (1911–12); Vailima edition, ed. by LLOYD OSBOURNE, with prefatory notes by FANNY VAN DE GRIFT STEVENSON, 26 vol. (1922–23); Skerryvore edition, 30 vol. (1924–26); Tusitala edition, 35 vol. (1923–24); South Seas edition, 32 vol. (1925), JAMES D. HART (ed.), *From Scotland to Silverado* (1966), brings together all of Stevenson's previously published and unpublished writings about his trip to California in 1879–80, with a biographical introduction. *The Letters of Robert Louis Stevenson to His Family and Friends* were edited by SIDNEY COLVIN (2 vol., 1899; 4 vol., 1911), but these are far from complete, and the text has sometimes been tampered with. Stevenson's lively and important correspondence with his lifelong friend Charles Baxter, *RLS: Stevenson's Letters to Charles Baxter*, was edited by DE LANCEY FERGUSON and MARSHALL WAINGROW (1956).

The official biography of Stevenson is by his cousin GRAHAM BALFOUR, *The Life of Robert Louis Stevenson*, 2 vol. (1901, reprinted 1968). This is valuable, but there are serious omissions and evasions. The standard modern biography is JOSEPH C. FURNAS, *Voyage to Windward: The Life of Robert Louis Stevenson* (1951); MALCOLM ELWIN, *The Strange Case of Robert Louis Stevenson* (1950, reprinted 1971); and RICHARD ALDINGTON, *Portrait of a Rebel* (1957), are lively retellings of the life. *Henry James and Robert Louis Stevenson: A Record of Friendship and Criticism*, ed. by J. ADAM SMITH (1948), gives an account of the relationship between the two writers and prints both their criticisms of each other's work and their letters to each other. Modern critical works on Stevenson are DAVID DAICHES, *Robert Louis Stevenson* (1946); LETTICE COOPER, *Robert Louis Stevenson* (1947); ROBERT KIELY, *Robert Louis Stevenson and the Fiction of Adventure* (1964); and EDWIN M. EIGNER, *Robert Louis Stevenson and Romantic Tradition* (1966).

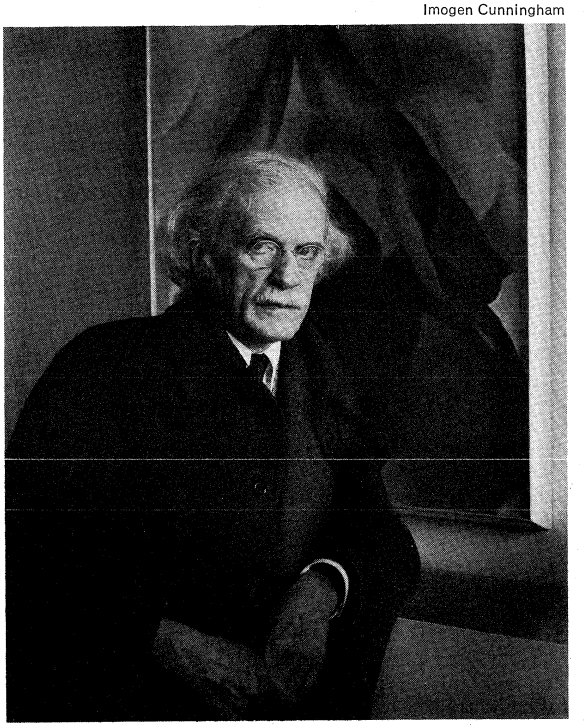
(D.Da.)

Stieglitz, Alfred

Often called the father of modern photography, Alfred Stieglitz was a slight, wiry man of intense passion, incredible energies, and ardent convictions. He devoted his life first to achieving the recognition of photography as an art and then to the introduction of modern art to the United States. As a renowned photographer, Stieglitz himself broke down the barriers against photography in American art museums. His prints were the first photographs accepted as art and received as such by major museums in Boston, New York, and Washington. In those museums his photographs were hung and exhibited in the same manner as other notable achievements in the graphic arts.

A born leader, Stieglitz had an infallible eye, a flexible and inquiring mind, and a temperament conducive to battling for his revolutionary beliefs. Almost single-handedly, he launched his country into the 20th-century art world, not only in the field of photography but in painting and sculpture as well. He awakened the public by exposing it, for the first time, to the best of modern art. With camera and pen, with fiery harangues in his own galleries, he goaded Americans away from their complacent artistic tradition, which largely had been confined to the imitation of European academicians since the 19th century, and worked for the recognition of the contemporary artist in the United States as a force in the international art world.

Imogen Cunningham



Stieglitz at his gallery "291" in 1934. Behind him is a painting by his wife, Georgia O'Keeffe.

Early years in America and Europe

Born in Hoboken, New Jersey, on January 1, 1864, Stieglitz was the oldest son of Edward Stieglitz, a retired woollen merchant. After his early schooling in New York, the Stieglitz family moved to Europe in 1881 to further their children's education. Stieglitz eventually registered for studies in mechanical engineering at the Berlin Polytechnic in 1883. A few months later, the purchase of a small camera led him to abandon engineering for photochemistry and to begin his photographic career.

While still a student at Berlin, where many of his friends were painters, Stieglitz decided to fight for the recognition of photography as a creative art medium equal to that of painting. The best way to gain this acceptance, he innocently believed, was to become a photographic authority, and to become such an authority, he believed he had only to set the highest standards for his own prints and to win all possible prizes and medals.

Both in Europe and the United States, where he returned to live in 1890, his early work (*c.* 1883–1910) reflects these aims. It was characterized by constant technical innovations, which at the time were believed impossible to achieve. Before the turn of the century, for example, Stieglitz had made the first successful photographs in snow, in rain, and at night and had undertaken the first serious use of a small, hand-held camera. By 1910 these photographs had won more than 150 important prizes.

The Photo-Secession group of photographers

Realizing that his own early fame could not alone bring about the recognition of photography as an art, Stieglitz believed that the work of a group might be more effective than that of an individual. After several ventures with established photographic organizations, in which Stieglitz usually led dissident factions, in 1902 he founded a new group of his own, the Photo-Secession, a title adapted from the German Secessionist painters who at that time also were in revolt against the traditional art world. As its leader, Stieglitz gathered around him a group of talented American photographers dedicated to his ideals.

Until 1905, however, the Photo-Secession had not had a centre to exhibit its work in. That year, largely at the urging of Edward Steichen, a close friend and associate, Stieglitz finally agreed that the group should have its own space. Initially called the Little Galleries of the Photo-Secession, located in New York City, the group soon became known by the gallery's street number ("291"); and Stieglitz himself was so closely identified with its work that he often signed his personal correspondence "291."

Soon after its establishment Stieglitz turned his energies to the cause of modern art. At first he did not claim to understand the modern movements, but he did consider them to be fascinating new ways of expressing reality.

The famous Armory Show of 1913 is often considered to have introduced modern art to the United States. Stieglitz in 1908, however, had already held the first showings of work by the sculptor Auguste Rodin and the painter Henri Matisse, and he continued to hold exhibitions of the painters Toulouse-Lautrec, Henri Rousseau, Paul Cézanne, and Pablo Picasso during the five years before the Armory Show. Even more revolutionary was Stieglitz' fervent belief that living American artists were of equal importance and could create significant works of art in their own lifetime. Thus, during the same years that he was bringing modern European art to the United States, he also was holding exhibitions of the work of American painters John Marin, Marsden Hartley, Arthur Dove, Alfred Maurer, and Max Weber. By the time "291" closed in 1917, Stieglitz also had given Americans their first shows of works by the sculptors Constantin Brancusi and Elie Nadelman and the painters Francis Picabia, Gino Severini, Stanton Macdonald-Wright, and Georgia O'Keeffe, who was to become his wife in 1924. For the most part these exhibitions were viewed by a hostile and derisive public. The history of these stormy years is documented in *Camera Work* (1903–17), the radical and beautiful magazine edited and published by Stieglitz.

Neither "291" nor *Camera Work* survived the difficulties of the World War I years, both coming to an end in 1917. With the closing of the gallery, Stieglitz once more had time for his own photography, neglected during the years of "291." His interests turned from his youthful preoccupation with technique and a wide subject matter to an intense preoccupation with what he saw and lived with daily—either at his summer home in Lake George, New York, or at his New York City residence. Two series of this period (1917–37), each consisting of more than 400 prints, are often considered his greatest photographs: the portraits of Georgia O'Keeffe, and the equivalents—photographs of abstract cloud shapes corresponding to all possible emotional experiences. In both he continued to extend the frontiers of photography.

The equivalents and the portraits

Stieglitz' preoccupation with his photography did not deter him from continuing to hold shows of American painters, thus helping to provide them a living and the freedom to work as they wished. Since by that time the once-ridiculed European artists had become a commercial success, he ceased working for them. (He was not a dealer and never profited financially from his activities for artists.) Fanatic until the end in his belief that significant art existed in the United States, his steadfast interest encouraged a group of American artists (principally Charles Demuth, Dove, Hartley, Marin, and O'Keeffe) to achieve this vision. His faith was demonstrated in more than 175 exhibitions held in his three galleries (1905–46), first at "291" (1905–17), then at the Intimate Gallery (1925–29), and finally at An American Place (1929–46). By the end of his life he had changed the climate for all American painting, gaining for it a truly independent status and acceptance.

Shortly after the end of the 1946 season at An American Place, Stieglitz, preparing for his yearly summer trip to Lake George, suffered a stroke and died on July 13, 1946, in New York.

BIBLIOGRAPHY. The most complete set of Stieglitz photographs is held by the National Gallery of Art, Washington, D.C. Other important collections are those of the Metropolitan Museum of Art and the Museum of Modern Art in New York; the Museum of Fine Arts, Boston; the Art Institute of Chicago; and the San Francisco Museum of Art. The main portions of the Alfred Stieglitz collection of paintings are at the Metropolitan Museum of Art and the Art Institute of Chicago. Works on Stieglitz include: DORIS BRY, *Alfred Stieglitz: Photographer* (1965); FRANK WALDO *et al.* (eds.), *America and Alfred Stieglitz: A Collective Portrait* (1934); DOROTHY NORMAN, *Alfred Stieglitz* (1960); and HERBERT J. SELIGMANN, *Alfred Stieglitz Talking* (1966).

(D.Br.)

Stilicho

Perhaps the most energetic, ambitious, and competent of military rulers as the Roman Empire ended in the West was Flavius Stilicho—half-Roman, half-Vandal by birth—who was regent in the West for the young and weak-willed Honorius, son of the emperor Theodosius I. Stilicho effectively ruled the Western empire from AD 395 until his death in 408 and was able to halt the advance of the Visigoth invaders. His rivalry with Flavius Rufinus, his counterpart in the East, to which Theodosius' son Arcadius had succeeded, signalled the complete and formal division of the empire into its eastern and western halves.

Alinari



Stilicho with wife and son, ebony diptych, *c.* 400. In the Cathedral Treasury, Monza, Italy.

Stilicho was born *c.* AD 365, the son of a mercenary captain. Making the army his career, in 383 he served on an embassy to the Persian king Shāpūr III, afterward marrying Serena, the favourite niece of Theodosius. He was appointed count of the domestics (commanding the Emperor's household troops) *c.* 385 and then, in or before 393, master of both services (*i.e.*, commander in chief of the army), an office he held until his death.

Little is known of Stilicho's military exploits before 395. He became the enemy of Rufinus because of a difference of opinion about the treatment of some barbarian invaders in 389. Before his death Theodosius appointed Stilicho and Rufinus to be the guardians of his sons. Stilicho had a great military advantage over his rival, for the army Theodosius had assembled to crush the usurper Eugenius was still concentrated in the West under Stilicho's command when Theodosius died.

Before either Stilicho or Rufinus could attack the other,

the Visigoths, a Germanic tribe living in Lower Moesia who had suffered severe losses in the campaign against Eugenius, rebelled under the leadership of their chieftain Alaric and began to devastate Thrace and Macedonia. Stilicho, who claimed for the Western government the prefecture of Illyricum, which was disputed between the emperors, went with his army to Thessaly. About to engage Alaric there, however, he was ordered by Arcadius, acting on Rufinus' advice, to send a number of his troops to Constantinople. Stilicho obeyed and thus enabled Alaric to penetrate into Greece; but the troops sent to Constantinople murdered Rufinus there on November 27, 395. The gifted poet Claudian began to publish his poems in praise of Stilicho, an important source of information about the politics of the time until 404, when the poet apparently died.

Conflict with Rufinus

In 397 Stilicho took another army to Greece but failed again to bring Alaric to battle and withdrew to Italy. In that same year, Gildo, the count of Africa, rebelled against the Roman government and refused to allow African grain ships to sail to Rome. Stilicho promptly imported grain from Gaul and Spain. In the following year he sent Gildo's brother, Mascezel, to Africa with an army and he easily overthrew Gildo and put him to death; but Mascezel died soon afterward and Stilicho was suspected of having had him murdered so that he might not become a rival. In 398 Stilicho's daughter Maria was married to the emperor Honorius, and Stilicho himself became consul in 400.

The Visigothic invasion

In 401 when Alaric and the Visigoths invaded Italy and threatened Milan, where Honorius was living, Stilicho summoned troops from the Rhine frontier and from Britain to strengthen his army; and in 402 he engaged the Goths at Pollentia (Pollenzo, near modern Bra). The battle, fought on Easter Sunday, April 6, resulted in an incomplete victory for Stilicho, but he captured the Gothic camp. Alaric then marched toward Etruria, but after negotiations he agreed to withdraw from Italy. In 403, however, he re-entered Italy and attacked Verona. Stilicho once more engaged him successfully and pursued him northward. Again an understanding was reached, and Alaric, though his army was in bad condition, was allowed to escape.

Late in 405 Italy was menaced by new invaders, a vast host of Germans, mainly Ostrogoths, led by a pagan called Radagaisus. Contemporary accounts numbered them in hundreds of thousands, though any such figure is impossible. They attacked Florence, but Stilicho compelled their withdrawal to Fiesole, where he cut off their supplies and massacred them. Radagaisus was executed on August 23, 406, and, in celebration, a triumphal arch was erected at Rome.

Attempt to gain Illyricum

Stilicho had never abandoned his plans to annex Illyricum, and in 407 he put them into operation. He closed the ports of Italy to all Eastern ships, instructed Alaric to hold Epirus for Honorius, and himself prepared to cross the Adriatic. But a false report reached him that Alaric was dead, and then he heard of the revolt of Constantine in Britain. Once more his plans had to be abandoned. Alaric then marched to the province of Noricum (roughly central Austria) and demanded compensation for the trouble he had been put to. His price was high—4,000 pounds (in weight) of gold—but despite considerable opposition Stilicho persuaded the Roman Senate to pay it.

By this time the empress Maria had died, but early in 408 Honorius married another daughter of Stilicho, Thermantia. Stilicho's influence, however, had declined. It was rumoured that he wished to have his son Eucherius elevated to the throne. Reports reached him early in 408 that his army was disaffected. Then came news of the death of the Eastern emperor, Arcadius, and Stilicho proposed to go to Constantinople. A certain Olympius, a palace official, spread the rumour that Stilicho was preparing to put his own son on the Eastern throne, and so the troops in Tictinum (Pavia) killed nearly all of the officials present on August 13. Stilicho went to Ravenna but was imprisoned by Honorius' orders. He was beheaded on August 22; Eucherius was put to death shortly thereafter.

BIBLIOGRAPHY. There is no recent biography of Stilicho; however, for information about him, J.B. BURY, *A History of the Later Roman Empire*, vol. 1 (1923), is adequate. The most important ancient source is the poet Claudian; ALAN CAMERON, *Claudian: Poetry and Propaganda at the Court of Honorius* (1970), is fundamental on Stilicho as well as Claudian.

(E.A.T.)

Stimulant

In general terms, a stimulant may be defined as a drug that excites some bodily function. Defined in this way, however, stimulants include half of all known drugs, ranging from cathartics, which stimulate the churnings of the large intestine; to psychedelics, which stimulate self-awareness; and hormones, such as insulin which stimulates the entry of glucose into cells. A more restricted but generally accepted definition refers to stimulants as agents that excite the central nervous system. Stimulation of the central nervous system frequently is defined as a mental state associated with heightened alertness and elevation of mood. Although this article is mainly concerned with stimulants of the central nervous system, drugs that stimulate functions of the peripheral nervous system also are considered briefly. Central nervous system stimulants may be classified as one of two types. Stimulants of one type affect alertness and have only secondary actions on emotions; stimulants of the second type, commonly called antidepressants, affect the emotions. Minor tranquillizers and stimulants (*e.g.*, amphetamine) are among the most commonly prescribed drugs. Methylxanthines, the major stimulant ingredients in coffee, tea, cocoa, and cola beverages, may be the most widely consumed drugs.

General features. *Nerve functions affected by stimulants.* In order to understand the mechanism of stimulant action, it is necessary to know the brain functions that are affected by stimulants. Information processing throughout the nervous system is primarily the function of the specialized nerve cells called neurons. A neuron has a long cell projection called an axon, which terminates in a nerve ending. Nervous function depends on the ability of a neuron to send an electrochemical current in one direction along its axon to the nerve ending. The nerve ending is in contact with other neurons either at the cell body or at dendrites originating at the cell body. A large number of nerve endings that may contact other neurons leave each axon, and cell bodies and dendrites are in contact with nerve endings from many different neurons. Thus the brain is a vast network of neuronal connections, and an almost infinite number of interactions among different neurons is possible. The purpose of nerve function is to convey information from the nerve ending of one neuron across a gap to the cell body or dendrites of another neuron. The gap is called a synaptic cleft, and the junction between two neurons is called a synapse.

Action of neurotransmitters

At most synapses the electrically excited state of a nerve ending stimulates the release of a chemical or neurotransmitter that diffuses across the synaptic cleft to another neuron. Different systems of neurons in the brain utilize different neurotransmitters. The chemical compositions of about five brain neurotransmitters are known; there may be others in addition to these.

Nervous function in the peripheral nervous system is essentially the same as that in the central nervous system (*i.e.*, brain and spinal cord). The peripheral nervous system includes nerves that innervate voluntary muscles (*e.g.*, those that move arms and legs) and those that innervate involuntary muscles (*e.g.*, the heart) and glands. In contrast to the central nervous system, in which at least five neurotransmitter substances are known, only two, the chemicals acetylcholine and norepinephrine, are known in the peripheral nervous system. Acetylcholine is a neurotransmitter both at the junctions of nerves with voluntary muscles and also at the junctions of some involuntary or autonomic nerves. The autonomic nerve endings that release acetylcholine are concerned with body functions associated with a calm state; *e.g.*,

acid secretion in the stomach, decrease in the heart rate, bowel function, constriction of the pupils of the eyes, dilation of blood vessels, and urination. These nerves are called collectively the parasympathetic nervous system. Norepinephrine, on the other hand, is the neurotransmitter for autonomic nerves whose stimulation produces a state of alarm; *e.g.*, increase in heart rate, constriction of blood vessels with an increase in blood pressure, dilation of the pupils and the bronchioles in the lung, and inhibition of bowel and urinary bladder function. These nerves are called collectively the sympathetic nervous system.

Mode of stimulant action. Drugs that either facilitate or antagonize the actions of acetylcholine and norepinephrine in parasympathetic and sympathetic nerves are widely used in medicine. The norepinephrine released by the stimulation of sympathetic nerves dilates the bronchioles in the lungs; therefore, drugs similar in structure to and capable of mimicking or facilitating the action of norepinephrine are valuable for treating bronchial asthma, a condition in which the bronchioles are constricted. Among the drugs used for this purpose are epinephrine, ephedrine, pseudoephedrine (Sudafed), and isoproterenol. Norepinephrine itself is one of the best known treatments for shock, the symptoms of which are vascular collapse accompaned by very low blood pressure; norepinephrine-like drugs used in treating shock include metaraminol and ephedrine.

Drugs may also block the actions of neurotransmitters; atropine, for example, antagonizes acetylcholine at parasympathetic nerve endings. The effects of atropine include dilation of the pupils, inhibition of acid secretion in the stomach, delayed urination, and constipation. Atropine and related compounds (*e.g.*, tincture of belladonna, scopolamine, banthine, and pro-banthine), therefore, are used to relieve symptoms of peptic ulcer and diarrhea.

After their release, neurotransmitter substances are normally inactivated so that the muscle, the gland, or the neuron can recover and receive another stimulation. A drug may act to enhance neurotransmitter effects by blocking neurotransmitter inactivation; acetylcholine, for example, is inactivated by enzymatic breakdown catalyzed by acetylcholinesterase. Drugs that inhibit acetylcholinesterase (*e.g.*, edrophonium chloride, physostigmine, and neostigmine) enhance the effects of acetylcholine at voluntary muscles and are used to treat diseases involving muscular weakness (*e.g.*, myasthenia gravis).

Effects on emotions

In the brain, norepinephrine is localized in neuronal systems concerned with emotional behaviour. Stimulation of these systems produces effects (alertness and euphoria, a feeling of intense well-being) analogous to the alarm effects associated with norepinephrine release in the sympathetic nervous system. In contrast to the enzymatic inactivation of acetylcholine, norepinephrine, after its release from a nerve ending, is inactivated by being pumped back into the same nerve ending by a specialized transport system. Drugs that inhibit this pumping process potentiate the effects (alertness and euphoria) of norepinephrine and are called euphoriants or stimulants. Both amphetamine and the so-called tricyclic class of antidepressant drugs inhibit the norepinephrine pumping process, thus potentiating its effects after it is released from nerve endings.

Although norepinephrine can be broken down by at least two enzymes, it is not likely that enzymatic breakdown plays a major role in synaptic inactivation. But if the activity of these enzymes is inhibited, the concentration of norepinephrine at nerve endings increases; thus, a drug that inhibits the enzymatic breakdown of norepinephrine stimulates the central nervous system. The enzyme monoamine oxidase, for example, found in nerve endings containing norepinephrine and in many other sites in the body (*e.g.*, the liver), catalyzes a reaction in which the synaptic activity of norepinephrine is lost. Drugs that inhibit the action of monoamine oxidase also inhibit the breakdown of norepinephrine, resulting in its accumulation at nerve endings; such drugs may be used as antidepressants.

Drugs also stimulate the central nervous system by releasing norepinephrine directly from nerve endings or by mimicking the action of norepinephrine at receptor sites. Amphetamine, for example, releases norepinephrine and may also mimic it at receptors. Amphetamine may stimulate the central nervous system in one of three ways: by inhibiting the pumping process that inactivates norepinephrine, by causing direct release of norepinephrine at nerve endings, or by activating norepinephrine receptors.

Absorption, metabolism, and excretion. Monoamine-oxidase inhibitors are readily absorbed when given by mouth. Although present in the body only a short time, the drugs have long-lasting effects because they irreversibly inactivate monoamine oxidase. Termination of an inhibitor drug's effects occurs after several weeks, when the enzyme again is formed. Amphetamine and its derivatives are effective when given by mouth, may act for long periods, and may be found in high concentrations in the brain and in cerebrospinal fluid. Since many of these drugs are resistant to inactivation by liver enzymes, substantial amounts are excreted unchanged in the urine. Methylxanthines are absorbed readily after administration orally, rectally, or by injection. The rate of absorption depends on the specific compound—caffeine, theophylline, or theobromine—and the kind of administration used. The methylxanthines, which may be partially metabolized in the body, are excreted largely as methyluric acids or methylxanthines (about 10 percent).

Antidepressant drugs. Since their introduction about 1957, antidepressant drugs, also called psychic energizers, have had a significant influence in the treatment of depressed patients. There are two types: monoamine-oxidase inhibitor drugs and tricyclic antidepressant drugs. An important feature of antidepressant drugs is the delay (up to four weeks) in the onset of their action in patients receiving them.

Monoamine-oxidase inhibitor drugs. Discovered by accident, the first monoamine-oxidase inhibitor drug, iproniazid (Marsilid), was introduced as a drug similar in function to the antituberculotic drug isoniazid (see CHEMOTHERAPEUTIC DRUGS). The moods of patients receiving the drug were elevated more than their infections were relieved, and within a year, iproniazid was being used to elevate the mood of depressed patients. After the discovery that iproniazid could inhibit monoamine-oxidase activity, other inhibitor drugs with a wide range of chemical structures were developed; some of them resemble norepinephrine while others do not, but all manifest some degree of antidepressant activity. Five monoamine-oxidase inhibitor drugs are commonly used; these include isocarboxazid (Marplan), nialamide, phenelzine, tranylcypromine (Parnate), and pargyline (Eutonyl).

Since they are absorbed readily from the stomach, monoamine-oxidase inhibitor drugs are given orally. Although they remain in the body for only a short time, these drugs may inhibit the activity of monoamine oxidase for one to two weeks or until new monoamine-oxidase enzyme molecules are formed.

Toxic effects

Some persons receiving monoamine-oxidase inhibitor drugs suddenly develop very high blood pressure, sometimes with fatal consequences, after ingestion of certain foods containing large amounts of amines (*e.g.*, tyramine) that act as powerful blood-pressure-elevating agents. These amines, normally broken down by monoamine oxidase in the liver, circulate in the blood stream of patients receiving inhibitor drugs and cause an increase in blood pressure. Paradoxically, the most common side effect of monoamine-oxidase inhibitor drugs in patients on amine-free diets is low blood pressure; in fact, one monoamine-oxidase inhibitor drug, pargyline, is used primarily to lower blood pressure. The most severe side effect caused by the monoamine-oxidase inhibitor drugs is liver damage; for this reason, iproniazid is no longer used.

Tricyclic antidepressant drugs. The antidepressant properties of tricyclic antidepressant drugs were discovered by accident after they were introduced as drugs similar in chemical structure and function to phenothiazine drugs (antischizophrenic tranquillizers). Imipramine (Tofranil), the first tricyclic antidepressant drug,

soon was used to treat depressed patients, and the tricyclic drugs now are among the most effective antidepressant drugs known. In addition to imipramine, amitriptyline, desmethylimipramine, and nortriptyline (Aventyl) are used.

Tricyclic drugs may inhibit the inactivation of the norepinephrine-pumping process, thus potentiating the effects of norepinephrine. In persons not suffering from depression, the tricyclic antidepressant drugs cause lethargy and sedation, effects probably related to a tranquillizing action of the drugs. In treating depression the tricyclic antidepressant drugs relieve symptoms in a high proportion (about 70 percent) of patients, produce few side effects, and do not require the avoidance of specific foodstuffs; they generally are preferred to monoamine-oxidase inhibitor drugs. As with the monoamine-oxidase inhibitor drugs, the antidepressant action of tricyclic antidepressant drugs may not become apparent until two to four weeks after treatment is begun. The possible side effects of tricyclic antidepressant drugs include dry mouth, blurred vision, constipation, dizziness, and difficulty in urination.

Amphetamine and related stimulant drugs. Unlike the monoamine-oxidase inhibitor drugs and the tricyclic antidepressant drugs, amphetamine was produced intentionally as a drug with norepinephrine-like effects; unexpectedly, however, amphetamine proved to affect the central nervous system more than the peripheral nerves. Amphetamine was introduced in 1935 as a treatment for narcolepsy, a condition in which patients spontaneously fall asleep; eventually it also was used to overcome fatigue in soldiers and to treat obesity. In recent years, use of amphetamines and related drugs as euphoriants has been abused.

Amphetamine appears to exert its actions by potentiating norepinephrine released in the brain, as do the antidepressant drugs. Unlike antidepressant drugs, amphetamine can potentiate norepinephrine in more than one way, but stimulation of the direct release of norepinephrine from nerve endings is probably its major mode of action. Although amphetamine potentiates norepinephrine, it is not used as an antidepressant drug; in most cases, its use results in a deterioration of the condition of a psychotically depressed patient. The interactions of amphetamine with norepinephrine differ in subtle ways from those of the antidepressant drugs. One difference may be that antidepressant drugs potentiate norepinephrine normally released by nerve endings, while amphetamine causes release of norepinephrine.

Amphetamine derivatives

The amphetamine molecule is chemically asymmetric; *i.e.*, there are two different forms of amphetamine, one the mirror image of the other. Although each form has the same effect on blood pressure (small doses cause a slight increase), one of them, dextroamphetamine (Dexedrine), has much greater stimulant activity on the central nervous system than the other. Several drugs similar in structure and function to amphetamine are used primarily as dieting aids because they decrease the appetite; these include methamphetamine hydrochloride (Methedrine), benzphetamine (Didrex), phentermine (Ionamin), diethylpropion hydrochloride (Tenuate), phenmetrazine (Preludin), and phendimetrazine (Plegine).

In contrast to the lag in onset of action with the antidepressant drugs, the principal psychological effects of amphetamine occur within minutes after the drug is taken. Effects of amphetamine include increased alertness; decreased fatigue; elevation of mood; and increased initiative, confidence and ability to concentrate. Speech and activity also increase. Although intellectual sharpness is not increased by amphetamine or drugs related to it, physical performance is enhanced, presumably because the drugs decrease fatigue. Amphetamine and its derivatives now are used by soldiers, pilots, and astronauts on long-term missions; the apparent beneficial effects of amphetamine drugs (particularly dextroamphetamine and methamphetamine), however, have resulted in their use by many individuals who want pep pills. Although amphetamine's stimulation of the central nervous system and its appetite-decreasing effect were once thought to be closely linked, there is an amphetamine derivative, fenfluramine, that is a potent appetite suppressant but does not affect the central nervous system.

Disadvantages of amphetamine usage

The initial enthusiasm for amphetamine and related stimulant drugs was based largely on studies with single doses or short-term use by emotionally stable individuals. Only after these drugs were consumed extensively by many people did disadvantages become apparent. Tolerance to these stimulant drugs develops rapidly, so that progressively larger doses must be consumed. The letdown effect that occurs after stimulant effects have abated is most pronounced when a chronic user of a stimulant drug stops taking it; this effect also occurs, however, after a single dose of the drug has worn off. The letdown effect in chronic users often is a deep mental depression. In the 1960s, a pattern of stimulant-drug abuse developed. Intravenous injection—especially of the amphetamine derivative methamphetamine (Methedrine)—informally called speed, crystal, or meth—resulted in rapid development of tolerance, so that addicts soon required extremely large doses (500 milligrams compared with the usual medical dose of 5 to 10 milligrams).

After injection, methamphetamine produces a short-lasting feeling of euphoria, which is followed by a profound and long-lasting (up to several days) depression. The most severe aftereffect of large doses of amphetamine, administered either intravenously or orally, is amphetamine psychosis. Psychotic reactions, which may occur after drug withdrawal or during its usage, resemble those of paranoid schizophrenia. Amphetamine addiction is more common than can be accounted for by the relatively few cases of intravenous users called speed freaks. Patients who use amphetamine derivatives for dieting usually increase dosage as tolerance develops and then may require barbiturates to overcome the resultant insomnia. In these cases the danger of amphetamine psychosis is added to the threat of overdoses of either drug (see also DRUG PROBLEMS).

Amphetamines and hyperactive children

Amphetamine and its derivatives have had a major role in the management of hyperactive children. The hyperactivity syndrome apparently is a true disease that results from a delayed maturation of brain function. Children with this condition may have normal or superior intelligence but are unable to sit quietly or to concentrate. Amphetamine derivatives calm such children and enable them to concentrate, usually within a day after treatment is begun. Children with the syndrome can tolerate large doses, up to 40 milligrams of dextroamphetamine or methylphenidate (Ritalin), with few ill effects; such large doses make adults or nonhyperactive children jittery, sleepless, and unable to eat. There is no evidence thus far that amphetamine treatment of hyperactive children affects them adversely as they mature. Dextroamphetamine has been used in the treatment of narcolepsy and Parkinson's disease. Amphetamine and its derivatives, as previously noted, also are used in weight-reducing programs because they effectively decrease the appetite.

Methylxanthine stimulants: caffeine, theophylline, and theobromine. Methylxanthines are the active ingredients of coffee, tea, cocoa, and cola beverages that stimulate the central nervous system. Since billions of pounds of coffee, tea, and cocoa are consumed each year, caffeine, theophylline, and theobromine probably are the most widely used drugs.

The chemical structure of methylxanthines is similar to that of the purine bases adenine and guanine, which are components of nucleic acids (DNA and RNA) and uric acid, whose overproduction and precipitation into nodules accounts for the symptoms of gout. The three methylxanthines in foodstuffs, caffeine, theophylline, and theobromine, are derivatives of the purine base xanthine; caffeine has three methyl groups (carbon and three hydrogens), and theophylline and theobromine have two.

Stimulant content in various beverages

The principal methylxanthine in coffee and cola beverages is caffeine; tea contains theophylline and caffeine, and cocoa contains caffeine and considerable amounts of theobromine. Caffeine is the most potent methylxanthine stimulant. One cup of coffee or tea each contains the same amount of caffeine, about 100–150 milligrams.

Assuming that caffeine is the major stimulant ingredient in these beverages, coffee and tea should be equally active; tea, however, has a greater stimulant effect since it also contains theophylline, a weaker central nervous system stimulant. The major methylxanthine in cocoa (theobromine) has very little stimulant activity on the central nervous system, but an average cup of cocoa also contains 50 milligrams of caffeine. There are about 50 milligrams of caffeine in a 12-ounce glass of cola beverage.

The popularity of beverages containing methylxanthines appears to be related to their stimulant properties. In addition to allowing sustained intellectual effort and association of ideas, caffeine enhances discrimination of sensory stimuli, quickens reaction time, and decreases drowsiness and fatigue. The methylxanthines also are used in medicine. Caffeine has the most potent action, which is manifested in all parts of the brain, including centres that regulate respiration; therefore, caffeine is a valuable antidote to respiratory depression induced by drug overdose (*e.g.*, from morphine or barbiturates). The methylxanthines also facilitate breathing by relaxing the bronchioles in the lungs. Theophylline, the most active agent in producing these effects, is valuable for the management of severe asthma. The methylxanthines also enhance the production of urine by the kidneys and acid secretion by the stomach. The rate of the heartbeat and the total output of blood by the heart are stimulated by the methylxanthines, especially theophylline, which is used to treat congestive heart failure. Although the methylxanthines dilate most blood vessels, they constrict those of the brain; thus methylxanthines, especially caffeine, are useful in treating headaches caused by dilatation of blood vessels in the brain. Caffeine is a constituent of many nonprescription headache pills, which probably account for its most widespread medical use.

BIBLIOGRAPHY. P.H. CONNELL, *Amphetamine Psychosis* (1958), a classic monograph describing the features of amphetamine psychosis in a series of patients; D.W. ESPLIN and B. ZABLOCKA-ESPLIN, "Central Nervous System Stimulants," in L.S. GOODMAN and A. GILMAN (eds.), *The Pharmacological Basis of Therapeutics*, 4th ed. (1970), an excellent review of actions and therapeutic uses of stimulants, particularly agents such as caffeine; J.M. RITCHIE, "Central Nervous System Stimulants," pt. 2, *ibid.*, an extensive review of the literature of stimulants and their action in the brain; D.F. KLEIN and J.M. DAVIS, *Diagnosis and Drug Treatment of Psychiatric Disorders* (1969), a thorough description of the properties and uses of drugs in psychiatry, with an extensive discussion of antidepressants and clinical aspects of depression; S.H. SNYDER, "New Developments in Brain Chemistry: Catecholamine Metabolism and the Action of Psychotropic Drugs," *Amer. J. Orthopsychiat.*, 37:864–879 (1967), a lucid review article, understandable to the layman, concerning ways in which drugs affect mental functioning by neurochemical interactions with neurotransmitters.

(S.H.S.)

Stockholm

Stockholm, the capital and largest city in Sweden, is situated at the outlet of Mälaren (Lake Mälar) on the Baltic. The city is built upon 22 islands as well as the mainland of Uppland and Södermanland. Stockholm is by virtue of its position regarded as one of the most beautiful capital cities in the world. The name means "the island at Stocksund." Stocksund was the old name for Norrström (North Stream)—the expanse of water by which Lake Mälar is connected with the Baltic.

The city, which was founded in the middle of the 13th century, came to be officially regarded as the capital in 1436. The Stockholm of today is the second largest city in Scandinavia (Copenhagen is the largest) and has a population of 746,560 (1970). A shortage of land reserves has resulted in a slow decrease in population in the city itself. Greater Stockholm, however, is rapidly expanding and in the early 1970s contained over 1,343,000 inhabitants. As a result of fortuitous developments in the last 100 years, Stockholm has no slum areas. The emblem of Stockholm is a picture of St. Erik, first recorded in 1376.

Layout and boundaries. The layout of the city is characterized to the north by the Stockholm Ridge, comprising Observatory Hill, Brunkeberg Ridge, and the higher areas of Stadsholmen (City Island). To the south there is a steeply sloping fault scarp: the Skinnarvik Hills and the Southern Hills. Stadsholmen, the oldest part of Stockholm, is surrounded to the north by the North Stream, to the south by the South Stream, to the west by Lake Mälar, and to the east by the Baltic.

The Old Town

The original nucleus of the city is the "city between the bridges"—Gamla Stan (Old Town), consisting of Stadsholmen, Helgeandsholmen (Island of the Holy Spirit), and Riddarholmen (Nobles' Island). The buildings in this area are primarily from the 16th and 17th centuries. The well-preserved medieval city nucleus, with the original old network of streets from the Middle Ages as well as a large portion of the buildings, is of great antiquarian and historical interest and is legally protected from change. Stadsholmen contains the Royal Palace, Storkyrkan, also known as the Cathedral or Church of St. Nicolas, the German Church, the House of Lords, the government offices, the Stock Exchange, and a number of other notable buildings. Riddarholmen is dominated by the Riddarholm Church. The House of Parliament and the National Bank are situated on Helgeandsholmen. The area of lower Norrmalm contains the Opera, the Concert House, the Foreign Ministry (the Palace of the Prince Royal), the Central Station, and the Central Post Office. The Blasieholm Peninsula is dominated by the National Museum and the Grand Hotel. Skeppsholmen (Ship Island), a naval base from the 17th century until 1960, now houses the Modern Museum and the East Asian Museum in the old naval buildings. Just to the east lies Kastelholmen, with its citadel and its salute battery.

The Vasa district (Vasastaden) contains a university area around the Observatory Hill (including buildings of the University of Stockholm, the Stockholm School of Commerce, and the City Library). Most foreign embassies are to be found in the Östermalm district. This district also contains the Ministry of Defense and other military administrative units. In the 1930s, blocks of modern apartment buildings were constructed in the district of Ladugårdsgärdet, or Gärdet, for many years a parade ground. Northern Djurgården contains the studios and offices of the Swedish Broadcasting Corporation as well as the TV tower (Kaknästornet), one of the tallest buildings in Scandinavia. South Djurgården, which takes the form mainly of a natural park, contains Skansen and the Nordic Museum, Liljevalchs Art Hall, Waldermarsudde, the Thielska Gallery, and Gröna Lund ("green arbour") with its amusement park, Tivoli. The oldest of the western districts of the city is Kungsholmen, with its famous City Hall, the Serafimer Hospital, the Central Police Station, and the City Court. The offices of two major newspapers are to be found in Marieberg. Södermalm (South Malm) is noted for its steep cliff facing the north.

The area between the southern parts of the original city island and Södermalm is called Slussen (the Sluice) because of the sluice in the South Stream here, which is used to regulate the outflow of Lake Mälar into the Baltic. This part of the city has well-preserved blocks of houses from the 17th and 18th centuries. Traffic is very heavy around Slussen, and the centre of the area consists of an internationally famous network of overpasses. Västerbron (Western Bridge), which connects Kungsholmen with the south, passes over Långholmen (well-known for its prison). The west, northwest, and the south all contain large complexes of houses and multistory apartment buildings. These areas have their own shopping centres, such as Appelviken, Vällingby, and Skärholmen. The outer suburbs also contain large office blocks for state enterprises, such as the Swedish Board of Telecommunications in Farsta and the National Power Administration in Råcksta.

Reconstruction of the city centre

The central parts of the city are undergoing reconstruction, and this contributes to the decrease in population of this area. Sergels Torg (Sergel Square) at lower Norrmalm will be the new city centre, containing a large complex of buildings for cultural purposes. The Sergel area also contains the largest concentration of department



Gamla Stan (Old Town), in central Stockholm, as seen from atop the City Hall. The openwork spire of the Riddarholm Church is on Riddarholmen (Nobles' Island) in the foreground. At the upper left are the Royal Palace and Storkyrkan, Stockholm's cathedral.
Refot—Rapho Guillumette

stores in Sweden as well as a number of new central bank "palaces." Stockholm proper, which includes the incorporated parishes of Bromma, Brännkyrka, and Spånga, also shows a general tendency toward a reduction in population. This is counterbalanced, however, by increases in Greater Stockholm; that is, the city together with the 17 other communities in the inner suburbs.

The environs of Stockholm, both toward Lake Mälar and toward the Baltic, are famous for their beauty. The Stockholm Archipelago, consisting of 10,000 islands of varying size, most of which are uninhabited, is especially worthy of note in this respect. The land area of Stockholm is 72 square miles (186 square kilometres). The area of Greater Stockholm amounts to 640,000 acres, with 224,000 acres of forest and 100,000 acres of arable land. An uncommon geographical feature of Stockholm is the blending of the city buildings with a natural landscape, which not even the city centre can entirely smother. Water, crags, and forest areas provide Greater Stockholm with a beautiful environment. No other city of comparable dimensions is dispersed over so wide an area as Greater Stockholm.

History. Lake Mälar has been the determining factor in the development of Stockholm. Its waters have been used since antiquity for communications and trade. Lillön (Little Island), also known as Helgö (Saints' Island), in Lake Mälar was inhabited as early as the first centuries AD. A township existed here for many centuries and had extensive international contacts. About 800, Birka on Björkö was founded—the first real town in Sweden. Birka, probably initially created by the indigenous royal family, developed gradually into an important trading centre with connections with both western Europe and Russia as well as with the Byzantine Empire. The first Christian town in Sweden was Sigtuna, which assumed Birka's previous role at the beginning of the 11th century. The German colonization of the coastal area from Holstein to the Weichsel resulted in a rapid expansion of trade in the Baltic.

The medieval period

According to the medieval Erik's chronicle, which was begun in about 1320, Birger Jarl was responsible for the building of the city, and this appears to be reliable. The name is first recorded in documents in 1252, at which time Stockholm seems to have existed as a town for only a short period; archaeological studies suggest that the first buildings on the island were scarcely of earlier origin than the latter part of the 13th century. There were, however, villages and farms nearby as early as Viking times.

The founding of Stockholm was in all probability one of the measures taken by the royal family in its attempts to maintain control over the whole country during a phase of opposition from the Separatists of Uppland, immediately to the north of the future city. It is clear that the rapid growth of the city was connected with a trade agreement made in about 1250 between Birger Jarl and the ambitious people of the German city of Lübeck. This agreement ensured the people of Lübeck freedom from customs charges for their trade in Sweden as well as the right to settle there. Political and military events centred upon Stockholm during the latter part of the Middle Ages, and the town fortifications were developed, especially by Gustav I. The town retained its medieval character until the first half of the 17th century. The old medieval castle was rebuilt during the later period of the Vasa dynasty (1521-1654) in order to serve as a centre for the government of the country. The tower was raised and was provided with a spire with three crowns.

Sweden's period as a great power in the 17th century resulted in a considerable increase in prestige for Stockholm. The first organized construction in the so-called malm areas immediately to the north and south of the original Stadsholmen took place during this period. Norrmalm was a town in its own right between 1602 and 1636. The Ladugård area was finally incorporated into Stockholm in 1682. Kungsholmen (until 1672, Munkläget) was donated by Queen Christina to the town.

A trade agreement of 1636 prohibited all towns north of Stockholm and Åbo from trading with foreign countries, so that trade might be concentrated in the capital. In 1634 Stockholm became an independent administrative unit governed by a lord mayor. The first incumbent of the office, Claes Fleming, initiated extensive rebuilding of the city; new town plans were developed and carried out in the Renaissance style.

When the medieval royal castle burned down in 1697, plans for a new palace were begun by the architect Nicodemus Tessin the Younger. Work was in progress for the greater part of the 18th century. The city changed little during the period from the death of Charles XII in 1718

until the coronation of Gustav III in 1772. Devastating fires, which struck the Klara district in 1751 and Södermalm in 1773 and 1759, accentuated the need for more careful regulations with respect to building activities. In commercial terms this period was extremely active for Stockholm. The large merchants financed the ironworks in Bergslagen and distributed their products to the European market. During the Gustavian period, from 1772–1809, Stockholm's position as the cultural centre of the country was clearly underlined. A number of architecturally excellent buildings and city areas were built during this period.

City improvements in the 19th century

Stockholm had earned a very bad reputation for poor hygiene during the first half of the 19th century. But the introduction of municipally organized cleaning (1859) and sanitation (1861) and other sanitary arrangements during the latter half of the 19th century radically changed the situation and contributed to a rapid increase in population. The city began to be redesigned according to the French and German pattern, on the basis of a general outline proposed in the Lindhagen city plan (after the legal administrator and town planner Albert Lindhagen).

In 1860 Stockholm received its first railway connection with the south and west of Sweden and thereby also with the rest of Europe, which countered the previous provincialism. The Central Station was completed in the city centre in 1871, and the railway lines to the north and south of Stockholm could then be linked up. The last three decades of the 19th century saw further innovations, such as street cars, telephones, gas, and electricity.

The prominent ridge of Brunkeberg, which runs in a north–south direction through Norrmalm, for many years prevented a consistent application of town planning and impeded communications through the centre in an east–west direction. It was not until 1911 that the first broad street was excavated through the ridge; this was Stockholm's famous Kungsgatan, which in the 1970s was a centre for wholesale and retail trade. The first skyscraper in Sweden and Europe was built here in 1923—Northern Kungstornet.

Population. It is impossible to make an exact computation of the population of the city during the Middle Ages and the 16th century. It has been estimated that the population at the beginning of the 16th century was about 7,000. The number would appear to have increased to about 15,000 a century later. The population increased very rapidly during the 17th century, culminating in about 1675 at a level of 50,000-55,000. Population growth was directly related to the commercial revolution during the 17th and 18th centuries.

Until the 1860s the number of deaths was greater than that of births, so that population increases could come only from other areas. Infant mortality in Stockholm was considerably higher than in the rest of the country. Thirty-five percent of all children died before their first birthday in the first decade of the 19th century, and the figure was still more than 25 percent as late as 1880. It was not until the 1920s that the figure for Stockholm was reduced to the level for the country as a whole. In 1900 the city had 300,000 inhabitants, and in 1922 the half-million mark was passed. Stockholm attained its largest population of 808,000 in 1961. This population has been decreasing by 5,000 to 10,000 annually in recent years, and this reduction is expected to continue. Vacant lots for building purposes are very limited. The city centre has decreased in population since the end of the 1930s.

Principal manufactures

Economic life. Stockholm is Sweden's leading industrial area, although it is not normally regarded as an industrial city. Industries of special importance are metal and machine manufacturing, paper and printing, foodstuffs, and chemicals. Stockholm also possesses the country's largest wholesale stores, and the city accounts for about 20 percent of the total retail trade in Sweden. Banks and insurance companies are also concentrated around Stockholm. In addition, many large industrial concerns are situated in neighbouring administrative districts.

Stockholm is the second largest port in Sweden, Göteborg being the first. The total length of quays is about 12 miles. The most important harbours are Skeppsbron, which handles traffic to Finland; Stadsgårdskajen; Värta Harbour; the Free Harbour and Oil Harbour at Loudden in Lilla Värtan; and the Hammarby Harbour to the south.

Transport. Public transport is provided by a large number of bus lines and a well-developed underground system. The underground network consists of 39.3 miles, of which 14 miles are actually underground. Traffic to the more peripheral outer areas is handled primarily by means of modern commuter trains. Stockholm is also the starting point for four main railway lines: Stockholm-Malmö, Stockholm-Göteborg, Stockholm-Ånge, and the Västerås line.

A number of impressive bridges and motorways have been constructed to ease the complicated traffic conditions that result from Stockholm's unusual position on a large number of islands. The Essinge Motorway (Essingeleden) brings automobile traffic from the west of Kungsholmen at Fredhäll by means of two bridges over Lake Mälar at Essingen to Nyboda and Brännkyrka, where it links up with European Highway No. 4. The Huvudsta Motorway provides facilities for traffic toward the north and west. Bridges to the south are at Skanstull, Danvikstull, and Liljeholmen, with Årsta Bridge being reserved for railway traffic. The city centre is connected with the western districts of the city by the Traneberg Bridge.

Stockholm's airport Arlanda, near Märsta in Uppland, has a worldwide network of connections. Domestic and charter flights as well as private flights have Bromma Airport as their base. Stockholm's radio and television transmitters are situated in Nacka to the west of the city.

Administration and cultural life. *Local government.* Stockholm is administratively incorporated in Stockholm County. The city courts in Stockholm are lower courts. The courts have 24 departments and consist of the mayor, 79 magistrates, and a number of extra magistrates and judges, together with 1,000 jurors appointed by the City Council.

Municipal decision making is conducted by the City Council's 100 members. The administration is run by the Central Board of Administration (12 members), which is divided into nine sections, each under the supervision of a commissioner. The head of the financial section, the economic commissioner, is chairman of the Commissioners' council, which prepares matters to be dealt with in the City Council and attempts to achieve uniform principles in the administration. A number of councils, boards, and administrative units are responsible for day-to-day administration.

Stockholm County consists not only of Stockholm but also of 50 communities surrounding the capital. The county is administered by a County Council of 149 elected members. The Council is responsible for matters concerning regional policy, medical services, collective traffic, and housing. The state is represented in the county by means of a county governor whose residence is in Stockholm.

Religion. In addition to the Evangelical Lutheran State Church, there are a number of other churches and denominations to be found in Stockholm, such as Protestant free churches, Roman Catholic and Greek Orthodox churches, and Jewish rites. The vast majority of Stockholmers are members of the official state church.

Education. Stockholm is the largest educational centre in Sweden. In addition to the University of Stockholm, the city's colleges and universities include the College of Gymnastics and Sports, Stockholm School of Commerce, Caroline Medical Institute, an art college, a teachers' training college, a college of music, a school of forestry, a college of dentistry, Institute of Technology, and a veterinary college. The university is gradually being moved to Freskati, a campus 2½ miles to the north of the city. The transfer is expected to be completed by the end of the 1970s.

There are a number of learned academies and scientific institutions based in Stockholm. The most important are

the Royal Academy of Sciences, founded in 1739; the Royal Academy of Letters, History and Antiquities; and the Swedish Academy, founded in 1786.

Newspapers. Several newspapers are published in Stockholm. They include *Dagens Nyheter*, the largest morning paper in the country, and the evening paper *Expressen*, which is the largest in Scandinavia.

Music and theatre. The musical life of Stockholm is dominated by the Royal Theatre (the opera), the Concert Association (Stockholm Philharmonic Orchestra), and the Swedish Broadcasting Corporation, which has its own symphony orchestra. The leading theatres are the Royal Dramatic Theatre and the City Theatre. Regular opera performances are given during the summer in the Drottningholm Theatre, situated by the Royal Palace of Drottningholm to the west of the city. This theatre, which has been preserved intact since the 1760s, is unique in Europe by virtue of its original decorations.

BIBLIOGRAPHY. The extensive literature on Stockholm is mainly in Swedish. A number of more recent scholastic works have, however, been provided with English or German summaries. In English, see ELSPETH HARLEY SCHUBERT, *Brief Encounter with Stockholm* (1965); and TORD OLSSON NORDBERG, *Buildings and Monuments in Stockholm: A Short History and 8 Excursions* (1949). The latter provides a good account of the historical development of the city and a detailed description of important areas of the city and of its buildings.

(H.S.A.)

Stoicism

Stoicism, a school of thought that flourished in Greek and Roman antiquity, was one of the loftiest and most sublime philosophies in the record of Western civilization. In urging participation in the affairs of man, Stoics have always believed that the goal of all inquiry is to provide man with a mode of conduct characterized by tranquillity of mind and certainty of moral worth.

NATURE AND SCOPE OF STOICISM

For the early Stoic philosopher, as for all the post-Aristotelian schools, knowledge and its pursuit are no longer held to be ends in themselves. The Hellenistic Age was a time of transition, and the Stoic philosopher was perhaps its most influential spokesman. A new culture was in the making. The heritage of an earlier period, with Athens as its intellectual leader, was to continue, but to undergo many changes. If, with Socrates, to know is to know oneself, rationality as the sole means by which something outside of the self might be achieved may be said to be the hallmark of Stoic belief. As a Hellenistic philosophy, Stoicism presented an *ars vitae*, a way of accommodation for people to whom the human condition no longer appeared as the mirror of a universal, calm, and ordered existence. Reason alone could reveal the constancy of cosmic order and the originative source of unyielding value; thus, reason became the true model for human existence. To the Stoic, virtue is an inherent feature of the world, no less inexorable in relation to man than are the laws of nature.

Fundamental tenets

The Stoics believed that perception is the basis of true knowledge. In logic, their comprehensive presentation of the topic is derived from perception, yielding not only the judgment that knowledge is possible but also that *certain* knowledge is possible, on the analogy of the incorrigibility of perceptual experience. To them, the world is composed of material things, with some few exceptions (*e.g.*, meaning), and the irreducible element in all things is right reason, which pervades the world as divine fire. Things, such as material, or corporeal, bodies, are governed by this reason or fate, in which virtue is inherent. The world in its awesome entirety is so ruled as to exhibit a grandeur of orderly arrangement that can only serve as a standard for mankind in the regulation and ordering of his life. Thus, the goal of man is to live according to nature, in agreement with the world design. Stoic moral theory is also based on the view that the world, as one great city, is a unity. Man, as a world citizen, has an obligation and loyalty to all things in that city. He must play an active role in world affairs, remembering that the world exemplifies virtue and right action. Thus, moral worth, duty, and justice are singularly Stoic emphases, together with a certain sternness of mind. For the moral man is neither merciful nor shows pity, because each suggests a deviation from duty and from the fated necessity that rules the world. Nonetheless—with its loftiness of spirit and its emphasis on man's essential worth—the themes of universal brotherhood and the benevolence of divine nature make Stoicism one of the most appealing of philosophies.

Rival schools in antiquity

Its chief competitors in antiquity were: (1) Epicureanism, with its doctrine of a life of withdrawal in contemplation and escape from worldly affairs and its belief that pleasure, as the absence of pain, is the goal of man; (2) Skepticism, which rejected certain knowledge in favour of local beliefs and customs, in the expectation that these guides would provide man with the quietude and serenity that the dogmatic philosopher (*e.g.*, the Stoic) could not hope to achieve; and (3) Christianity, with its hope of personal salvation provided by an appeal to faith as an immanent aid to human understanding and by the beneficent intervention of a merciful God.

Along with its rivals, Stoicism enabled the individual to better order his own life and to avoid the excesses of human nature that promote disquietude and anxiety. It was easily the most influential of the schools from the time of its founding through the first two centuries AD, and it continued to have a marked effect on later thought. During the late Roman and medieval periods, elements of Stoic moral theory were known and used in the formulation of Christian, Jewish, and Muslim theories of man and nature, of the state and society, and of law and sanctions—*e.g.*, in the works of Marcus Cicero, Roman statesman and orator; in Lactantius, often called the "Christian Cicero"; and in A.M.S. Boethius, a scholar transitional to the Middle Ages. In the Renaissance, Stoic political and moral theory became more popular to theorists of natural law and politcal authority and of educational reform—*e.g.*, in Hugo Grotius, a Dutch jurist and statesman, and in Philipp Melanchthon, a major Reformation scholar. More recently, Stoicism has become popular again for its insistence on the value of the individual and the place of value in a world of strife and uncertainty —*e.g.*, in Existentialism and in Neo-orthodox Protestant theology. Stoicism also plays an important role in reassessments of the history of logic—*e.g.*, in Jan Łukasiewicz, a Polish logician, and in William and Martha Kneale, mid-20th-century British logicians.

ANCIENT STOICISM

Historical and philosophical setting

Early Greek Stoicism. With the death of Aristotle (322 BC) and that of Alexander the Great (323 BC), the greatness of the life and thought of the Greek city-state (polis) ended. With Athens no longer the centre of worldly attraction, its claim to urbanity and cultural prominence passed on to other cities—to Rome, to Alexandria, and to Pergamum. The Greek polis gave way to larger political units; local rule was replaced by that of distant governors. The earlier distinction between Greek and barbarian was destroyed; provincial and tribal loyalties were broken apart, first by Alexander and then by Roman legions. The loss of freedom by subject peoples further encouraged a deterioration of the concept of the freeman and resulted in the rendering of obligation and service to a ruler whose moral force held little meaning. The earlier intimacy of order, cosmic and civic, was now replaced by social and political disorder; and traditional mores gave way to uncertain and transient values.

Stoicism had its beginnings in a changing world, in which earlier codes of conduct and ways of understanding proved no longer suitable. But it was also influenced by tenets of the older schools. The earliest Greek philosophers, the Milesians, had called attention to cosmic order and the beauty of nature. Later, the monist Parmenides of Elea stressed the power of reason and thought, whereas Heracleitus of Ephesus, precursor of the philosophy of becoming, had alluded to the constancy of change and the omnipresence of divine fire, which illumined all things. A deeper understanding of man himself came with Socrates,

symbol of the philosophic man, who personified *sophia* and *sapientia* (Greek and Latin: "wisdom"). Of the several schools of philosophy stemming from Socrates, the Cynic and Megarian schools were influential in the early development of Stoic doctrine: the Cynics for their emphasis on the simple life, unadorned and free of emotional involvement; and the Megarians for their study of dialectic, logical form, and paradoxes.

The Early Stoa

Stoicism takes its name from the place where its founder, Zeno of Citium (Cyprus), customarily lectured—the Stoa Poikile (Painted Colonnade). Zeno, who flourished in the early 3rd century BC, showed in his own doctrines the influence of earlier Greek attitudes, particularly those mentioned above. He was apparently well versed in Platonic thought, for he had studied at Plato's Academy both with Xenocrates of Chalcedon and with Polemon of Athens, successive heads of the Academy. Zeno was responsible for the division of philosophy into three parts: logic, physics, and ethics. He also established the central Stoic doctrines in each part, so that later Stoics were to expand rather than to change radically the views of the founder. With some exceptions (in the field of logic), Zeno thus provided the following themes as the essential framework of Stoic philosophy: logic as an instrument and not as an end in itself; human happiness as a product of life according to nature; physical theory as providing the means by which right actions are to be determined; perception as the basis of certain knowledge; the wise man as the model of human excellence; Platonic Ideas—or the abstract Forms that things of the same genus share—as being unreal; true knowledge as always accompanied by assent; the fundamental substance of all existing things as being a divine fire, the universal principles of which are (1) passive (matter) and (2) active (reason inherent in matter); belief in a world conflagration and renewal; belief in the corporeality of all things; belief in the fated causality that necessarily binds all things; cosmopolitanism, or cultural outlook transcending narrower loyalties; and man's obligation, or duty, to choose only those acts that are in accord with nature, all other acts being a matter of indifference.

Cleanthes of Assos, who succeeded Zeno as head of the school, is best known for his *Hymn to Zeus*, which movingly describes Stoic reverence for the cosmic order and the power of universal reason and law. The third head of the school, Chrysippus of Soli, who lived to the end of the 3rd century, was perhaps the greatest and certainly the most productive of the early Stoics. He devoted his considerable energies to the almost complete development of the Zenonian themes in logic, physics, and ethics. In logic particularly, he defended against the Megarian logicians and the Skeptics such concepts as certain knowledge, comprehensive presentation, proposition and argument, truth and its criterion, and assent. His work in propositional logic, in which unanalyzed propositions joined by connectives are studied, made important contributions to the history of ancient logic and is of particular relevance to more recent developments in logic.

In physics, Chrysippus was responsible for the attempt to show that fate and free will are not mutually exclusive conceptual features of Stoic doctrine. He further distinguished between "whole" and "all," or "universe," arguing that the whole is the world, while the all is the external void together with the world. Zeno's view of the origin of man as providentially generated by "fiery reason" out of matter was expanded by Chrysippus to include the concept of self-preservation, which governs all living things. Another earlier view (Zeno's), that of nature as a model for life, was amplified first by Cleanthes and then by Chrysippus. The Zenonian appeal to life "according to nature" had evidently been left vague, because to Cleanthes it seemed necessary to speak of life in accord with nature conceived as the world at large (the cosmos), whereas Chrysippus distinguished between world nature and human nature. Thus, to do good is to act in accord with both human and universal nature. Chrysippus also expanded the Stoic view that seminal reasons were the impetus for animate motion.

He established firmly that logic and (especially) physics are necessary and are means for the differentiation of goods and evils. Thus, a knowledge of physics (or theology) is required before an ethics can be formulated. Indeed, physics and logic find their value chiefly in this very purpose. Chrysippus covered almost every feature of Stoic doctrine and treated each so thoroughly that the essential features of the school were to change relatively little after his time.

The Middle Stoa

Later Roman Stoicism. The Middle Stoa, which flourished in the 2nd and early 1st centuries BC, was dominated chiefly by two men of Rhodes: Panaetius, its founder, and his disciple Poseidonius. Panaetius organized a Stoic school in Rome before returning to Athens, and Poseidonius was largely responsible for an emphasis on the religious features of the doctrine. Both were antagonistic to the ethical doctrines of Chrysippus, who, they believed, had strayed too far from the Platonic and Aristotelian roots of Stoicism. It may have been because of the considerable time that Panaetius and Poseidonius lived in Rome that the Stoa there turned so much of its emphasis to the moral and religious themes within the Stoic doctrine. Panaetius was highly regarded by Cicero, who used him as a model for his own work. Poseidonius, who had been a disciple of Panaetius in Athens, taught Cicero at his school at Rhodes and later went to Rome and remained there for a time with Cicero. If Poseidonius admired Plato and Aristotle, he was particularly interested—unlike most of his school—in the study of natural and providential phenomena. In presenting the Stoic system in the second book of *De natura deorum* (45 BC), Cicero most probably followed Poseidonius. Because his master, Panaetius, was chiefly concerned with concepts of duty and obligation, it was his studies that served as a model for the *De officiis* (44 BC) of Cicero. Hecaton, another of Panaetius' students and an active Stoic philosopher, also stressed similar ethical themes.

If Chrysippus is to be commended for his diligence in defending Stoic logic and epistemology against the Skepticism of the New Academy (3rd–2nd century BC), it was chiefly Panaetius and Poseidonius who were responsible for the widespread popularity of Stoicism in Rome. It was precisely their turning of doctrine to themes in moral philosophy and natural science that appealed to the intensely practical Romans. The times perhaps demanded such interests, and with them Stoicism was to become predominantly a philosophy for the individual, showing how—given the vicissitudes of life—one might be stoical. Law, world citizenship, nature, and the benevolent workings of Providence and the divine reason were the principal areas of interest of Stoicism at this time.

The Late Stoa

These tendencies toward practicality are also well illustrated in the later period of the school (in the first two centuries AD) in the writings of Lucius Seneca, a Roman statesman; of Epictetus, a slave freed by Nero; and of Marcus Aurelius, a Roman emperor. Both style and content in the *Libri morales* (Eng. trans., *Moral Essays*) and *Epistulae morales* (*Moral Letters*) of Seneca reinforce the new direction in Stoic thought. The *Encheiridion* (*Manual*) of Epictetus and the *Meditations* of Marcus Aurelius furthered the sublime and yet personal consolation of the Stoic message and increasingly showed the strength of its rivalry to the burgeoning power of the new Christianity. The mark of a guide, of the religious teacher, is pre-eminent in these writings. It is difficult to establish with any precision, however, the extent of Stoic influence by the time of the first half of the 2nd century AD. So popular had these ideas become that many specifically Stoic terms (viz., right reason, comprehension, assent, indifference, *logos*, natural law, and the notion of the wise man) commonly were used in debate and intellectual disputes.

Stoic elements in Pauline and patristic thought. There is much disagreement as to the measure of Stoic influence on the writings of St. Paul, the Apostle of Christ. At Tarsus, Paul certainly had opportunities for hearing Stoic lectures on philosophy. And it may be that his discussion of nature and the teaching of it (I Cor. 11:14) is Stoic in origin, for it has a parallel in the *Manual* of Epictetus 1.16, 10. Although not a Stoic technical term, *syneidēsis*,

which Paul used as "conscience," was generally employed by Stoic philosophers. In I Cor. 13 and in the report of Paul's speech at Athens (Acts 17), there is much that is Hellenistic, more than a little tinged by Stoic elements—*e.g.*, the arguments concerning man's natural belief in God and the belief that man's existence is in God.

The assimilation of Stoic elements by the Church Fathers was generally better understood by the 4th century. Stoic influence can be seen, for example, in the relation between reason and the passions in the works of St. Ambrose, one of the great scholars of the church, and of Marcus Minucius Felix, a Christian Apologist. Each took a wealth of ideas from Stoic morality as Cicero had interpreted it in *De officiis*. In general, whereas the emerging Christian morality affirmed its originality, it also assimilated much of the pagan literature, the more congenial elements of which were essentially Stoic.

Stoic concepts in Tertullian and Origen

Earlier, in the 3rd century, Quintus Tertullian, often called the father of Latin Christian literature, seems to have been versed in Stoic philosophy; *e.g.*, in his theory of the agreement between the supernatural and the human soul, in his use of the Stoic tenet that from a truth there follow truths, and in his employment of the idea of universal consent. Even in his polemical writings, which reveal an unrelenting hostility to pagan philosophy, Tertullian showed a fundamental grasp and appreciation of such Stoic themes as the world *logos* and the relation of body to soul. This is well illustrated in his argument against the Stoics, particularly on their theme that God is corporeal being and identified with reason as inherent in matter—also to be found in his polemics against Marcion, father of a heretical Christian sect, and against Hermogenes of Tarsus, author of an important digest of rhetoric. Yet in his doctrine of the Word, he appealed directly to Zeno and Cleanthes of the Early Stoa. Another important polemic against the Stoics is found in the treatise *Contra Celsum*, by Origen, the most influential Greek theologian of the 3rd century, in which he argued at some length against Stoic doctrines linking God to matter.

Also, St. Cyprian, bishop of Carthage in the 3rd century, revealed the currency of Stoic views; *e.g.*, in his *Ad Demetrianum*, a denunciation of an enemy to Christianity, in which Cyprian castigates the ill treatment of slaves, who, no less than their masters, are formed of the same matter and endowed with the same soul and live according to the same law. The beliefs in the brotherhood of man and in the world as a great city, commonly found in early Christian literature, were current Stoic themes. The Christian attitude appears in what St. Paul said of Baptism: "You are all sons of God through Faith. For as many of you as were baptized into Christ have put on Christ" (Gal. 3:26–27).

STOICISM IN MEDIEVAL AND MODERN PHILOSOPHY

Stoic undercurrents in medieval thought. During the period when Christian institutions and doctrines were developing (AD 230–1450), Stoicism continued to play a popular role. The *De consolatione philosophiae* (524) of Boethius (died AD 524/525) was widely known and appreciated as a discourse on the mysterious questions of the nature of good and evil, of fortune, chance, or freedom, and of divine foreknowledge. If the plan of Boethius was to serve as an interpreter of Plato and Aristotle, he succeeded only in working through some logical theories of Aristotle, together with several commentaries on those theories. In the *Consolatione*, however, the themes are quite different; in the fifth book, for example, he attempted to resolve the apparent difficulty of reconciling human freedom with the divine foreknowledge, a problem that among Stoic thinkers—though by no means uniquely among them—had been in general currency for a long time. This work of emancipation from worldly travail through the glories of reason and philosophy, which included Stoic doctrines as found in the writings of Cicero and Seneca, was much more influential for later medieval thought than that of Lactantius, of the late 3rd to early 4th century, who was largely concerned with the writing of a history of religion—a summary statement of Christian doctrine and life from earliest times. Lactantius also wrote a not unimportant work, *De ira Dei* (313), on the possibility of anger in God. It poses a problem of how to deal with the essentially Greek, or philosophic, view that God cannot feel anger because he is not subject to passions and that *apatheia* ("apathy," or "imperturbableness") is not merely the mark of the wise man but is also a divine attribute. This view, which had been most thoroughly developed among Stoic thinkers and particularly by Epictetus, raised a peculiarly Christian problem, the concern of the power of God to reward the righteous and punish the transgressor; thus, it challenged the very idea of Providence. Other manifestations of anthropopathism, the attributing of human feelings to God, had also been charged against the early Christian religionists; and the writers of the time—Lactantius and Tertullian among them—took great pains to refute the largely Stoic formulations of these charges. Although the refutations took the form—in St. Augustine, for example—of denying that the wrath of God is a perturbation of the soul and of holding that it is rather a judgment, the concept of the divine essence excludes all passions. Within the monastic tradition, there remained more than a residue of concern over apathy as a divine attribute and as a model for the truly religious.

Stoicism in legal and political theory

Other significant Stoic influences appeared in medieval discussions of the popular origin of political authority, *res publica* and *res populi*, and on the distinctions made in law between *jus naturale, jus gentium, jus civile*—doctrines of Stoic origin—found in 3rd-century Roman juridical texts gathered together by Isidore of Seville (died AD 636), a Spanish encyclopaedist and theologian. The Stoic belief—as against Aristotle—that men are by nature equal was an integral part of the knowledge that certain rules of law are universally recognized, laws that all people might naturally follow. In this way, the Romans—whose genius lay in organization and in law—fostered the conception of natural, or common, law, which reason was supposed to make evident to all men. Thus, in the second half of the 11th century, the Stoic texts of Cicero and Seneca became important doctrinal sources for the initial discussions of social and political philosophy. These early theories of law, of the natural equality of men, and of the rights of prince and populace were to become the basis for 13th-century systems of social and political privilege and obligation.

In the 12th century, John of Salisbury, an English critical scholar, produced, in his *Policraticus* (1159), the first complete attempt at a philosophy of the state since Classical times. Stoic doctrines of natural law, society, state, and Providence were important elements in his effort to construct a social philosophy on ethical and metaphysical principles. The impact of these doctrines and the lengthy history of their use in the earlier Middle Ages can also be found in the views of Thomas Aquinas on the philosophy of the state and of man.

Renascence of Stoicism in modern times. If the influence of Stoic doctrines during the Middle Ages was largely restricted to the resolution of problems of social and political significance, it remained for the Renaissance, in its passion for the rediscovery of Greek and Roman antiquity, to provide a basis for the rebirth of Stoic views in logic, epistemology, and metaphysics, as well as the documentation of the more familiar Stoic doctrines in ethics and politics. Late in the 16th century, Justus Lipsius, a Flemish scholar and Latin Humanist, was responsible for the first restatement of Stoicism as a defensible and thoroughgoing (Christian) philosophy of man. His treatises *De constantia* (1584) and *Politicorum sive civilis doctrinae libri V* (1589) were widely known in many editions and translations. His defense of Stoic doctrine in *Manuductio ad Stoicam Philosophiam* (1604) and *Physiologia Stoicorum* (1604) provided the basis for the considerable Stoic influence during the Renaissance. Around the turn of the 17th century, Guillaume du Vair, a French lawyer and Christian philosopher, made Stoic moral philosophy popular, while Pierre Charron, a French theologian and Skeptic, utilized Stoic themes in *De la sagesse* (1601; Eng. trans., *Of Wisdome*, 1608), as did the Skep-

Stoic concepts in Boethius and Lactantius

Stoic concepts in Lipsius and certain Skeptics

tic Michel de Montaigne in his *Essais* (1580; Eng. trans., 1603). Through the work of Lipsius, Stoic doctrines were to influence the thought of Francis Bacon, a precursor of modern philosophy of science, and, later, the *De l'esprit des lois* (1748; Eng. trans., *The Spirit of Laws*, 1750), by the political theorist Charles-Louis, baron de Montesquieu. In the continuing and relentless war against the Aristotelianism of the later Middle Ages, the doctrines of Stoicism influenced many prominent figures of the Renaissance and Reformation periods.

Pietro Pomponazzi, an Aristotelian of early-16th-century Italy, in defending an anti-Scholastic Aristotelianism against the Averroists, who viewed the world as a strictly necessitarian and fated order, adopted the Stoic view of Providence and human liberty. The 15th-century Humanist Leonardo Bruni absorbed Stoic views on reason, fate, and free will. Pantheism, the view that God and nature are unitary in the sense that God is an impersonal being, and naturalism, the view that nothing is supernatural, both of which identify God with the cosmos and ascribe to it a life process of which the world soul is the principle, were widely held Renaissance notions. Such a pantheistic naturalism was advocated—though from diverse standpoints—by Francesco Patrizi, a versatile Platonist, and by Giordano Bruno, defender of an infinite cosmos; and in both authors the inspiration and source were fundamentally Stoic. In the development of a philosophy of public law based upon a study of man, Stoic elements are found in the *Utopia* (1516; Eng. trans., 1551), by Thomas More, and the *De Jure Belli ac Pacis* (1625; Eng. trans., 1682), by Hugo Grotius. This latter work is one of the most famous Renaissance treatises on the theory of natural and social rights.

The foremost Swiss reformer of the early 16th century, Huldrych Zwingli, who regarded justification by subjective belief as the foundation of the new Christianity, utilized Stoic views on the autonomy of the will, on the absolute predestination of the good and evil man, and on moral determinism.

Another Stoic influence of considerable importance in the tradition of Christian Humanism was the view that all religions have a common basis of truths concerning God —a universal Deism. Among those who favoured such a view were Zwingli and Desiderius Erasmus, the great Renaissance Humanist and scholar. More and Grotius also laid special stress on this view, and its influence was felt in the moral, social, and even the artistic life of the 16th century. Later, Herbert of Cherbury, often called the father of Deism, further developed the idea of religious peace and the reduction of opposing religious views to common elements. This view became one of the most popular ideas of the 17th century.

Philipp Melanchthon also cultivated Humanism and the philosophy of antiquity as a basis for a reborn Christianity. Although Aristotle was his chief inspiration, Melanchthon made telling use of the Stoic theory of knowledge, with its notions of innate principles and the natural light of reason, which teach man the great truths of metaphysical and moral order. Stoicism thus became the basis for the natural-law theory, which holds that the state is of immediately divine origin and independent of the church —a Protestant view opposed by Catholic writers.

Stoic concepts in Cartesian and Spinozistic thought

The Cartesian revolution in thought in the 17th century brought forward several Stoic notions: that morality consists of obedience to the law of reason, which God has deposited within man; that ethics presupposes a knowledge of nature, because man must learn to know his place in the world, for only then may he act rightly; that self-examination is the foundation of ethics; and that the innateness and commonality of truths bespeak the view that only thoughts and the will belong properly to man, for the body is a part of the material world. Such views were particularly developed by René Descartes, often hailed as the father of modern philosophy, in his dualism of mind (or soul) and body.

Benedict de Spinoza, a freethinking Jewish Rationalist, made similar use of Stoic views on the nature of man and the world. That aspect of Spinoza's thought that is debatably labelled pantheist is essentially Stoic in character. Together with the Cartesians, Spinoza insisted upon the importance of internal and right reason as the sole means by which to attain to indubitable truths and to the possibility of human freedom.

Blaise Pascal, a French scientist and religious writer, also was sympathetic to Cartesian conceptions of man and nature. Though he turned his back on philosophy, his religious thought retained the Cartesian and Stoic insistence on the independence of human reason, holding that man is fundamentally a thinking being, innately capable of making right decisions. There is an important and crucial difference and conflict between Pascal's views and those of Spinoza and the Cartesians: for Pascal, though the use of (the Stoic) right reason might result in proofs and demonstrations that lead to the God of truth, it would never lead to the God of love, the one true God. Thus, the Stoic exaltation of reason to an entity in its own right—indeed, to a divine entity—as exemplified among the Cartesians and in the thought of Spinoza, was rejected by Pascal in the Jansenist Christianity that he finally adopted—a rejection that, because it also repudiated free will, distinguishes Pascal from those who held Stoic as well as alternative conceptions of human freedom and responsibility.

Stoic concepts in Christianity and modern philosophical thought

Christianity in general, in spite of striking contrasts with Stoicism, has found elements within it that parallel its own position. As the Stoic, for example, feels safe and protected in the rational care of some immanent Providence, so the Christian senses that a transcendent though incarnate and loving God is looking after him. And in general, Stoicism has played a great part throughout the ages in the theological formulation of Christian thought as well as in the actual realization of the Christian ideals.

Contemporary philosophy has borrowed from Stoicism, at least in part, its conviction that man must be conceived as being closely and essentially connected with the whole universe. And contemporary Humanism still contains some obviously Stoic elements, such as its belief in the solidarity of all peoples based upon their common nature, and in the primacy of reason. It is perhaps just because Stoicism has never become a full-fledged philosophic system that, many centuries after the dissolution of the Stoic school, fundamental themes of its philosophy have emerged again and again, and many have become incorporated into modern thinking.

BIBLIOGRAPHY

Primary sources: HANS VON ARNIM, *Stoicorum Veterum Fragmenta*, 4 vol. (1905–24), is the standard text collection. The following works of classic Stoic authors and substantive issues in Stoicism in antiquity are most conveniently available in the "Loeb Classical Library": MARCUS AURELIUS, *Meditations;* EPICTETUS, *Discourses;* CICERO, *De fato, De finibus, De natura deorum,* and *Academica;* DIOGENES LAERTIUS, *Lives of Eminent Philosophers*, vol. 2, book 7; SENECA, *Epistulae Morales* and *Moral Essays;* SEXTUS EMPIRICUS, *Outlines of Pyrrhonism* and *Against the Dogmatists*, 4 vol.

Classical and Roman Stoicism: GEORGE MURRAY, *The Stoic Philosophy* (1915), the classical statement of the grandeur of the Stoic philosophy; MAX POHLENZ, *Die Stoa*, 2 vol. (1948–49); and PAUL BARTH, *Die Stoa* (1903), are representative of scholarly studies of Stoic philosophy; JOHNNY CHRISTENSEN, *An Essay on the Unity of Stoic Philosophy* (1962), is a recent and comprehensive essay; LUDWIG EDELSTEIN, *The Meaning of Stoicism* (1966); and JOSIAH B. GOULD, *The Philosophy of Chrysippus* (1970), are among the best of recent studies of Greco-Roman Stoicism; EDUARD ZELLER, *Die Philosophie der Griechen in ihrer geschichtlichen Entwicklung*, 5 vol. (1856–68; in part translated as *Stoics, Epicureans, and Skeptics*, rev. ed., 1962); EDWYN R. BEVAN, *Stoics and Sceptics* (1913, reprinted 1959); and ROBERT D. HICKS, *Stoic and Epicurean* (1910, reprinted 1962), are illustrations of Stoic philosophy in the Hellenistic period. See also JASON L. SAUNDERS, *Greek and Roman Philosophy After Aristotle* (1966); and HARRY A. WOLFSON, *Philo: Foundations of Religious Philosophy in Judaism, Christianity and Islam*, 2 vol. (1962). EDWARD V. ARNOLD, *Roman Stoicism* (1911, reprinted 1958); and FREDERICK W. BUSSELL, *Marcus Aurelius and the Later Stoics* (1910), are excellent statements of Stoicism in the later Roman period.

Later Stoic influences: PIERRE DE LABRIOLLE, *Histoire de la littérature latine chrétienne* (1920; Eng. trans., *History and*

Literature of Christianity from Tertullian to Boethius, 1924), an excellent presentation of the influence of Stoic views in late antiquity and the patristic period; ETIENNE GILSON, *History of Christian Philosophy in the Middle Ages* (1955); HARRY A. WOLFSON, *The Philosophy of the Church Fathers*, 3rd ed. rev. (1970); MAURICE DE WULF, *Histoire de la philosophie médiéval*, 2 vol. (1925; Eng. trans., *History of Mediaeval Philosophy*, 2 vol., 1926); and ISAAC HUSIK, *A History of Medieval Jewish Philosophy* (1940), are careful introductions to Stoic influences in patristic and medieval times; JASON L. SAUNDERS, *Justus Lipsius: The Philosophy of Renaissance Stoicism* (1955); HAROLD HOFFDING, *A History of Modern Philosophy*, 2 vol. (1950); and FREDERICK C. COPLESTON, *A History of Philosophy*, rev. ed., vol. 3–6 (1962), trace the Stoic influence from its revival in the Renaissance into modern philosophy.

Stoic logic and physics: BENSON MATES, *Stoic Logic* (1953); SAMUEL SAMBURSKY, *The Physical World of the Greeks*, 2nd ed. (1960; orig. pub. in Hebrew) and *Physics of the Stoics* (1959).

(Ja.L.S.)

Stone Age European Peoples, Arts of

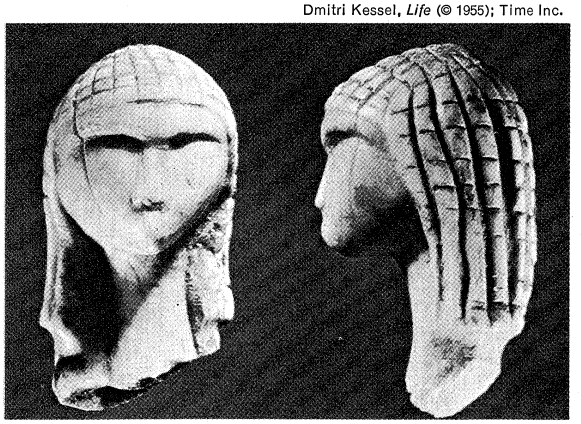
The earliest known example of something definable as art dates from the appearance of Cro-Magnon man (*Homo sapiens*) in the last glacial and post-glacial phases of the Paleolithic period (32,000–10,000 BC). Subsequent phases of the arts of the Stone Age peoples include the Mesolithic, which began around 10,000 BC, and the Neolithic, which began around 6,500 BC in the eastern Mediterranean area and around 3,000 BC in Spain and northern Europe. Although other peoples passed through similar phases of Stone Age culture, this article includes coverage only of the arts of Stone Age European peoples. Arts of Stone Age Asian peoples are treated in history articles and in SOUTHEAST ASIAN PEOPLES, ARTS OF; SOUTH ASIAN PEOPLES, ARTS OF; and CENTRAL ASIAN PEOPLES, ARTS OF. The arts of the Stone Age American Indian, African, and Oceanic peoples are dealt with in AMERICAN INDIAN, ARTS OF; AFRICAN PEOPLES, ARTS OF; OCEANIC PEOPLES, ARTS OF.

UPPER PALEOLITHIC AND MESOLITHIC AGES

Although throughout the Paleolithic period (from about 600,000 or 700,000 years ago until about 10,000 BC) man lived by fishing, gathering plant food, and hunting, he made great cultural progress during the Upper Paleolithic period. The hand axes and flake tools of the earlier assemblages were replaced by diversified and specialized tools made on blades struck from specially prepared cores. Many important inventions appeared, such as needles and thread, skin clothing, stone and bone tools provided with handles, the harpoon, the spear-thrower, and special fishing equipment. Bone, ivory, and antler, in addition to flint, were used extensively. The earliest man-made dwellings, consisting of semi-subterranean pit houses, are found in this period. Of prime importance and interest is the beginning of the basic techniques of drawing, modelling, sculpting, and painting, as well as the earliest manifestations of dancing, music, the use of masks, ceremonies, and the organization of society into patterns that were apparently fairly complex. Indeed, the location of certain settlements suggests a fairly complex social life, perhaps including collective hunting. There is evidence for fertility magic, private property, and possible social stratification. Furthermore, primitive types of early man disappeared, and the remains of men of modern type (*Homo sapiens*) alone are found in Upper Paleolithic sites.

Beginning of art

Dmitri Kessel, *Life* (© 1955); Time Inc.



Views of an ivory female head (fragment of a "Venus" figurine), Aurignacian-Gravettian (c. 22,000 BC). From the Grotte du Pape, Brassempouy, Landes, France. In the Musée des Antiquités Nationales, Saint-Gérmain-en-Laye, France. Height 3.5 cm.

The chronology of this interval in Europe shows a succession of cultures known as Lower Perigordian (*c.* 32,000–*c.* 28,000 BC), Aurignacian (*c.* 28,000–*c.* 22,000 BC), from which the oldest known art is dated, Upper Perigordian (*c.* 22,000–*c.* 18,000 BC), Solutrean (*c.* 18,000–*c.* 15,000 BC), and Magdalenian (*c.* 15,000–*c.* 10,000 BC), each characterized by its distinctive types of artifacts along with gravers (or burins), end scrapers, points, and so on, which are common to all levels. The graver itself is an important tool, the invention of which made possible the extensive working of bone and facilitated the development of art. The climate of the Upper Paleolithic varied from cold steppe or even Arctic tundra to north temperate (taiga), similar to parts of Siberia and Canada of the present day.

Mesolithic Age

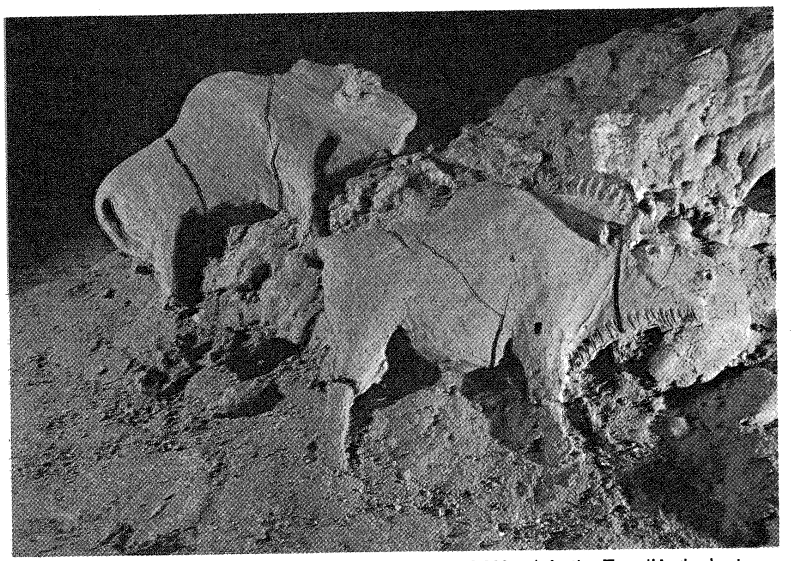
The retreat of the ice sheet resulted in radical changes of climate, fauna, and flora and gave rise to a tundra-like vegetation that was later replaced by temperate forest land. The first human cultures of the Mesolithic period appear to have been degeneracies of the Upper Paleolithic cultures; adaptation to the new environment proved difficult. Consequently the equipment of the Azilians of France, Switzerland, and southern Germany (descendants of the Magdalenians) was relatively poor; their antler harpoons were large and clumsy, their flint implements, known as microliths, became extremely small, and Magdalenian cave art degenerated into conventionalized designs painted on pebbles. Full adaptation to the environment was not achieved until the Boreal period. The many local cultures that flourished at that time (6,800–5,000 BC) are collectively referred to as Tardenoisian, known largely from their carefully trimmed microliths. They lived mainly on unforested sandy ground and left some of the earliest traces of habitations—skin or birch-bark tents in the summer and half-subterranean timber-framed shelters in the winter.

At the same time in northern Europe, communities living in a forest environment developed elaborate hunting and fishing equipment. These Maglemosians were among the first people to build canoes and to tame a domestic animal, the dog. Toward the end of the Mesolithic period in western Europe (*c.* 4,000 BC), they could grind stone and make rudimentary pottery, techniques probably derived from the advanced civilizations of the Near East. After the retreat of the glacier, the cultures of eastern Europe became more differentiated. In the south and to some extent in the central wooded belt, Stone Age cultures resembled those of Western Europe (microlithic stone industries), while their northern neighbours were more like the Mesolithic Scandinavians (heavy stone tools, bone points, and so on).

The northern peoples may have developed further in what can be termed the Epi-Mesolithic period. They already had acquired some cultural traits characteristic of the Neolithic (mainly pottery of the pitted-ware type and ground stone axes) but were ignorant of agriculture and the breeding of domesticated animals. The origin and chronology of their cultures has not yet been fully cleared up; it is believed, however, that they survived very late and were contemporary with the central and west European Bronze Age. (Ed.)

Venus figures

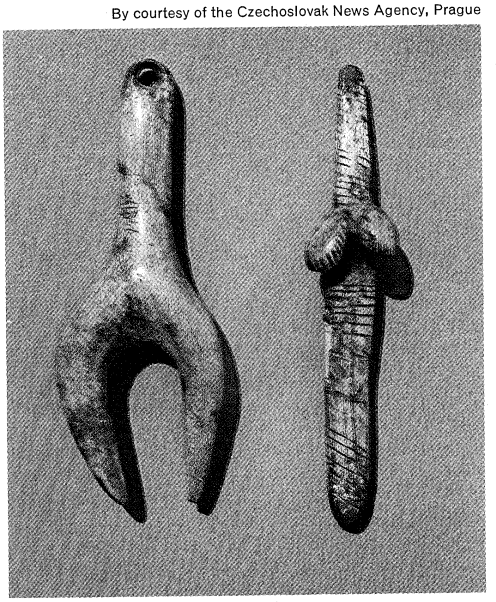
Paleolithic art: carved and incised objects. Although human beings are rarely represented in the cave art of western Europe, the earliest plastic figures represent mature, large-buttocked females with pendulant breasts. These figurines, which show remarkable consistency in style over their total distribution from France to southern U.S.S.R., are carved in bone or schist, once coloured, and



Two bison modelled in clay, Magdalenian (c. 15,000–c. 10,000 BC). In the Tuc d'Audoubert cave, Ariège, France. Length of bison at right 61 cm.
J. Vertut—Ziolo

are known now as "Venuses." The earliest examples come from middle Aurignacian contexts; those from the later Aurignacian–Perigordian period are of higher quality. About 130 examples are known, many from encampment sites of mammoth hunters at Gagarino, Ukraine, U.S.S.R., where special niches to house them were built into the hut walls. One of the most noted examples is from Willendorf, Austria (Naturhistorisches Museum, Vienna); and another, the head of such a figure with retted hair and beautifully marked facial features, is a famous Venus from Brassempouy, Landes, France (Musée les Antiquités nationales, Saint-Germain-en Laye). In the Ukraine and in Moravia, Czechoslovakia, these figurines became highly schematized, reduced often to pegs of ivory or bone with highly developed buttocks and breasts. Few bodily features are marked, although some examples have extremely thin arms folded over the chest. Moravian examples have incised geometric spiral designs on the back and buttocks that might represent tattooing. Male statuettes are much rarer and are artistically inferior, but a male head from Dolní Věstonice in Moravia (Moravian Museum, Brno) has realistically carved features.

By courtesy of the Czechoslovak News Agency, Prague



Stylized "Venus" figurines carved in ivory, Aurignacian-Gravettian (c. 24,800 BC). From Dolní Věstonice, Mikulov, Moravia, Czechoslovakia. In the Moravian Museum, Brno, Czechoslovakia. Height (left) 8.3 cm and (right) 8.6 cm.

Larger scale sculpture is almost entirely confined to France and probably evolved from natural rock forms of suggestive shape. A parallel to the Aurignacian Venus figurines is a low relief from Laussel, Dordogne, of a large-buttocked female figure holding a "cornucopia." (Musée de L'Homme, Paris). Similar Perigordian reliefs were carved in partly exposed rock shelters at Cap Blanc, Dordogne, and some of ibex and wild boar have been found on the Roc-de-Sers, Charente. High-relief animal sculptures at Angles-sur-l'Anglin, Vienne, are Magdalenian. The only piece of sculpture in the round is a head of a musk ox from Laugerie-Haute, Dordogne, nine inches high (Musée des Antiquités Nationales, Saint-Germain-en-Laye); but Magdalenian clay sculptures of a bear from the Montespan cave, Garonne, and copulating bison from the Tuc d'Audoubert, Ariège, approach the roundness of sculpture. Some of the stone reliefs are the only art attributable to the ill-defined Solutrean period.

Miniature carving of animals

In the Magdalenian period, miniature carving of animals predominated, particularly in western Europe. Archaeologically stratified ivory carvings have come from Stetten ob Lontal, Baden-Württemberg, Germany (wild horse, panther, mammoth; Institut für Vor- und Frühgeschichte, Tübingen), and from late Aurignacian mammoth-hunting stations in Moravia, where clay models accompanied ivory (mammoth, bear, wild horse, reindeer, owl). In French sites, miniature sculpture comes mostly from Magdalenian strata. These pieces, particularly those of horse and bison heads, surpass anything in eastern Europe. Particularly fine are a horse's head carved of reindeer antler from Le Mas d'Azil, Ariège, and a standing horse in ivory from Lourdes (both Musée des Antiquités Nationales). Some have harpoons engraved upon them and must have been used as hunting talismans; some of the best examples are carved on throwing sticks and other hunting implements.

In this category are carved finials of spear-throwers of Le Mas d'Azil in the form of a chamois (small, goatlike antelope; Péquart Collection, Saint Brieuc) or a black cock (Musée des Antiquités Nationales), handles such as one carved in the form of two crouching deer from Bruniquel, Tarn-et-Garonne (British Museum), and another in the form of a single reindeer from Laugerie-Basse. Perforated staves of ivory are sometimes carved with animal terminals; although known as *bâtons de commandement*, they were probably used as arrow straighteners. They are often covered with designs of deer, horses, birds, and fish.

The vast majority of Magdalenian finds are isolated engravings on bone, flat tines (pointed branches) of antler, small slabs of stone, or flat pebbles, executed some-



Head of an ibex, engraving on a bone staff, Magdalenian (*c.* 15,000–*c.* 10,000 BC). From Isturits, Basses-Pyrénées, France. In the Sainte-Périer Collection, Morigny, France. Height 13 cm.
J. Vertut—Ziolo

times with surety, sometimes as though they were practice pieces. Notable are Dordogne examples of hairy bisons' heads from Limeuil and Laugerie-Basse on bone and small limestone blocks. Unusual perspective and impressionistic effects were sometimes achieved by the Magdalenian artists. From Teyjat are experiments in group drawing: a herd of reindeer with only the outer animals drawn in detail, the herd effect achieved by central multiplications of legs and antlers. A herd of galloping horses is treated less summarily. A more detailed style is seen in antler engravings from Lorthet in the Pyrenees where, on a spear-thrower, reindeer are shown swimming. The hooves and body hairs are drawn with precision, as are details of salmon swimming between the beasts' feet. It is typical of a group of Upper Magdalenian miniatures in which the patterning of hair, skin, and fur forms a significant element.

Bone engravings of Pyrenean quality are not common elsewhere in Europe, but from the Pekárna cave in Moravia comes a masterly representation of bison in combat (Moravian Museum, Brno). Here too were found perforated staves parallel to the French type and bone pieces engraved with horse, bison, and chamois. Outline engravings of mammoths have been found at Kostienki I and Maltà, near Lake Baykal, Siberia. In these minor and applied arts of the Magdalenian period no interest is shown in scenic background or the human figure; the few examples of what appear to be human beings, including an engraved bone from La Colombière, France (Université de Lyons), are badly executed and difficult to discern. These minor pieces, though scarcely surpassed in treatment and skill, are greatly surpassed in scale by engravings and paintings in caves.

Paleolithic art: cave engravings and paintings. Geographical limitations confine cave art largely to southwestern France and northeastern Spain, the so-called Franco-Cantabrian province, but isolated finds come from southern Italy, southern Germany, and southern U.S.S.R. In the absence of firm chronology, it is difficult to reconstruct stylistic development, even in France.

Franco-Cantabrian cave art

Cave decoration was made either with stone gravers or with masses of liquid paint. While massive sculpture was usually placed in open rock shelters, paintings and engravings were usually hidden in deep recesses of caves where clay modelling was also practiced.

Cave art evolved from simple Aurignacian line engraving to late Magdalenian polychrome painting. Incised on hard rock surfaces, painted, or drawn in wet clay with a finger, the lines were originally of uniform thickness, later revised and modulated. Detailed drawing of pelt, muscles, and other anatomy tends to be zonal.

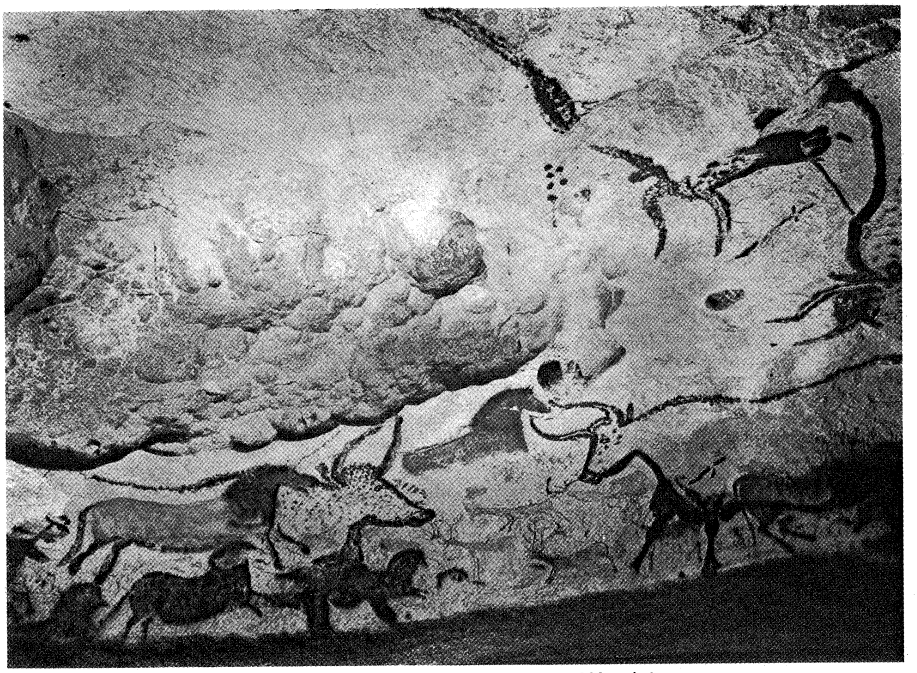
Painting and engraving often appear together in caves and probably came into use simultaneously. Distinctions in style suggest that painters and engravers practiced separate crafts. Only one known cave is decorated exclusively with engraved designs: Les Combarelles, Dordogne, has thousands of superimposed engravings from late Aurignacian through middle Magdalenian, suggesting that this cave was the centre of a hunting cult. Bison and reindeer are the main repertoire, but there are also grotesque part human, part animal figures, which are perhaps decoys or hunting magicians. Some of the engraved animals are lively, others stiff; chronological distinctions between the styles cannot be accurately made. From stratified caves, however, it would appear that Aurignacian style was close contoured and stiff and lacked the three-dimensional qualities of the later period. At Altamira, Teyjat (Dordogne), and Font-de-Gaume, engravings accompany Magdalenian wall paintings; it is possible to appreciate the plastic qualities of Magdalenian engraving and efforts to represent dimension, although the position of feet (often completely omitted) is embarrassing. Action is depicted at Font-de-Gaume, where a pregnant lioness is engraved ready to spring and horses canter in file. At Altamira leaping hinds and baying stags have lifelike qualities, antlers drawn in fine perspective. These epitomize the extremely flowing style of Magdalenian engraving, seen in plodding mammoths at Font-de-Gaume and some bears depicted at Les Combarelles.

Lascaux and Pech-Merle caves

At Lascaux, in a cave with passages and chambers on various levels, the best preserved traditions of later Aurignacian art survive; the Pech-Merle cave, Cabrerets, should be associated with it. Paintings in both caves are executed in thick black outline little refined. The bodies of some of the animals are washed in with red or brown paint, shading off toward the underparts. Most horses are painted in solid-black silhouette, but some horses, bison, and aurochs (extinct long-horned wild ox) in both caves are painted with an overall dot pattern. Friezes of horses and bulls traverse the lower edge of the arched roof, while the upper portions are left for depicting larger animals. Although the paintings are superimposed, at Lascaux there seems to have been a consistent attempt to decorate the roof in an orderly fashion. In other caves, animals are painted without significant relationships.

Altamira Cave remains pre-eminent in Magdalenian style. Natural bulges of the seven-foot- (two-metre-) high rock roof were painted with standing and squatting bison, thickly outlined in black and filled in with red, producing an effect of unparalleled beauty and realism. About 20 of them cover a distance of about 15 yards (14 metres). The rock formation was used to give a round effect in the body; the legs were painted from underneath (the alternative in the visual dilemma was to paint the legs from the side), giving the animals a slightly concave appearance countered by the rock surface. Font-de-Gaume Cave closely parallels the painted-in technique of Altamira and indeed, in parts, probably represents an earlier late Aurignacian style, but unfortunately it is badly preserved. The Niaux caves, Ariège, belong to the same black-outline tradition but do not have painted fillings. The outlines, however, are painted with thick strokes and adjusted with thin and are more carefully drawn than elsewhere; and the perspective of the feet of bison and wild horse is more successfully achieved.

Besides introducing painting as a pictorial technique, the primitive art of the Paleolithic begins an expressional and magical use of colour. Black, red, brown, and a dull yellow were chiefly used. Besides animals and anthropoid figures, areas and brush marks of colour and curiously shaped tectiforms (wedge-shaped markings perhaps representing huts) were sometimes painted in, as were colour-blotched outlines of stumpy human hands. Among the pigments at Altamira and Lascaux, natural oxides of iron and manganese supplied red and black; iron carbon-



Cave painting of bulls, horses, and deer, Aurignacian (*c.* 28,000–*c.* 22,000 BC), Lascaux, Dordogne, France. Photograph covers a span of about 9.14 cm.
Ralph Morse, *Time* (© 1959); Time Inc.

ate gave yellow; at Lascaux blue-black and dark-brown manganese oxides were found; other blacks came from soot. When finely ground, these pigments were mixed with water rather than a fatty base, allowing quick application to damp rock. Pigments were stored in such hollow receptacles as bones, stones, and skulls, as at Les Cottés, Vannes. Bone mortars and pestles for grinding paints and stone lamps for lighting dark recesses have been found. Fingertips, brushes, and daubing pads were used to apply the colour.

Function of the cave art

The purpose of this art is not clear. Executed far from normal gaze, the work was apparently done in great discomfort. That it was directed at the magical possession of favourable hunting conditions is occasionally indicated. An outline-painted mammoth in the Pindal cave, Asturias, Spain, has a red blotch indicating the position of the heart and a batch of arrows painted in front of it. Outlined bison in the Niaux cave, Pyrenees, have superimposed painted barbs pointing at the heart and midrib and blotches of red (blood?) on the body. Patchwork squares painted in front of such naturally curious animals as oxen and goats at Lascaux might represent pitfalls covered with sheets of skins. In the same cave is a scene in which a man with a broken spear is worsted by a bison. Comblike patterns painted near and on top of some animals might represent fenced enclosures to which they were lured. A human-shaped figure (known as the "sorcerer") wearing an antler mask, in Les Trois Frères Cave, Ariège, might be a decoy or a hunting magician. Hunting rites are evidenced by clay model animals, some with unmistakable signs of use as targets. Other models suggest magical increase of fertility (many painted animals are pregnant, and pregnant women are occasionally engraved on bone).

Toward the end of the Magdalenian period, naturalistic realism weakened and gave way to formalized arrangements of horns, antlers, and fur reduced to a pattern. Schematic female figures have been found from the earlier Paleolithic in Moravia; in the west, chamois and reindeer were reduced to formalistic ornaments. In western Europe, art was not to achieve such heights of abstraction until the appearance of formalized Celtic art of La Tène.

A branch of Franco-Cantabrian art is represented at Parpalló Cave, Valencia, where engraved bone pieces are stratified in Gravettian, Solutrean, and Magdalenian contexts; and where stones with animal paintings occur in the Solutrean. At La Pileta Cave, Málaga, a peripheral Magdalenian style of painting seems to have been introduced. The animals depicted, unlike those of France, are not Ice Age fauna but such forest dwellers as deer, wild boar, the smaller deer, and oxen; doubtless the southerly latitude is as much responsible for the difference as is the later date. The art of these two caves is possibly linked with engraved animals in caves in Liguria, Apulia, and west Sicily and may thus belong to an ill-defined "Mediterranean" province, possibly with African connections, which continued Paleolithic artistic tradition into Mesolithic times.

Mesolithic (Epipaleolithic) cave art of Spain. Another group of cave paintings, while difficult to date, can be largely attributed to the Mesolithic, or Epipaleolithic period. Found in southeastern Spain, they differ sharply in style and content from the paintings of the Franco-Cantabrian province. They incorporate an African style as seen in largely undatable paintings and rock engravings from Algeria, Tunisia, the Atlas Mountains, outer Sahara, and Bushman art in South Africa. These African examples, like those of southeastern Spain, are tribal rather than individual; hunting by small family groups in Paleolithic times had been taken over by organized males of the tribe. Spanish Levantine painted caves, discovered mostly in the early 20th century, are grouped in Albacete (Alpera and Minateda caves), in Teruel (Alcañiz and Tormón), in Murcia (Cogull and Yecla), in Valencia (Bicorp), in Castellón de la Plana, and in several shelters in the Valltorta gorge between Albocácer and Tirig. New paintings were brought to light at such sites as the Cueva del Polvorín, Castellón. The group has one outlyer, the cave of Las Batuecas between Salamanca and Cáceres.

Spanish Levantine caves

Engravings are very rare and the absence of stratified pieces of minor art makes the dating and culture attribution of the paintings difficult, but their beginning is probably to be associated with the Upper Paleolithic Capsian industry of southeastern Spain. While large-scale single animals are the rule in Franco-Cantabrian paintings, small-scale compositions of many figures were chosen by the southern Spanish artists. In most of the paintings, human groups play a large part, hunting single animals with bow and arrow. All figures are painted in solid silhouette; those of animals are true to nature while those of humans are highly distorted, with elongated bodies and thick legs. The colours are chiefly red, brown, and

black, largely derived from manganese and bound with blood, egg, or honey.

With tribalism came war, and many of the scenes depict lively arrow battles or mass execution of prisoners. Women, often drawn in long dresses, are shown dancing or walking with children. There is no evidence of fishing or trapping, but the dangers of the chase are vividly shown in Cueva Remigia paintings in which oxen and ibex turn

Instituto de Estudios Editoriales, Barcelona



Woman gathering honey, water-colour copy by F. Benitez Mellado of the original painting in the Cueva de la Araña, near Bicorp, Valencia, Spain, Mesolithic (*c.* 10,000/8,000–*c.* 3,000 BC). In the Museo de Prehistoria de la Diputación Provincial, Valencia, Spain.

on fleeing hunters. Little domestic life is shown, but a painting in the Cueva de la Araña, Valencia, depicts a woman gathering wild honey; one in the Gasulla gorge shows a spider surrounded by flies.

Paleolithic and Mesolithic art of northern Europe. A third group of Stone Age art, largely from the northern limits of Europe, while again showing a generic relationship with Ice Age art of western Europe, cannot be related to an Upper Paleolithic culture. In Finland, northern Norway (Finnmark, Nordland, Troms), central Norway (Trøndelag), and on the upper Yenisey River at Shalobolinsk in western Siberia, the custom arose of engraving scenes of hunting and magic on largely inaccessible slabs of rock or cliff surfaces. The linear style of Arctic reindeer, elk, and bear lacks Magdalenian dimensional qualities but is very definite and simple. More naturalistic representations and those with perspective as seen in French Magdalenian art are found only in remotest northern Norway and Siberia. Groups of animals and disguised human decoys are sometimes shown, and great interest in the sea is evidenced by engravings of halibut and seal; and, on the White Sea coast, by boats of a type ancestral to those in later Bronze Age engravings from Scandinavia. Although hunting scenes are not shown, depiction of inner parts of animals strongly indicates that hunting magic was the primary purpose of these drawings. For decorative purposes, superb small animal models were carved in Danish Maglemose (Mesolithic) settlements. (W.Cu.)

Linear style

NEOLITHIC AGE

The origins and history of European Neolithic culture are closely connected with postglacial climate and forest development. The increasing temperature during the period from around 8,000 to 5,500 BC caused a remarkable change in late glacial flora and fauna. Thus, the Mediterranean zone became the centre of the first cultural modifications leading from the last hunters and food gatherers to the earliest farmers. This was established by some important excavations in the Near East, which unearthed the first stages of early agriculture and stock breeding (7th and 6th millennia BC) with wheat, barley, dogs, sheep, and goats. Early pre-pottery Neolithic finds (probably 6th millennium BC) have been made in the Argissa Magula near Larissa (Thessaly, Greece), and excavations in Lepenski Vir (Yugoslavia) have brought to light some sculptures of the same period. The independent origin of European Neolithic was established, and it was thought highly probable that the cradle of farming in the Near East had not been the only one: there were others in Europe, too.

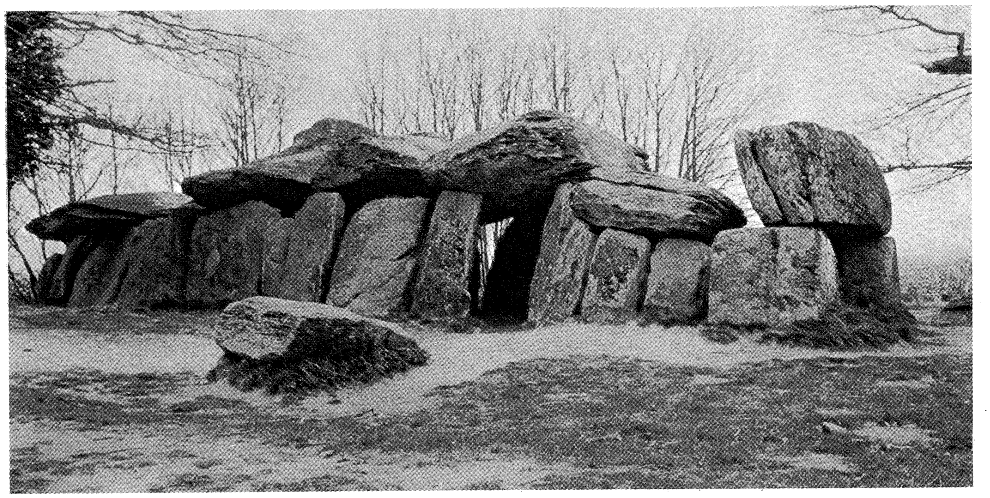
Early pre-pottery Neolithic finds

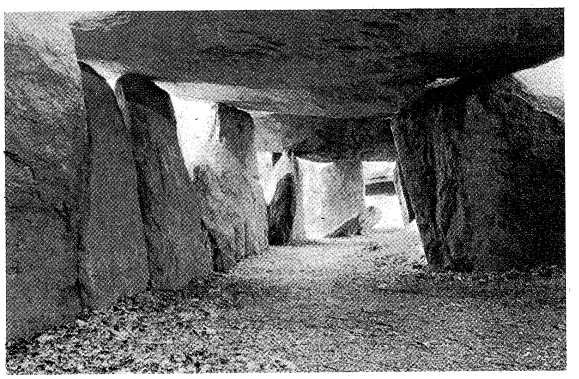
Neolithic farming in Europe developed on its own lines in the four different ecological zones: the Mediterranean zone of evergreen forest and winter rains; north of the Pyrenees, the Alps, and the Balkans, the temperate zone of deciduous forest and evenly distributed annual rainfall; still farther north the circumpolar taiga, or coniferous forest (the only zone to remain free of agriculture and stock breeding); and to the southeast the western end of the Eurasiatic steppes. Each zone itself is subdivided into natural regions by physiographic boundaries and peculiarities of climate or soil. Three major divisions of the temperate zone are western Europe, from the Atlantic to the Vosges and Alps and including the British Isles; the loesslands of central Europe, including the Ukraine and limited by the Balkans and the Harz; and the northern province, that portion of the Eurasiatic plain lying between the Rhine and the Vistula and including Denmark and southern Sweden. The substantial Neolithic communities that arose by the 6th millennium BC must have been largely recruited from indigenous Mesolithic hunters and fishers, attested to so abundantly by remains in western and northern Europe. (Some communities indeed seem to have been composed entirely of such Mesolithic stocks, though they had adopted a Neolithic equipment from immigrant farmers; such are sometimes termed "secondary Neolithic.") From these Mesolithic survivors, too, must be derived much of the science and equipment applied in Neolithic times to adapting societies to European environments. Upon the resultant distinctively European technology and economy was reared a no-less-original ideological superstructure expressed in distinctive sepulchral monuments, styles of pottery decoration, and fashions in personal ornaments.

Architecture. Dwelling places in Greece, Sicily, and the Iberian Peninsula were built of mud brick on stone foundations. But in the Balkans and throughout the temperate zone, wood was used for the construction of gabled houses, stout posts serving to support a ridgepole and walls of split saplings or wattle and daub (a framework of woven twigs covered and plastered with clay). The earliest houses on the loessland of central Europe were very large, up to 135 feet (41 metres) in length and large enough to accommodate a whole lineage or small clan together with stalled cattle and grain stores. In the sequel, these communal houses gave place to smaller two-roomed dwellings, 24½ to 33 feet (7.5 to ten metres) long but still entered through one end. Finally in late Neolithic times clusters of one-roomed huts became the most widespread fashion. Round the Alps such two-roomed houses, and, less often, one-roomed huts were raised on piles above the shores of lakes or on platforms laid on peat mosses. These are known as Swiss lake dwellings (*Uferrandsiedlungen*).

Swiss lake dwellings

In northern Europe, as in other sections of the continent, the earliest villages consisted of two parallel long communal houses, but these were subdivided into 20 or more apartments, each with a separate door. Here again the communal houses eventually broke up into free-standing one-roomed huts. Finally, the Stone Age village of Skara Brae on the treeless island of Orkney illustrates an ingenious adaptation of one-roomed wooden huts to an





Megalithic gallery grave.
(Left) La Roche aux Fées, Essé, Ille-et-Vilaine, France, Neolithic (*c.* 3,000–*c.* 1,800 BC).
(Right) Interior of La Roche aux Fées.
D. Lesec—Ziolo

inhospitable environment. Skara Brae also shows how commodiously such huts must always have been furnished, despite their limited size.

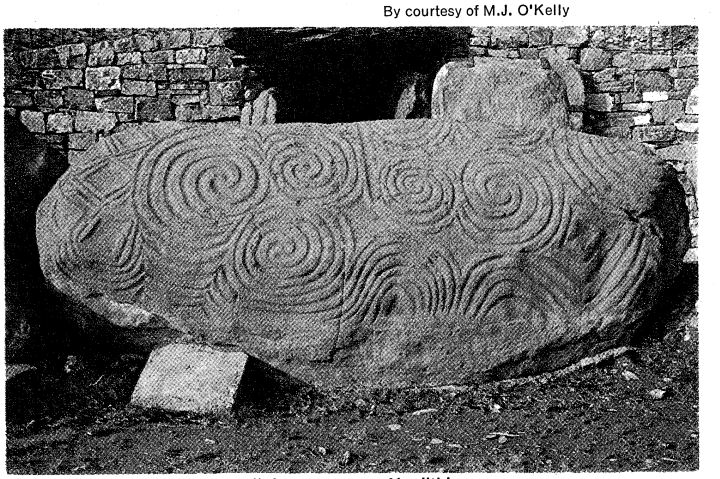
Megalithic tombs. Neolithic farmers were generally buried in a flexed or contracted position on the side except in northern Europe, where the corpse lay extended on its back, as was the custom also among the surviving hunter-fishers of the taiga and Russia. In Britain, Secondary Neolithic folk practiced cremation, burying the ashes in circular enclosures, one of which was destined in the Bronze Age to become the world-famous sanctuary of Stonehenge. But the most celebrated and imposing funerary monuments of the European Neolithic are collective tombs or family vaults, some built of huge undressed stones and thus rightly called megalithic, others walled with dry stone masonry and roofed by corbeling (arch supported by a bracket projecting from the face of the wall), others simply cut in the rock. Save for a cluster on the northern slopes of the Caucasus, such tombs are virtually confined to the Mediterranean coasts, the Iberian Peninsula, and western northern Europe and are most densely concentrated near the coasts. The simplest but not necessarily the oldest type of megalithic tomb, the true dolmen (consisting of four or more upright stones supporting a horizontal stone slab), is confined to northern Europe and was originally designed to contain a single extended skeleton. Passage graves, consisting of a polygonal chamber entered by a distinct passage and covered with a round tumulus, probably originated in Portugal and spread thence by seaways to Brittany, Britain, and Denmark.

Collective tombs

Finally gallery graves, which are long, narrow chambers with no distinct entrance passage, often covered by a long cairn (heap of stones) with a crescent-shaped forecourt at one end, are distributed from the Gulf of Lyons across France to Scotland and Northern Ireland. A special variant, generally subterranean, was developed in the Seine–Oise–Marne Basin and carried with a distinctive assemblage of pots, weapons, and personal ornaments by migrating bands to the Channel Islands and across Belgium and Westphalia to southern Sweden. (Ri.P.)

Chief expressions of Neolithic art. The migration of Neolithic agriculturalists from the Near East to Europe and the rise of European peasant communities brought the beginnings of a folk art expressed chiefly in weaving, basketry, and pottery making. Cave art was now confined to the engraving of simple magico-religious patterns on rocks, as at Newgrange and Lochcrew in Ireland, and standing stones at Morbihan in Brittany.

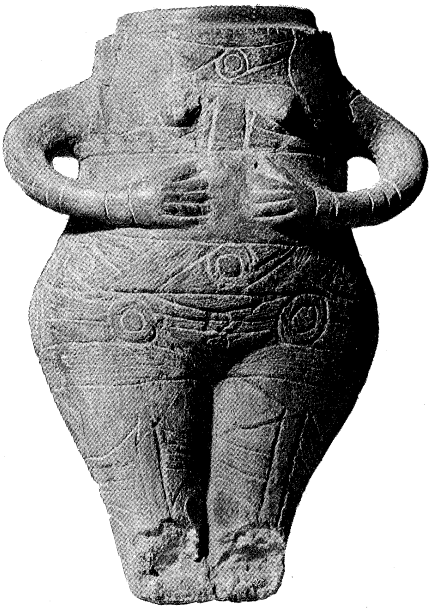
By courtesy of M.J. O'Kelly



Engraved rock with magico-religious patterns, Neolithic (*c.* 3,000–*c.* 1,800 BC), New Grange, County Meath, Ireland. Length 3.2 m.

From Anatolia the idea of pottery painting spread into southern and eastern Europe. A variety of techniques of painting and incision are known in the chief areas of Neolithic painted-pottery culture at Tripolye, U.S.S.R.; Cucuteni, Romania; Vučedol, Yugoslavia; and the Dimini and Sesklo cultures, Greece. In all of these, designs are abstract, often spiriform; the pottery tradition also includes figurines of women with painted or incised designs indicating elaborately embroidered clothing. Decorations

Pottery



Female figurine in the form of a jar, incised and with traces of paint, clay, Neolithic, *c.* 3,000 BC. From Vidra, Bucharest, Romania. In the National Antiquities Museum, Bucharest, Romania. Height 42.5 cm.
Holle Bildarchiv, Baden-Baden

on unpainted pottery from the rest of Europe show the predominant influence of basketry. In the process of perfecting techniques of shaping and polishing stone tools, the Neolithic craftsman brought to earlier Stone Age traditions a new combination of utilitarian and aesthetic feeling. (W.Cu.)

BIBLIOGRAPHY. Major modern studies of Paleolithic art should begin with the work of the pioneer, the Abbé HENRI BREUIL, *Quatre cents siècles d'art pariétal* (1952; Eng. trans., *Four Hundred Centuries of Cave Art*, 1952). His biography by ALAN H. BRODRICK, *The Abbé Breuil, Prehistorian* (1963), serves as an introduction to his work and method. There are two large and lavishly illustrated works on Paleolithic art: PAOLO GRAZIOSI, *L'Arte dell' antica età della pietra* (1956; Eng. trans., *Paleolithic Art*, 1960), which is indispensable and particularly good on the Mediterranean engraved caves; and ANDRE LEROI-GOURHAN, *Préhistoire de l'art occidental* (1965; Eng. trans., *The Art of Prehistoric Man in Western Europe*, 1967), which attempts a reclassification of styles and chronological evolution. Although the statistical evidence does not give strong support to many of the author's stylistic divisions, the work has great value both for its tabulation of motifs (particularly the abstract forms) and diagrams of cave layout. NANCY K. SANDARS in *Prehistoric Art in Europe* (1968), besides giving an excellent coverage of the works themselves, seeks to place the art of Paleolithic man in the context of his experience and spirituality. The author also treats technical problems of representation and perspective. The place of Paleolithic art in the total aesthetic experience of man is the subject of SIEGFRIED GIEDION, *The Eternal Present, Part One: The Beginnings of Art* (1962). Many caves have been given individual monographs and articles: reference should be made to the site index (for both caves and portable objects) in LEROI-GOURHAN above. Lascaux cave has been given detailed and well-illustrated treatment in FERNAND WINDELS, *Lascaux* (1948; Eng. trans., *The Lascaux Cave Paintings*, 1949); and GEORGES BATAILLE, *La Peinture préhistorique: Lascaux, ou la naissance de l'art* (1955; Eng., trans., *Lascaux; or, the Birth of Art*, 1955). For prospective visitors, ANN and GALE SIEVEKING, *The Caves of France and Northern Spain: A Guide* (1962), is an excellent companion. The pieces of Paleolithic mobiliary art in Britain (mostly from Cresswell Crags, Derbyshire) are described in DOROTHY GARROD, *The Upper Paleolithic Age in Britain* (1926). For Russia, EUGENE A. GOLOMSHTOK, "The Old Stone Age in European Russia," *Trans. Am. Phil. Soc.*, vol. 29, pt. 2 (1938), is one of the few source works in English.

The Epipaleolithic, or Mesolithic, art of peripheral regions is treated by SANDARS (*op. cit.*); more fully for Scandinavia by HERBERT KUHN, *Die Felsbilder Europas* (1952; Eng. trans., *The Rock Pictures of Europe*, 1956); and J.G.D. CLARK, *The Mesolithic Settlement of Northern Europe* (1936). For late Spanish cave painting the best illustrations are in RAMON MENENDEZ-PIDAL (ed.), *Historia de España*, ch. 3 and 6 (1963), where a bibliography of individual caves will be found. The fundamental work on Spain is JUAN CABRE AGUILO, *El Arte rupestre en España* (1915).

The instruments and pigments used by the Stone Age artists are treated by L.S.B. LEAKEY in ch. 7 of *A History of Technology*, ed. by CHARLES SINGER, vol. 1 (1954); and in more detail by KURT HERBERTS, *The Complete Book of Artists' Techniques* (1958), with interesting descriptions of experiments.

(W.Cu./Ri.P.)

Storage and Warehousing

Storage, an organized method of holding and protecting articles or commodities for future use, dates from earliest times, when it was involved in the preservation of agricultural products, later in commerce, and in increasingly complex ways with the introduction of mass production. Warehousing is the dynamic aspect of commercial storage; it includes material handling systems and control and information systems whose function is to facilitate the efficient placing of stored materials for convenient or scheduled retrieval.

ECONOMICS OF STORAGE AND WAREHOUSING

Functions of warehousing Commerce of almost any kind creates the need for ways to accumulate goods for subsequent distribution to the marketplace on the basis of fluctuating demand. Although the earliest emphasis in warehousing was doubtless on security from the weather, pests, and theft, in more recent years, increasing importance has come to be attached to the reservoir function of the warehouse in levelling out the cycles of commerce. This function applies particularly to (1) products such as crops, which are produced during a limited period but sold year round; (2) products such as seasonal or sports equipment, produced year round but consumed only during a limited period; (3) products for which demand is unpredictable, such as pharmaceuticals and drilling equipment; and (4) products that are manufactured in large production runs for economy's sake and dispensed over a long period. This is the most important single function of warehousing. If all products were manufactured as they were ordered, rather than held in large quantities for possible future sales, there would be little need for warehousing.

Transportation and warehousing. Like commerce, transportation is intermittent, or cyclical; a truck, railroad car, airplane, or boat is loaded and starts on its journey. Hours, days, or weeks may pass before another shipment is made; in the meantime, therefore, a means for accumulating the product in a systematic way is required so that when the next shipment is made the transportation equipment can be quickly loaded. At the receiving end each shipment equals a supply for a certain period and must be unloaded and stored systematically to permit delivery to users on demand. Warehousing accommodates transportation surges in the same way that it accommodates surges of production and market demand, and, as in its relation to the production cycle, warehousing reduces transportation costs, which also vary with volume; the more goods of particular type that are shipped at one time, the lower the unit transportation costs. Large numbers of small shipments over long distances incur high transportation costs; the availability of a warehouse permits the manufacturer to ship in large volume and then to dispense his goods locally from the warehouse.

Warehousing also permits economies through flexibility of transportation. Water carriers can move the largest volumes at the least cost, followed by railroad, truck, and air transport in that order. A manufacturer with sufficient volume can use warehousing to achieve the most economical combination while, at the same time, satisfying other requirements.

Place-time values. Although neither transportation nor warehousing adds value to products in the sense that manufacturing does, they create equivalent values (called place-time values) in the marketing process by facilitating the delivery of goods when and where they are need-

ed. Equally significant are the place-time values provided by warehousing in the production process. Raw materials, components, and subassemblies must be fed to production lines in a steady flow. To shut down production because of the unavailability of materials is extremely costly, and the warehouse gives insurance against this contingency.

The distribution centre. Although the warehouse as a marketing channel—a place to store goods pending their sale—is as old as commerce itself, the emergence in the 20th century of numbers of companies manufacturing at multiple points and distributing nationally or internationally has given rise to a new type of warehouse called a "distribution centre." Although the typical worldwide multiplant firm seldom if ever manufactures its complete product line at all of its plants, it is usually able to sell to its customers a selection that includes products from all of its plants. It is the distribution centre that makes this feasible. A manufacturer of appliances, for example, may produce stoves at one location, refrigerators at another, washers and dryers at a third, and small appliances at a fourth. Since dealers seldom order a single item in substantial volume lots, if plants at each location were to ship directly to a single dealer, the high cost of the multiple small shipments from the different plants would be a serious economic disadvantage. Given a number of dealers, however, the manufacturer can build a distribution centre at a point intermediate to his plants and principal market areas, shipping to this distribution centre volume shipments of the individual items for later mixing into specific orders.

A factor in this practice is the provision in most carrier rates that permits storage in transit, so that a single origin-to-destination rate, substantially lower than the rate for two separate shipments, is applied.

Forecasting

An important element in exploiting the mixing function of the distribution warehouse is accurate forecasting of market requirements, so that the right quantities can be shipped from each of the plants, assuring the distribution centre of adequate stocks on hand to fill customer orders. As in any warehouse, the risks and costs of not having all products in stock at all times must be weighed against the risks and costs of having an oversupply and paying the penalties of obsolescence and deterioration, as well as the cost of the products themselves and the warehousing expense of holding them.

Warehouses as either mixing points or distribution centres play a large role in chain-store retail operations, receiving products in volume from many different manufacturers and assembling orders for their different stores. A distributor's or a wholesaler's warehouse performs a similar service for a number of customers, while a public warehouse may serve as a distribution centre for many manufacturers. One advantage of the public warehouse is that by handling the merchandise of a number of different manufacturers it can consolidate orders for products destined for a single buyer and so obtain lower delivery costs. The wholesaler or distributor who handles products of different manufacturers has a warehousing operation that is similar to the public warehouse, with the important exception that the wholesaler can control the flow of goods into the warehouse.

Although the warehouse-distribution centre is thought of generally as handling finished products, it may also serve an important function in stocking parts and maintenance supplies. In the automotive industry, for example, the warehouse handles only parts and supplies, while the assembled product (automobiles) moves directly to the distributors or dealers. The ready availability of parts in the warehouse is important in the case of appliances, office equipment, and machinery of all types. When equipment is rented or leased, rather than sold outright, the ready availability of parts is essential to keep machines operative.

Large-scale warehousing of operating supplies is useful in large business firms. Central stockpiling and distribution of such supplies as stationery, printed forms, and office supplies may involve warehousing for both economy and efficiency.

Custody storage and warehousing. In contrast to the dynamic warehouses used in production and marketing are the warehouses used primarily to hold goods in long-term storage. The primary function of such warehouses is to protect the contents from the elements and from loss, theft, or deterioration. Depending on the products to be stored, warehouses may range from general purpose buildings to highly secure, climate-controlled structures. Since the emphasis is on capacity rather than on product movement or accessibility, and goods in custody storage may remain in the same place for months or even years, warehouses of this type need relatively little materials handling equipment and few permanent employees.

Custody storage generally involves household goods, high value goods, and business records. A bank vault is essentially a high-security storage facility; the term "repository," originally applied to banks, is also used to denote public warehouse facilities that specialize in records retention and, if required, phased destruction under secure conditions. The production and marketing applications of custodial storage are limited but do exist: in the case of a high volume production item (*e.g.*, canned food) during a limited season, the goods may be moved into custody storage for some months before going into normal marketing channels. Similarly, certain agricultural products may be held in custody storage in anticipation of price improvement or, in some countries, as part of government programs.

Since the basic requirement for warehousing is protected space, almost any warehouse is likely to contain some dynamic and some static storage. Dynamic warehousing activity, however, tends to take place in more modern and more expensive buildings designed for high turnover, while static, or custody, storage is more likely to be conducted in older buildings.

Alternatives to warehousing. Broadly speaking, a product is either in use or in storage. If it is being transported from the manufacturer to the buyer, from the manufacturer to the warehouse, or from the warehouse to the buyer, it is, in effect, being stored in a "warehouse on wheels." By definition, storage is the setting aside of goods for future use and is not limited to placement in a fixed structure. Since warehousing in many cases is a substitute for transportation, it follows that under some circumstances transportation may be an economical substitute for warehousing.

Transportation as an alternative

One of the principal attractions of transportation as an alternative to warehousing is its availability on an as-needed basis, without the necessity of large investment in facilities or commitment of stocks. In the chemical industry, for example, prompt response to customer orders is important for competitive reasons. Chemical firms can load railcars with material prior to receipt of orders and consign these carloads to themselves at distant points as "rolling warehouses." When an order is received, the firm can then check through computer connections with the railroads to determine the location of a carload moving in the customer's area. Through these same connections they can issue instructions to "divert and reconsign" the carload shipment to the customer, thus providing fast response without the additional expense of regional stockpiling.

Similar practices are followed with lumber and produce. Firms on the west coast of North America, for example, regularly consign uncommitted carloads to themselves at eastern points and, as markets change, divert the carload shipments to areas in which they can command the highest prices.

With products of higher value, truck or air shipment may provide a practical alternative to regional warehousing. Although air freight was formerly considered applicable mainly to "emergency" shipments, in recent years it has been recognized that the high cost of air freight may be more than offset by savings in warehousing, inventory, packaging, and insurance when goods of relatively high value are moved by surface carriers. The electronics industry regularly uses air shipment as a substitute for warehousing, since its products are generally small and have high value. Similarly, products that are perishable

and have a time value, such as expensive cut flowers, frequently move by air. In international trade, air freight now occupies an increasingly important place in commerce in electronics, high-value automobile replacement parts, and similar items.

On-site manufacturing

Another alternative to regional warehousing is on-site manufacturing or assembly. This approach is useful, for example, in the automotive industry, in which the regional assembly plant makes possible a limited storage of parts and components, with a high volume of inbound movement of these items and a similarly high volume of outbound movement of assembled cars consigned to distributors and dealers.

Finally, careful scheduling of production and close control of inventories can also provide a partial alternative to warehousing, particularly if there is also an effective coordination with transportation facilities.

In practice, freight cars and truck trailers have been widely used for short-term storage by receivers, who find that the charges assessed by the carriers for retention of their equipment (called demurrage charges) are usually less than the cost of building or leasing space. Carriers generally oppose such use of their equipment because the return from demurrage is substantially less than that received when the equipment is in actual transportation service. There has been some interest in the development of demountable containers, which permits short-term storage without tying up rolling stock. Some buildings have even been designed so that such containers may be physically attached to the outside and entered directly from the building, thus providing miniature warehouses from which parts and supplies can be drawn.

Factors influencing number, location, and size of warehouses. As the scale and geographical range of a business increase, so do the scale and range of its warehousing needs. Manufacturers distributing on a wide or international scale, generally warehouse in multiple locations. Firms operating multiple locations, either plants or retail outlets, develop their warehousing patterns accordingly. Most manufacturing plants require on-site warehousing to serve as a buffer or reservoir between production and shipping. It is generally impractical to attempt to manufacture goods in such a way that they may be moved directly from the end of the production line to the transportation equipment. Typically, a manufacturing firm must produce a large quantity of one item before changing its production lines over to another, and the plant warehouse functions as an interim reservoir, enabling shipment of goods in mixtures specified by customers or regional warehouses.

If the plant manufactures the firm's complete product line, it is likely that the plant warehouse also serves as a distribution warehouse for the area. If it does not manufacture the complete product line, its plant warehouse serves primarily as an intermediate point in the movement of goods to regional distribution warehouses, where they are mixed with other products from other plants.

In high volume, consumer-oriented businesses, the pattern of warehousing that emerges will be determined primarily by the location of plants and markets. Usually it will be the case that adequate service can be provided from a limited number of market-oriented warehouses located where there are good transportation facilities to and from plants and customers. But cost factors are so complex that careful studies will usually be undertaken to determine the exact number and location of such facilities. Such a study normally has two phases: first, determining the range of possible locations that would satisfy the customer service requirement, and then narrowing that range to the actual number and precise locations at which this requirement can be met at least total cost. The second step involves a study of transportation costs in comparison with warehousing costs, and may be performed by computer simulation techniques, by which a computer model is developed to represent the firm's distribution and used to test various combinations of transportation and warehousing.

Warehousing location study

While transportation is a major consideration in selecting a warehouse location, labour supply, power, and other factors that are important in locating manufacturing plants are only minor considerations. Terrain is a factor only as it affects access and construction costs. Climate may be important, since heating or cooling the typical warehouse is likely to be inefficient and costly. Because warehouse labour productivity drops sharply in extremes of climate, warehouses in cold climates are often equipped with special door-area heaters, while warehouses in warmer regions increasingly are air-conditioned.

The size of a warehouse is essentially a function of anticipated volume plus an allowance for growth. Operating efficiency tends to increase with size, within reasonable limits. Because of the dynamic character of warehousing, considerable technique is required to arrive at a precise determination of capacity requirements. A study may begin with measurement of the package sizes to be handled by the warehouse, proceeding next to a determination of the most efficient unit load dimensions and floor space requirements, with consideration of different storage alternatives such as floor stacking, racking, and binning. Computers have been applied to such studies. A final factor influencing the number, location, and size of warehouses is the availability of alternatives, such as flexible and economical transportation.

Public versus private, or captive, warehousing. A public warehouse offers warehousing services to the general public; a private, or captive, warehouse is maintained by a manufacturer for his exclusive use.

Operationally, a public warehouse functions much like its private counterpart. Goods are received in volume, placed in storage, and dispensed as orders are received at the warehouse. But because the public warehouse usually serves a number of stores that sell to the same customers, it can achieve important economies both in assembling orders and in consolidating deliveries or shipments. Each of four manufacturers selling to a given retail outlet might, for example, have an order amounting to a quarter truckload, in which case if the goods are stored in a public warehouse, the public warehouseman can consolidate all four orders into a single shipment at a lower rate.

Fungible goods

Such economies become even more significant in the case of so-called fungible goods; *i.e.*, products like grain and certain chemicals that have identical specifications and quality standards and are stored together without differentiation as to ownership. Thus a liquid chemical produced by manufacturer "A" may be stored in a tank along with chemicals of the same standard produced by manufacturers "B" and "C." The customer of "A" receives a product that is a mixture of the products of manufacturers "A", "B," and "C," but since the quality is identical, this is of no concern, and joint storage and economy of scale allow a lower price to be set than would otherwise be possible.

Although branded goods stored in a public warehouse are not fungible in the sense of standard commodities, a public warehouse may undertake to maintain in joint storage all products of a manufacturer placed in storage and subsequently sold but not removed from the premises. The public warehouse does not physically segregate the merchandise sold to retailer "A" from that sold to retailer "B" but merely insures that an adequate inventory is on hand to serve both as store orders come in. By eliminating the intermediate order picking and shipping directly to the stores of a chain rather than to its distribution centre for rehandling, the public warehouse can offer important advantages to both the storer and his customers.

From the user's point of view, another major advantage of public warehousing is its flexibility and availability for use on an as-needed basis. This is particularly attractive to the firm with highly seasonal items or fluctuating storage requirements, where facilities of sufficient capacity to accommodate the storage peaks would be seriously underutilized during storage valleys. The somewhat higher storage and handling costs involved in using a public warehouse are usually more than offset by the savings realized by not investing in an inefficient facility.

Public warehousing is also advantageous in test marketing in a single country or region. The availability of public warehousing makes it possible for a manufacturer to maintain an adequate supply of his product to meet the demand generated by advertising and promotion in the test market without a permanent commitment to large fixed facilities.

Many public warehouses perform a number of so-called accessorial services in addition to those directly associated with handling, storing, and documenting the goods of their customers. Among these are bottling, bagging, and drumming from products shipped in volume; repair and reclamation of damaged products, recall, storage and disposition of defective or contaminated products; equipment servicing; provision of office and display space; and performance of export-import services. Since these services, too, are available on an as-needed basis, the storer who requires them often finds it more practical to use the services of a public warehouse than to maintain his own staff and facilities.

Financial services available at public warehouses

A number of useful financial services are also often available from public warehousing firms. Important among these are loans on goods held in storage and "field warehousing," in which a public warehouse firm takes custody of goods at the manufacturer's location so that they may be used as collateral for a loan without having to be physically removed. A number of public warehousing firms have entered the business of financing and constructing warehouses to be operated by manufacturing firms as captive facilities, in some instances even assuming responsibility for their operation.

Though many firms find public warehousing advantageous, either as an alternative to operating their own facilities or as a supplement to existing captive warehouses, it must also be recognized that there are some disadvantages. One of these is curtailed control over the physical handling of their products and over the customer service function. While the public warehouseman may in actual fact perform both of these activities as well as the firm itself, the storer must compete for both service and attention with other customers of the public warehouse; he must recognize that if he is a small or intermittent customer his products and his customers may not get the same care and attention they would if he operated his own warehouse.

Advantages of private warehouses

An advantage of the private, or captive, warehouse is the degree of control the manufacturer can exercise over its operations. This is particularly important when the warehouse operation must be coordinated with both plant operations and customer requirements. If the product requires special handling or storage equipment, protection or climate control, the firm may have to operate its own facilities simply because there are no adequate public facilities available. For the retail firm, private warehousing is usually both desirable and necessary; by controlling both the warehousing and the actual distribution of its products and their movement to sales areas in the stores, the firm can exercise controls and achieve economies that would not normally be possible in a segmented operation using public warehousing.

A disadvantage of private, or captive, warehousing is that costs are frequently high because of periodic underutilization of personnel and facilities and fluctuating volumes. Too much emphasis on customer service may burden the warehouse with an excessive number of "emergency" orders resulting in high levels of overtime during some periods and equally high levels of idle time at others. Some management problems are inevitable; the firm that operates its own warehouses is in effect in the warehousing business on a relatively small scale, but with management problems comparable to those it would have were it in the warehousing business on a large scale.

TECHNOLOGY OF STORAGE AND WAREHOUSING

Warehouse design. The technology of warehousing is undergoing important changes in both building design and handling and storage equipment. Although most modern warehouses are of single story design, there is increasing recognition that these are inefficient in land use and impose a cost penalty in interior travel time; *i.e.*, the excessive amount of horizontal travel required because goods are spread out over a large area. Typically, in the past the multistory warehouse was handicapped by floor load limitations on upper stories, low ceilings, and slow elevators. In some instances these disadvantages have been partially offset by the installation of faster elevators with automated loading and unloading eliminating the bottlenecks of the older system. In modern structures there is increasing emphasis on high-ceiling warehouses permitting storage to heights of 20 feet or more for forklift trucks, and much higher for stacker-crane systems. Western European warehouses have pioneered the high-ceiling concept, with storage heights in certain West German installations of 95 feet (28 metres), and one in Switzerland of 103 feet (31 metres).

Building shape is an important design element in both construction cost and operating efficiency. A building 150 feet by 200 feet (45 by 60 metres) has the same floor area as a building measuring 100 feet by 300 feet (30 by 90 metres), but its perimeter measures only 700 feet (210 metres) as compared to the second building's 800 feet (240 metres). Besides adding a construction-cost penalty, the shape of the second building would also increase interior travel time and make supervisory control more difficult by dispersing personnel and operations over a wider area. Heating and air-conditioning costs are greater in long, narrow buildings, and generally the shape does not lend itself to efficient arrangement of aisles and storage slots. Buildings of irregular shape, built to fit into available land areas, are equally inefficient, but they may be necessary if location is more important than operating efficiencies. Refrigerated space usually costs three to four times as much as ordinary space.

Safety considerations

Since protection of contents is a paramount consideration, most modern warehouses employ fireproof materials and sprinkler systems. Lightweight concrete, fireproof and susceptible to prefabrication, is widely employed. This material also has good insulating and acoustical properties. Steel buildings, shop-fabricated and field-assembled, are also widely used. Although clear span construction is an asset in warehousing because it permits maximum use of floor space without obstructions, most warehouses are too large in floor area for clear span construction to be economically feasible; so, instead, conventional column supports, spaced to approximate the dimensions of storage bays and aisles, are utilized.

An important design question is the number of truck doors and railcar positions. Besides adding to construction cost, a large number of doors also adds to heating costs and creates security problems. There is a trend in more recent buildings to provide fewer door positions and improve their utilization by better scheduling of transportation equipment and mechanized loading and unloading. One large distribution centre warehouse—by scheduling trucks in at 20-minute intervals—is able to handle all shipments through a single truck door. There is also a trend to enclosed shipping and receiving docks for both railcars and trucks, in which equipment is loaded and unloaded inside the building, an advantage for both climatic and security reasons.

Because efficient layout and space utilization are critical to economical warehousing operations, warehouse design is based on forecasts of requirements and determination of the aisles and the shapes and volume of goods the warehouse will be handling. This may start with a study of package dimensions and determination of the most efficient unit loads, with stacks arranged in a specific pattern on a pallet or banded together, for handling by forklift trucks. Unit load dimensions of high-volume items, plus the quantities to be handled over a period of time, provide the basis for determining bay size and general storage plan, as well as overall dimensions and cubic capacity of the warehouse. Handling and storage methods to be used must also be considered at this juncture, since they will influence such layout features as aisle placement and width, storage racks, bins, and work areas.

Warehouse equipment. Two principal activities that are common to all warehouses are handling and storage.

Handling includes the steps involved in moving goods to and from storage, while storage is the condition of the goods at rest in designated areas of the warehouse. Major handling activities include unloading goods from railcar or truck; sorting; placing on pallets; moving by truck, forklift, or conveyor; placing goods in storage racks, slots, bins, or shelves; picking or assembling items ordered from storage positions; moving goods to the shipping area; checking; and loading goods into railcar or truck.

Non-automated equipment. Many mechanical devices are used to perform handling functions. The most common are the two- or four-wheeled hand truck, or cart, and the electric, gasoline, or liquid-propane forklift truck. The forklift truck, the workhorse of the warehousing industry, can transport loads both horizontally, for movement to and from the storage area, and vertically, to position and retrieve loads in upper positions of the storage system. The forklift is essentially a tractor with a mast and powered lifting mechanism that raises or lowers two flat prongs, or forks, running on tracks on the mast. The operator positions the forks underneath the pallet, or skid, by inching the truck forward. He then raises the forks until the load is clear of the floor and drives to the storage position where the load is raised or lowered as required, and placed in position. This completed, he backs the truck away and clears the forks, lowers the forks, and is ready for the next operation.

There are numerous attachments for forklift trucks: hooks, rams, clamps, suction devices, push-pull devices, side-shifters, drum-handling clamps, and other devices for special handling situations. Forklift trucks used in warehousing operations are generally in the range of 2,000 to 6,000 pounds' (900 to 2,700 kilograms') capacity.

Four-wheel carts are also widely used; the worker pushes the truck through the storage area, picking items from their storage positions and placing them on the truck. In some cases, these trucks are assembled into a tractor train and in still others are moved from the storage area to the shipping dock by a towline system, generally an underfloor chain that engages a vertical bar on the truck through a slotted track in the floor and switches or releases the truck at predetermined positions on the shipping dock. Fixed conveyers with wheels, rollers, or belts are also used for moving items in cartons; they operate either on gravity principles or with power drives, or a combination of the two. Special portable conveyors are sometimes used for loading and unloading operations.

Pallet stacking

The basic storage system in warehousing involves stacking pallet loads one on top of another. This has certain disadvantages, however; the weight of a high stack may crush packages on the bottom; stock rotation is difficult, because the oldest merchandise is on the bottom; and space utilization is poor because loads can seldom be stacked to the full available height. To overcome these problems, many warehouses use storage racks that permit placement or retrieval of a pallet load without disturbing neighbouring loads. These pallet racks consist of rails or track, welded to uprights, that support the edges of the pallet, keeping the pallets vertically separated. Drive-in space permits merchandise to be stored in depth without preventing access to the loads in the rear. Bins are typically used for small items, and shelves for larger and odd shaped items.

An important development in storage systems has been the gravity flow rack. This is similar to the conventional storage rack except that the rails are at a 3 to 4 percent grade and incorporate wheels or rollers. A pallet load placed in the rear of the rack rolls forward each time a pallet is removed from the front. A single line in the flow rack may hold 20 or more pallet loads, front to rear, and a typical flow-rack assembly may contain 75 or more such lines, allowing a total storage capacity of 1,500 or more pallet loads. As a pallet load is removed from the front of its line, another moves automatically into the forward position. This arrangement permits orders of merchandise to be assembled from the front of the system simultaneously with restocking operations at the rear. It also provides excellent space utilization by reducing aisle space requirements, and it ensures stock rotation.

Automated equipment. Automated warehousing using stacker cranes in combination with gravity flow racks is another important development in high-volume storage operations where space is at a premium. The stacker crane operates on principles similar to a forklift truck but moves by remote control in a fixed path by means of guide rails in the floor and on the warehousing ceiling, either to feed pallet loads into the gravity flow rack system from a production line, or to retrieve pallet loads and move them to the shipping area, or both. Feeding and retrieval lanes are preselected by an operator working at a console equipped with a card reader or computer that directs the crane to the proper location for pickup or delivery, and from there to its next assignment. A modern stacker-crane operation, operated by a brewing company in Kyōto, Japan, handles up to 2,400 pallets a day, each pallet loaded with 24 cases, in a warehouse with a total capacity of 8,748 pallets. This rapid turnover is made possible by the efficiency of ten stacker cranes, divided between two sides of the warehouse, one for receiving and one for loading.

Stacker-crane operation is adaptable to a wide variety of stored products; in Gütersloh, West Germany, a warehouse was specially designed for books. Its capacity of 7,000,000 volumes makes it one of the world's largest book repositories. The stacker crane is suspended from the rack framing. The system is particularly well adapted to refrigerated warehousing where space utilization is an important factor because of high construction costs, and where the work force can only work in the refrigerated area for short periods.

For smaller units of shipment, such as individual packages, automated warehousing takes the form of conveyor systems that divert packages at predetermined points to conveyor spurs, where they meet other packages drawn from other parts of the warehouse but consigned to the same buyer or scheduled for shipment on the same transportation equipment. The packages may be diverted by remote control by an operator visually inspecting each package as it passes him, or the system may employ a sensing or optical-scanning device that interprets package labels and activates the diverting mechanisms.

Computerized operations

Most automated warehousing systems employ some level of computerization. A typical example is a chemical-storage warehouse in Monthey, Switzerland, in which nine stacker cranes serve 20,000 pallet positions, with an annual handling volume of 120,000 pallets. Papers for inbound pallets are transmitted by television from control desks in the office to the control room for computer recording; the computer gives instructions to conveyors and to the stacker cranes and retains control of positioning and future movement of the pallets. A pharmacy warehouse at Nottingham, serving a chain of 1,300 drugstores in the United Kingdom, stores products according to size and turnover frequency in three differently equipped departments. The department for full cartons with high turnover is the most extensively automated, with a computer-controlled order-picking tower that controls dispatching, stockkeeping, and sorting of products according to customers' orders.

Although automation is still relatively expensive, it has proven to be substantially less expensive than conventional methods in a number of installations when the total cost of the installations has been taken into account. By permitting high-density storage, with loads stored both higher and deeper than is possible with manual methods, stacker-crane systems reduce warehouse space requirements by as much as 40 percent below that required by conventional warehouses, crisscrossed with aisles. Certain conditions must exist before automated warehousing is economically feasible; these include high construction costs, a high and relatively stable level of handling volume, and a limited range of package sizes. One disadvantage of automated systems is their relative inflexibility. They cannot readily be adapted to changing conditions in the warehouse in the way work crews can be shifted about, reassigned, or taught new skills. If an automated system reaches capacity, problems that arise cannot be solved by scheduling overtime or increasing the work force. Yet if an automated system has too much excess

capacity the added cost makes its operation uneconomical. Technological advances are overcoming many of these problems, however.

Terminaling

Other forms of warehousing and storage. Similar in principle to conventional warehousing, although employing somewhat different technology, is the storage of bulk products in either liquid or dry flowable form. Intermediate storage of these products comparable to warehousing is usually referred to as terminaling, because the point of storage is usually served directly by rail or water or both, and frequently by pipeline. Dry products requiring protection are typically stored in silos or elevators, while others are stored on the ground outdoors. Because of the high costs of ships, barges, and specialized rail equipment used to transport such products, most storage locations are equipped with high-speed loading and unloading systems employing conveyors and gravity and pneumatic devices. Some export coal-loading points are equipped with massive devices that grasp a railcar loaded with coal and turn it completely upside down, emptying the contents into a high-speed conveyor feeding directly into the ship's hold.

Liquid products are typically stored at "tank farms," which are either public or private operations. Often the tank farm is the terminus of a pipeline and serves as the distribution centre for the region, using tank trucks for final delivery. The use of pipelines for transporting an increasing range of liquid products, plus some limited success with slurried products (solids, such as coal, carried in a liquid), has caused a significant growth in this type of storage.

Household-goods storage follows the same general line as merchandise storage, although because of its static nature there is limited need for sophisticated handling or storage equipment. Often such goods placed in storage are put into pallet vaults—a large box with a pallet base, which both protects the contents and permits efficient handling and stacking by forklift truck.

Storage and warehousing occupy an increasingly important position in the modern, mass-production-oriented economy. Although the technology of warehousing has, in general, lagged behind production technology, sharp increases in the cost of land, buildings, and labour—plus increased dollar value of products—have led in recent years to the application of a systems approach, both to the operations of the warehouse itself and to its unique function in the procurement-production-distribution cycle underlying all manufacturing activity.

BIBLIOGRAPHY. C.H. JENKINS, *Modern Warehouse Management* (1968), is one of the few textbooks devoted to this field. Another is A.J. BRIGGS, *Warehouse Operations Planning and Management* (1960). An excellent chapter on warehousing is included in R.C. COLTON and E.S. WARD, *Practical Handbook of Industrial Traffic Management*, 4th rev. ed. (1965). W. BLANDING, *The ABC's of P.D.M.* (1967), is a booklet explaining warehousing and other physical distribution disciplines in layman's language. CONTROL FLOW SYSTEMS, *Economics of Automated Warehousing* (1970), examines the cost factors in automated versus conventional warehousing in an objective fashion. N.L. ENRICK, *Inventory Management* (1968), is a concise explanation of the inventory concepts and techniques underlying modern warehouse management. E.R. SIMS, JR., *Planning and Managing Material Flow* (1968), is a comprehensive treatment of industrial engineering approaches involved in the handling and storage of goods.

(W.Bl.)

Strabo

The work of Strabo, the Greek geographer and historian, is a major source for contemporary knowledge of antiquity. His *Geography* is the only extant work covering the whole range of peoples and countries known to both Greeks and Romans during the reign of Augustus (27 BC–AD 14). Its numerous quotations from technical literature, moreover, provide a remarkable account of the state of Greek geographical science, as well as of the history of the countries it surveys. Yet, though chiefly a compiler, Strabo left his personal imprint upon his work through his shrewd criticism of the sources he used. Other features of his personality were a keen sense of the practical usefulness of scientific knowledge (under the influence of Stoical teaching) and his admiration for Roman policies.

Formative influences

Born in 64 or 63 BC in Amaseia (now Amasya), the former capital of the kings of Pontus, Strabo belonged on his mother's side to a famous family, whose members had held important offices under Mithradates V (around 150–120 BC), as well as under Mithradates the Great, the opponent of Rome (132–63 BC). His first teacher was the master of rhetoric Aristodemus, a former tutor of the sons of Pompey (106–48 BC), in Nysa (now Sultanhisar in Turkey), on the Maeander. He moved to Rome in 44 BC to study with Tyrannion, the former tutor of Cicero, and with Xenarchus, both of whom were members of the Aristotelian school of philosophy. Under the influence of Athenodorus, former tutor of Octavius, who probably introduced him into the future emperor's circle, he turned toward Stoical philosophy, the precepts of which included the view that one unique principle ceaselessly pervading the whole universe causes all phenomena.

It was in Rome, where he stayed at least until 31 BC, that he wrote his first major works, his 47-book *Historical Sketches,* published around 20 BC, of which but a few quotations survive. A vast and eclectic compilation, it was meant as a continuation of Polybius' *Universal History*. The *Historical Sketches* covered the history of the known world from 145 BC—that is, from the conquest of Greece by the Romans—to the Battle of Actium (31 BC), or to the beginnings of the principate of the Roman emperor Augustus (27 BC).

In 29 BC, Strabo visited the island of Gyaros (today known as Yiáros, or Nisós) in the Aegean Sea, on his way to Corinth, Greece, where Augustus was staying. In 25 or 24, together with Aelius, prefect of Egypt who had been sent on a military mission to Arabia, he sailed up the Nile as far as Philae. There are then no further references to him until AD 17, when he attended the triumph of the Roman general Germanicus Caesar (15 BC to AD 19) in Rome. He died shortly after AD 23, having devoted his last years to compiling his second important work, his *Geographical Sketches*. Judging by the date when he wrote his personal notes, he must have worked on the book after his stay in Egypt and then have put it aside from 2(?) BC to AD 14, when he started the final edition, which he brought to an end in AD 23.

The first two books, in effect, provide a definition of the aims and methods of geography by criticizing earlier works and authors. Strabo found fault with the map designing of the Greek scholar Eratosthenes, who lived from *c.* 276 to *c.* 194 BC; Eratosthenes had combined astronomical data with coast and road measurements, but Strabo found his work lacking in precision. Although Strabo closely followed the treatise against Eratosthenes of the Greek astronomer Hipparchus, who had lived in the 2nd century BC, he blamed Hipparchus for neglecting the description of the Earth. On the other hand, he appreciated Polybius, who, in addition to his historical works, had written two books on European geography that Strabo admired for their descriptions of places and peoples. Although he praised Posidonius, the Greek historian and philosopher who lived from about 135 to 51 BC, for his knowledge of physical geography and ethnography, he rejected Posidonius's theory of climatic zones and particularly his hypothesis that the equatorial zone was habitable. This critical study led him logically to decide in favour of a descriptive type of geography, based on a map with an orthogonal (perpendicular) projection. The problem of projecting the sphere on a flat surface is not dealt with at any length, for his work, as he said, was not designed for mathematicians but for statesmen who must know countries, natural resources, and customs.

In books III to VI, Strabo described successively Iberia, Gaul, and Italy, for which his main sources were Polybius and Posidonius, both of whom had visited these countries; in addition, Artemidorus, a Greek geographer born around 140 BC and author of a book describing a voyage around the inhabited Earth, provided him with a description of the coasts and thus of the shape and size of countries. Book VII was based on the same authorities and described the Danube Basin and the European coasts

His books on Greece

of the Black Sea. Writing about Greece, in books VIII to X, he still relied upon Artemidorus; but the bulk of his information was taken from two commentators of Homer —Apollodorus of Athens (2nd century BC) and Demetrius of Scepsis (born around 205 BC)—for Strabo placed great emphasis on identifying the cities named in the Greek epic the *Iliad*. Books XI to XIV describe the Asian shores of the Black Sea, the Caucasus, North Iran, and Asia Minor. Here Strabo made the greatest use of his own observations, though he often quoted historians who dealt with the wars fought in these regions and cited Demetrius on problems of Homeric topography in the region about ancient Troy. India and Persia (Book XV) were described according to information given by the historians of the campaigns of Alexander the Great (356 to 323 BC), whereas his descriptions of Mesopotamia, Syria, Palestine, and the Red Sea (Book XVI) were based on the accounts of the expeditions sent out by Mark Antony (about 83 to 30 BC) and by the emperor Augustus, as well as on chapters on ethnography in Posidonius and on the book of a Red Sea voyage taken by the Greek historian and geographer Agatharchides (2nd century BC). Strabo's own memories of Egypt, supplemented by the writings of Posidonius and Artemidorus, provided material for the substance of Book XVII, which dealt with the African shores of the Mediterranean Sea and with Mauritania.

Obviously, personal travel notes formed only a small part of the material used in this considerable work, although Strabo prided himself on having travelled westward from Armenia as far as the regions of Tuscany opposite Sardinia, and southward from the Black Sea as far as the frontiers of Ethiopia. Even on the subject of Italy, where he lived for a long time, Strabo did not himself contribute more than a few scattered impressions. His material, accordingly, mostly dates from the time of the sources he used, although the reader is not made aware of this. The value of firsthand observations, chosen from the sources with care, compensates, however, for his lack of originality and contemporaneousness. Strabo showed himself equally competent in selecting useful information—giving distances from city to city and mentioning the frontiers between countries or provinces as well as the main agricultural and industrial activities, political statutes, ethnographic peculiarities, and religious practices. He also took interest in the histories of cities and states, and—when he knew them—mentioned the circumstances under which they were founded, related myths or legends, wars they had instigated or endured, their expansion or recession, and their celebrities. Geological phenomena were reported when they were in some way unusual or when they furnished an explanation for other phenomena—such as the Atlantic tides in Iberia, the volcanic landscapes to be seen in southern Italy and Sicily, the fountains of naphtha occurring near the Euphrates River, and the rise and fall of the Nile waters. Paradoxically, although the description of Greece fills three whole books, such elements are virtually neglected in them. In this part, indeed, Strabo was more attracted by the problem of identifying the localities mentioned in Homer's works than in the geographical realities. These books, however, illustrate another side of his thought, based on the conviction that Homer was perfectly acquainted with the geography of the Mediterranean area and that the correct critical interpretation would reveal his vast learning. This classical thesis is abundantly defended in Strabo's introduction, which attacks the skepticism of Eratosthenes; moreover, it represents, in Strabo's work, the specific contribution to learning of the Greek cultural tradition.

Descriptions of natural phenomena

BIBLIOGRAPHY. The *Geography* was seldom used in ancient times but was often copied in Byzantium; it first appeared in Western Europe as a Latin translation issued in Rome in about 1469, and was then published in the Aldine edition of 1516, based on a faulty manuscript. The first critical edition was that of I. CASAUBON (1587, then 1620), which was followed by that of G. KRAMER—the first to be based upon an analysis of the manuscripts, 3 vol. (1844–52). Various scientific editions by F. SBORDONE, book 1–6 (1963–70); F. LASSERRE, G. AUJAC, and R. BALADIE, book 1–6, 10 (1966–71), with French translation; and W. ALY, book 1–6 (1972), are in progress. These include information from a newly discovered palimpsest from the library of the Calabrian theologian Cassiodorus (born *c.* AD 485). Meanwhile, H.L. JONES' edition, *The Geography of Strabo*, with an English translation, notes, and maps, remains usable ("Loeb Library," 8 vol., 1917–32); it also includes a select bibliography. Fragments of the *Historical Sketches* were last edited and commented upon by F. JACOBY, *Die Fragmente der griechischen Historiker*, vol. 2A (1925). W. ALY, *Strabonis Geographica*, vol. 4 (1957); and G. AUJAC, *Strabon et la science de son temps* (1966), are the best up-to-date standard works. E.H. BUNBURY, *A History of Ancient Geography Among the Greeks and Romans, from Earliest Ages till the Fall of the Roman Empire*, 2 vol. (1879); and J.O. THOMSON, *History of Ancient Geography* (1948), can also be read with profit.

(F.La.)

Strafford, Thomas Wentworth, 1st Earl of

As one of Charles I's chief advisers before the English Civil War (1642–51), Thomas Wentworth, named earl of Strafford in 1640, was foremost in the struggle to consolidate the sovereign power of the king, as against Parliament. He was ultimately sacrificed by Charles, however, to appease Parliament.

By courtesy of Earl Fitzwilliam



Thomas Wentworth, oil painting by Sir Anthony Van Dyck, 1636. In the collection of Earl Fitzwilliam.

Wentworth was born in London on April 13, 1593, the eldest surviving son of Sir William Wentworth, a wealthy Yorkshire landowner. Educated at St. John's College, Cambridge, and at the Inner Temple, he was knighted by James I in 1611. His marriage to Lady Margaret Clifford, daughter of the impoverished Earl of Cumberland, established a link with an ancient and noble family still influential in the north. Following his marriage he completed his education by a leisurely journey through France, returning to England shortly before his father's death in 1614.

Wentworth represented Yorkshire in the Parliaments of 1614 and 1621, and Pontefract in 1624. His wife died childless (1622) and he married Arabella Holles, daughter of John, earl of Clare, a peer out of favour at court who brought Wentworth into touch with the critics of the King's expensive and inefficient policy of war against Spain and, from 1627, against France. Along with other critics of the court he was prevented from sitting in the Parliament of 1626, and later in the year he refused to subscribe to the forced loan, imposed to pay for the war, and was for some time under arrest.

Early opposition to the crown

Despite his record of opposition to the King's policy, Wentworth was approached by the crown—anxious to strengthen its position in the north—with the offer of a barony (1628). He was appointed lord president of the north (virtually governor of England north of the Humber) and in 1629 was given a seat on the Privy Council.

Wentworth's return to the service of the court, coming so soon after his vehement opposition to it in Parliament, startled even some of his closest friends. His conduct was no doubt partly inspired by personal ambition, though he had logical reasons for his change of front since in the summer of 1628 the King gradually abandoned his war policy.

On the Privy Council Wentworth seems to have advocated the paternalist government that distinguished the early years of the King's personal rule: closer supervision of justices of the peace, more effective implementation of the Poor Laws, of laws against enclosure, and of measures for dealing with famine, though he was not above privately making profit out of the corn shortage of 1631. As lord president of the north he quelled all defiance of his authority and made many enemies by his insistence on the honour due to him as the King's representative, but his administration was on the whole just and efficient; he supervised the local justices and curbed the often tyrannous excesses of local magnates. In 1631 he was deeply distressed by the death of his much-loved wife, though he provoked scandalous rumours not long afterward by secretly marrying (October 1632) Elizabeth Rodes, the young daughter of a neighbouring squire.

Lord deputy of Ireland

The King meanwhile had appointed him lord deputy of Ireland. Taking up his office in the summer of 1633, he immediately set himself to consolidate the royal authority, break the power of the dominant clique of "new English" landowners, extend English settlement, improve methods of agriculture, increase the productivity of the land, and stimulate industry and trade. His ultimate goal was to assimilate Irish law and customs to the English system and to make a prosperous Protestant Ireland into a source of revenue to the English crown.

Wentworth continued his effective and firm-handed administration of Ireland until 1639, when he was recalled to England by King Charles. The King needed advice and support in handling a Scottish revolt precipitated by an ill-conceived attempt to enforce episcopacy on the Scots. Wentworth was created earl of Strafford (1640) and was expected to resolve the crisis. But his policy of making war on Scotland proved disastrous for both himself and the King. The English Parliament, called especially to vote money for the war, proved recalcitrant, and Strafford, in command of the English army, failed to prevent the Scots from overrunning the northern counties. The King, unable to pay his own troops or to buy off the Scots, was compelled by joint English and Scottish action to call a new Parliament in November 1640.

Strafford was the chief target of attack from both nations. He was advised to leave the country, but the King relied on his help and assured him that he should not suffer in life or fortune. Detained by illness, he reached Westminster on November 10 with the intention of impeaching the King's opponents in Parliament for treasonable correspondence with the Scots. The leader of the Commons, John Pym, acted first by impeaching Strafford before he could take his seat in the House of Lords. His trial began in March 1641. The basic accusation was that of subverting the laws and was supported by a charge that he had offered to bring over the Irish army to subdue the King's opponents in England. More detailed charges rested on his administration in Ireland and the north. He conducted his defense with great skill, and it looked at one point as though he might be acquitted. Pym therefore introduced a bill of attainder (*i.e.*, a summary condemnation to death by special act of Parliament). The Commons passed it by a large majority; the Lords, intimidated by popular rioting, passed it, too, but by a much smaller majority. While an angry mob surged around Whitehall, Strafford wrote to the King releasing him from his promise of protection, and Charles, afraid for the safety of the Queen, gave his consent to the bill. Strafford went to the scaffold on May 12, 1641, in the presence of an immense and jubilant crowd. In his last speech he once more professed his faith in "the joint and individual prosperity of the king and his people," for which, in his view, he had always worked.

Strafford's execution

He remains an enigmatic figure in English history: ambitious, greedy for power and wealth, ruthless, and sometimes dishonest, but with a vision of benevolent authoritarian government and efficient administration to which he often gave persuasive expression. He made innumerable enemies, but his few close friends were deeply attached to him. In the last weeks of his life his dignity, eloquence, and loyalty to the King made a deep impression even on some of his enemies.

BIBLIOGRAPHY. C.V. WEDGWOOD, *Thomas Wentworth, First Earl of Strafford, 1593–1641: A Revaluation* (1961), is the fullest biography at present available, including an extensive bibliography. The *Earl of Strafford's Letters and Despatches*, ed. by WILLIAM KNOWLER, 2 vol. (1739–40), is still the largest printed collection of Strafford's papers, but it contains only about a twelfth of the manuscript material in the possession of Strafford's descendant, Earl Fitzwilliam. J.P. COOPER, "The Fortune of Thomas Wentworth, 1st Earl of Strafford," *Economic History Review*, Second Series, 11:227–248 (1958), is of great importance for Wentworth's financial interests and character.

(C.V.W.)

Stratigraphic Boundaries

Stratigraphy has been defined broadly as the branch of geology that is concerned with the order and relative position of the stratified rocks of the Earth's crust. Such concerns are dependent upon the correlation of stratigraphic rock units and, hence, upon the stratigraphic boundaries that delineate these units. Stratigraphic correlations (*i.e.*, the tracing and matching of primarily sedimentary rock units of equivalent age) serve as the foundation for interpretations of past geologic history and may involve many facets of the Earth sciences. They are intimately related to such practical considerations as the exploration for ore bodies and fuels, and to such abstract questions as the origin of the Earth, the evolution of life forms, and the history of other planetary bodies on which stratified rocks occur. Most groundwater reservoirs, fuel deposits of coal and petroleum, many fissionable atomic substances, and substantial quantities of other economically important minerals are contained in the stratified rock succession.

Stratigraphic boundaries may be based upon purely physical differences, such as the nature of the rock types comprising the strata, or upon the kinds and distribution of fossils within the rock strata, or upon all physical and organic differences that may be related within a certain time span. The boundaries of some types of units may be delineated much more precisely than others.

Elements in the classification of stratigraphic units

A rock unit possesses many physical attributes, such as thickness, horizontal extent, or colour, that reflect some of the conditions prevailing at the time of its deposition as well as some qualities that may have come into existence after its deposition. The rock unit classification is based on such attributes. The fundamental rock units are formations, which may be subdivided into two lesser divisions—members or beds—or into the larger divisions—subgroups, groups, or supergroups—when combined with other formations.

The deposition or removal of rock units from the geologic record by geologic events required varying amounts of geologic time; the same is true for the evolutionary changes apparent in fossils. Early geologists gradually became aware of these time requirements, and thus developed time-stratigraphic and geologic time classifications. The large rock successions deposited during a specific interval of geologic time were called systems. They were recognized not on the kinds or thickness of rocks comprising them but by the inference of their existence within the specified time interval. Systems were combined to form larger divisions, erathems or terrains, smaller subdivisions called series, or still lesser stages, and many stratigraphers recognize a still lesser subdivision, the

zone. Units that dealt with time rather than the time-rock concept were (in descending order of magnitude) eras, periods, epochs, and ages.

Only rarely would rock units be deposited within a limited time interval over wide areas to permit their use as rock-stratigraphic, or lithochronologic, indicators. Ashfalls that result from very large volcanic eruptions during a brief interval of time might serve as such lithochronologic markers, but the actual time at which a geologic event took place can only be determined by radiometric dating or by relative dating based on the fossil content—biochronology. The most common method of dating geologic events is by biochronology, which involves biostratigraphy. In this scheme, rock or stratal units are defined and identified by their contained fossils and are subdivided into a variety of zones. The fourfold scheme of classification devised for time-stratigraphic, geological time, rock-stratigraphic, and biostratigraphic units is indicated in Table 1.

Table 1: Classification of Different Kinds of Stratigraphic Units

time–stratigraphic units	geological time units	rock–stratigraphic units	biostratigraphic units (no hierarchial order intended)
erathem (terrain)	era	supergroup	zones
system	period	group	assemblage
series	epoch	subgroup	faunizone
stage	age	formation	florizone
zone (Oppelian)		member	range zone
		bed	biozone
			teilzone
			ecologic zone
			community–zone
			zonule

Stratigraphic relationships

Stratigraphic data represent various attributes of layered rocks within a framework consisting of the two geographic dimensions (length and breadth) plus that of geologic time. For many years the notion prevailed that geologic time was conveniently divided into eras, periods, and epochs by some sort of worldwide convulsions of the geologic record that had left universally recognizable indicators behind. It has become increasingly evident, however, that the rock record is far from uniform. While erosion and removal of strata were taking place in one part of the world, an adjacent region might be receiving deposits and accumulating a stratigraphic record, and still a third region might remain in a state of equilibrium, undergoing neither deposition nor removal of strata. Illustrating such three-dimensional relationships is difficult. The great majority of interpretations of time-correlative strata are based on data that are sporadic in their distribution in space and time, such as fossils or radiometrically datable minerals, and many attempts at depicting such relationships have been unsuccessful in varying degrees because potentially barren or negative intervals were neglected in plotting.

This article treats the historical background and current practice of stratigraphic classification, and the stratigraphic boundaries and boundary problems that emanate from several kinds of stratigraphic division of the rock record. For further information on the chronology of the record, see GEOLOGICAL TIME SCALE; DATING, RELATIVE AND ABSOLUTE; PLEISTOCENE EPOCH; and ROCK MAGNETISM. For further detail on some of the principles and data germane to stratigraphic classification, see FOSSIL RECORD; SEDIMENTARY ROCKS; IGNEOUS ROCKS, PYROCLASTIC; SEDIMENTARY FACIES; PALEOGEOGRAPHY; UNIFORMITARIANISM; EVOLUTION; and CLASSIFICATION THEORY.

HISTORICAL REVIEW OF STRATIGRAPHIC CLASSIFICATIONS

The principle of superposition of strata. Such Greek scholars as Xenophanes of Colophon (6th century BC), Xanthus of Sardis (*c.* 500 BC), and Herodotus of Halicarnassus (5th century BC) began to record observational data and to discover the principles that governed natural phenomena. Although these classical scholars did not envision the necessity for stratigraphic classifications, they did provide several concepts basic to stratigraphy. Classic uniformitarianism (the concept that presently occurring Earth processes are similar in kind to those that occurred in the past) suggested that the presence of marine fossils on land was an indication of the greater extent of past oceans. Unfortunately, such sensible notions failed to be augmented or re-evaluated until the 16th century, except for the brilliant observations of a few scholars like Roger Bacon (*c.* 1220–*c.* 1292) or Leonardo da Vinci (1452–1519). It was da Vinci's rediscovery of the classical concept of uniformitarianism—in particular that fossils were remains of once-living animals, and that marine types in sedimentary rocks inland indicated the former presence of ocean waters—that seemed to mark the prologue to an era of geology whose first chapter was to be written by Nicolaus Steno in the 17th century. Steno's classic work in geology, the *Prodromus* (1669), laid down the foundation for the fundamental notion of superposition of rocks, a basic principle of stratigraphy vital to an understanding of the sequential relationships of rocks. It is, simply, that those rock layers superpositionally highest in a sequence of strata were formed at a later time. This made possible determination of the proper superpositional order of rock strata and the assignment of relative ages to each layer.

Contributions of Steno, Lehmann, and Arduino

Making use of Steno's principle, further efforts at recognizing rock successions during the 17th and early 18th centuries became increasingly common. Among the most important were those of the Englishman John Strachey, who made the earliest geological map in the modern sense (1725) and described the succession of strata in the coal district of southwestern England (1719, 1725).

Johann Gottlob Lehmann, a German geologist, was probably the first actually to measure the thickness of rock units and to record their sequence. He formulated a superpositional classificatory scheme that long influenced his contemporaries and successors and that is perhaps ancestral to the erathems and eras of the modern geological time scale. This subdivision encompasses: (1) an older, superpositionally lower category, the Primary (*Urgebirge*), made up of rocks of chemical origin thought to be older than "the origin of life," (2) a Secondary (*Flötzgebirge* or "flat-lying") group comprising fossiliferous strata, some derived from superpositionally earlier rocks, and (3) alluvial deposits (*Angeschwemmtgebirge*) resulting from "local revolutions and in part from the deluge of Noah." Lehmann thought that this threefold classification applied to mountains and their origin. Three contemporaries helped to demonstrate the superpositional succession and the utility of Lehmann's threefold classificatory scheme. Giovanni Arduino published in 1759 a work on northern Italy, George Christian Füchsel published in 1762 and 1773 on east central Germany, and Peter Simon Pallas published in 1777 on the Ural Mountains of central Russia. Arduino recognized a Primary and Secondary division that were essentially synonyms of the *Urgebirge* and *Flötzgebirge*, and a Tertiary division, composed of highly fossiliferous and more poorly consolidated sediments occurring on the flanks of the mountains, that was superpositionally older than the *Angeschwemmtgebirge* of Lehmann. The term Tertiary survives as a period name in modern geological classification. Füchsel worked only with the middle category of Lehmann, the *Flötzgebirge* or stratified fossiliferous rocks. He recognized that they could be divided into natural geologic entities, the formations of modern usage, and that such formations represented episodes in the history of the Earth marked by fossils that were distinct in stratigraphic position and unique in kind. This is one of the earliest statements noting differential time relating rock strata and fossils. Pallas recognized the threefold succession of rocks in the Urals but made no other contribution to rock-sequence classification.

Abraham Gottlob Werner (1750–1817), a German mineralogist, was perhaps one of the most renowned and influential geologists of the last quarter of the 18th century. He was a very popular professor of mining and mineralogy at the Freiburg Mining Academy; many of his students also achieved renown and so served to pro-

Werner, Neptunism, and the transition strata

mulgate many of his theories of the Earth. His theory (Neptunism) that formations were uniformly distributed over the globe and that all rocks had been precipitated from a universal ocean provoked one of the great scientific controversies of the 18th century until it was proven false before the century's close. Werner did emphasize the differences in age of successive rock strata and further advanced Füchsel's ideas of the diagnostic potential of fossils in strata. He recognized an additional subdivision within Lehmann's scheme, namely the Transition or *Übergangsgebirge*. This was said to occupy a position between the Primary and Secondary and to consist, he thought, of the earliest fossil-bearing strata of geologic time, one that marked the transition between the Earth's habitable and chaotic stages. This division would become what is now termed the Paleozoic Era. The essentially threefold classification of Lehmann thus had become a fourfold one with several potential subdivisions. Crudely, these divisions marked the pre-Paleozoic (Primary or *Urgebirge*), the Paleozoic (Transition or *Übergangsgebirge*), Mesozoic (Secondary or *Flötzgebirge*), and Cenozoic (Tertiary plus the *Angeschwemmtgebirge*).

By the first quarter of the 19th century the subdivisions of Primary, Transition, Secondary, Tertiary, and post-Tertiary, or Alluvium, were in general use throughout Europe and in eastern North America. In the second quarter of the 19th century the modern terms for geologic eras were substituted. The English geologist Adam Sedgwick first used the name Paleozoic in a lecture in 1838; another Englishman, John Phillips, proposed in 1840 both Mesozoic and Kainozoic (the latter subsequently changed to Cenozoic). All the eras of the geological time scale presently in use except for the Precambrian were thus formally proposed and soon supplanted usage of Transition, Secondary, Tertiary, and Alluvium in their era sense. Thus, although initial attempts by geologists had been to achieve a rock classification, it had now been transformed into a chronologic one and increasing attention was devoted to paleontologic aspects of the geologic record and methods of classifying them in their stratigraphical contexts.

Catastrophism versus uniformitarianism

A contemporary of Steno, Robert Hooke (1635–1703) actually made the suggestion almost a century before its final acceptance as a principle that fossils might be used "to raise a Chronology" and be of assistance in stating "the intervals of the time wherein such . . . catastrophies and mutations have happened." In the succeeding years of the 17th and particularly in the late-18th centuries, workers like Füchsel and Werner had clearly seen the importance of fossils in their stratigraphic context. Interest in them continued to grow apace in the 19th and 20th centuries. These investigations substantiated their importance in determining relative ages of their entombing strata, as well as in their usefulness for correlation and for other stratigraphical purposes.

The world of the 17th and 18th centuries was largely conditioned to thinking of geologic surface features as being caused by catastrophic processes, and of these the biblical Noachian Deluge was thought to have wrought all changes after the Genesis world creation. A few brilliant and brave souls challenged this view after the time of Leonardo da Vinci. Bernard Palissy, a French citizen, published a book in Paris (1580) asserting that fossils were indeed the remains of once living animals. Palissy did not survive his "heresy" but was burned at the stake, as also was Giordano Bruno in Rome in 1600 for denying that there had been a deluge and brashly stating that the positions of land and sea had often interchanged in the past.

For better than three hundred years the fossil controversy raged, and uniformitarian concepts of a dynamic Earth changed by natural, commonplace causes were suspect. Georges-Louis Leclerc, comte de Buffon (1707–88), strongly championed the cause of uniformitarianism, although his ideas of how changes in the geologic past had taken place were not convincing. He maintained that natural processes of a commonplace rather than catastrophic nature played a major role in producing geologic surface features of the Earth.

It remained for the Scotsman James Hutton finally to synthesize (1788, 1795) the data into a convincing hypothesis and demonstrate conclusively that uniformitarianism was a fundamental principle of the science of geology. This important principle and that of the superposition of strata were necessary before stratigraphy could progress beyond a mere classification of rock successions, or before a geological time scale could be devised as a framework for stratigraphic classifications. It now was inevitable that one of the directions taken by subsequent geological investigation would be that related to devising a time scale and determining the relationships of fossils with strata.

The work of Cuvier and William Smith

The idea of a fossil succession. Baron Georges Cuvier (1769–1832) stated that it would be possible to determine the greater or lesser age of some fossil species as compared to others. Baron Cuvier had seen the succession of fossil species in the stratigraphic record and realized their potential value in correlation, even remarking (1817) ". . . but for them, [the fossils] we could never have suspected that there had existed any successive epochs in the formation of our earth. . . ." Cuvier espoused a catastrophist doctrine of periodic destruction and recreation of organic life to explain the changes he saw in the fossil succession. In 1791–93 the Englishman William Smith was engaged in surveying and canal building in England, where his work led him to observe and be impressed with the constant relationship between certain strata and certain fossils. In 1799 Smith dictated a list of fossils common to certain strata that clearly demonstrated his hypothesis that there is not only a succession of rock strata but also a fossil succession, and that one may recognize the relative age of rocks by their contained fossils. In Smith's *Geologic Map of England and Wales with Part of Scotland* (1815), the hypothesis of fossil succession was demonstrated as a valid scientific principle that has since been recognized as one basic to geology in its historical aspects.

The application of the principle of fossil succession to rock successions previously classified on their lithologic (rock characteristics) superposition alone indicated that errors had been made in initial relative-age assignments. In several instances in which lithological characteristics had suggested the placing of rocks in the Primary, their fossil remains indicated that they belonged with much younger rocks. Other strata consisting of little-deformed rocks that had been assumed to be young were now placed with older successions on the basis of fossil content.

Jean-Baptiste-Pierre-Antoine de Monet, chevalier de Lamarck, departed from his colleagues' beliefs that the fossil succession was caused by periodic catastrophic annihilations of living organisms followed by repopulation by newly created forms. Lamarck believed that organisms had evolved by means of their responses to changes in their environment. After each successive environmental change, some organisms, through inability to adapt, did not survive. Others that proved to be more adaptable were able to change their form slightly to survive. Such long-continued adaptation resulted eventually in simple organisms evolving into more complex ones. On studying fossil shells collected by Cuvier and Alexandre Brongniart, a fellow Frenchman, Lamarck determined that almost all were of species unknown in modern seas, and that their nearest living kinship was to types still living in warmer waters of tropical areas. The publication of pictures and scientific descriptions of these shells by Lamarck did much to suggest the correlation of strata with similar fossils in the London Basin and in the region of the sub-Apennines of Italy.

The term Tertiary proposed by Arduino for those fossiliferous sedimentary rocks lying above the Secondary was applied to other rocks traced over broad regions of northern Italy, partially on the basis of lithologic similarity and fossil content. Because fossils from them compared favourably to those in the rocks studied by Cuvier and Brongniart in the Paris Basin, the name Tertiary was extended there and ultimately to rocks of similar age in the London Basin. The name Tertiary became a formal

Lyell and the subdivision of the Tertiary

period of geologic time through the work of Sir Charles Lyell and Gérard-Paul Deshayes. During analysis of the fossil mollusks collected from Tertiary rocks of France, England and Italy, Lyell noted that the stratigraphically oldest beds of the Tertiary (2,500,000 to 65,000,000 years ago) in the Paris and London basins contained only a few species that Deshayes had identified as still living. Stratigraphically higher, and therefore younger beds, on the other hand, contained increasing numbers of still-living species. These percentage differences suggested to Lyell a method for analyzing relatively large fossil aggregates from rock successions of Tertiary age by comparing the numbers of still-living species in them.

The facies concept. The method propounded by Lyell was that percentages of still-living species in large fossil aggregates collected from successive beds of strata could be used to make a subdivision into epochs. His demonstration of the principle resulted in an initial threefold subdivision (Eocene, Miocene, and Pliocene epochs) in 1833. This was subsequently modified into a fourfold classification (Eocene, Miocene, Older and Newer Pliocene epochs). The fossil aggregates used numbered more than 40,000 specimens, all identified by Deshayes. Although the fossils were collected from rock formations vertically arrayed, no attempt was made to discriminate fossil position within formations by horizon of occurrence. Thus, the fossil collections formed sort of a composite of organisms contained within certain formational units that could be distinguished on lithological grounds. Lyell's methods of collection and analysis unwittingly avoided some of the problems inherent in correlation by fossils. The distribution of organisms is a complex matter, and only in the last part of the 19th and first half of the present century have biologists and paleontologists begun to understand the complex patterns and some of the reasons for the sporadic distribution of fossils in rock strata. Organisms tend to adapt to their environment to the extent that their own genetic composition allows them. Populations of organisms, then, reflect these different environments as they are controlled by such physical factors as light, depth (or height), salinity of water, or temperature. Organic factors also influence the nature of the environment, such as the interaction of individuals within a species population or with other species for space in which to live, food to eat, or other factors. Organisms, then, live within an area called a habitat, which is defined by its limiting factors for their particular habitat adaptation. The net effect of this to the stratigrapher is to find fossils that were once living organisms also adapted to their particular environments, and in strata of similar age, different fossils might be present if environmental conditions had been different in both. If Lyell had collected his fossils by horizons of occurrence through successive formational units, he would have found that many stratal units that would seem to be equivalent in age to strata in some adjacent area on the basis of superpositional relationships, would not have seemed to be correlational by the different fossils because of differing environments. Lyell's collections had been made without discriminating horizons of occurrence, and hence were actually a crude random sampling of most of the species adapted to the varying habitats (or facies) represented by the stratal deposition.

Amanz Gressly (1838) demonstrated that although some lithologic types and some fossil faunas of the same age might be remarkably widespread and constant in character, there were many others in which both the lithology and fossil content varied greatly. The variation in lithology of strata of similar age he called facies. He inferred that rocks of a given stratigraphic unit will show more than a single aspect, and that this aspect reflects a variation in the original environment of deposition. That fossil facies exist also is implied, though not specifically stated, by Gressly. Facies, then, may be considered as the differing paleontologic or lithologic aspects of rocks of the same age but different geographic positions. For the discrimination by fossil content of subdivisions of greater refinement than those made by Lyell for the Tertiary, attention henceforth would have to be directed toward obtaining a better understanding of the evolution and causes of distribution of organisms within their stratigraphic and biogeographic context.

The stage concept. The French paleontologist and stratigrapher Dessalines Alcide d'Orbigny had made detailed studies of the exceedingly fossiliferous strata of Jurassic (136,000,000 to 190,000,000 years ago) and Cretaceous age (65,000,000 to 136,000,000 years ago) throughout France. The Jurassic had been subdivided for some years into three epochs on the basis of a fortuitous threefold lithology, and it was also possible to recognize these subdivisions on the basis of their fossil content. D'Orbigny's attempts at more refined correlations from one place to another often met with frustrations because the lithology and fossil content of stratigraphic units that seemed to be similar in age on superpositional grounds would prove to be dissimilar. He determined to devise a more refined classification of the epochs, and finally (1842) achieved a scheme using the principle of fossil succession that gave each successive layer of strata its characteristic fossils. Extensive travel throughout France and familiarity with fossil content of strata of similar age elsewhere convinced him that, despite lithologic variations, similar assemblages of fossils would always appear in the same superpositional relationships. He called these levels of similar fossils *étages* or stages.

D'Orbigny's subdivision of the Jurassic

D'Orbigny's observation was accurate in its general outlines as far as France was concerned, and some faunal elements could be recognized elsewhere in Europe. Facies still played a role in distribution of the fossils, however, and only coincidentally did it happen that the period of time involved (Jurassic) was one in which unusually uniform conditions of environment were to be found in many of the fossil-bearing strata investigated by him, and also that many of the fossils he used were the commonly occurring, extinct cephalopod mollusk, the ammonite. Ammonites were distantly related to nautiloids and were evolving with great rapidity during the Jurassic. Many kinds had remarkable swimming ability and wide ecological tolerances, and some probably also became widely distributed after death because gases buoyed the shells. The net effect was to allow some types to achieve distribution in many different facies. Although d'Orbigny's mode of implementing his hypothesis had many flaws (not the least of which was the fact that there had been little attempt at determining precise ranges of fossils), his notion of stages found quick acceptance among geologists on the Continent and in England. The concept was without question a brilliant one, and, when coupled with an adequate understanding of the geographic and stratigraphic distribution of fossils, has become the standard method of subdividing epochs. Of the ten Jurassic stages d'Orbigny proposed, most are still in use today, although the stratigraphic boundaries have been redefined.

Quenstedt, Oppel, and the refinement of stages

Recognition of zones. About ten years later (1856–58) a professor at the University of Tübingen, F.A. Quenstedt, criticized several of the more glaring flaws in d'Orbigny's work. One of the more important of these criticisms was that d'Orbigny's fossils had not been collected systematically through strata, and it was thus often impossible to recognize whether strata belonged to one stage or another. Quenstedt felt that the stage boundaries should be based upon the precise stratigraphic range of species. Distribution was not only a matter of vertical or time distribution but of horizontal or geographic distribution as well. He recommended that stratigraphers could better spend their time investigating the bed-by-bed occurrences of fossils than cataloging stages from the most remote corners of the world. The check lists of fossil occurrences would show range of species in each stratigraphic section. An evaluation and synthesis could then be made of the check lists, showing the range of each species in strata from many localities. Only in this way, he argued, could a true understanding be obtained of the ranges of individual species in strata. An understanding of the succession of faunas in strata would result from such studies, and this would provide an idea of the development of evolution within fossil lineages.

Quenstedt's pupils and a following generation of their

pupils continued to make detailed investigations of Jurassic strata throughout the world. The most important of these pupils, Carl Albert Oppel, not only refined the stage but also subdivided it into Zones. (The use of an initial capital letter in "Zone" is a convention indicating that the term is being used in the sense defined by Oppel.) His achievement was more than just the subdivision of stages into Zones, however, for he established a principle based upon the ranges of lineages of fossils that could be applied to units of any magnitude within the geologic time scale. He was able to define distinctive faunal aggregates within stratal successions by the overlap of ranges of several species. By carefully delineating both the upper and lower boundaries of these subdivisions on the basis of the appearance of new species or the disappearance of old species, his Zones could be made both contiguous with each other and mutually exclusive. The boundaries could be more precisely located than any previously devised. He redefined the stage as being made up of groups of Zones, and he therefore defined on the same basis—that is, by overlap of species ranges. In his method of delineating Zones, he determined the range of species throughout the region of outcrop of Jurassic strata in Europe by employing Quenstedt's method of making a synthesis of the range of a species in many local stratigraphic sections. He then selected the Zones on the basis of the overlapping ranges of species that were also of widespread occurrence. Contiguous zones were selected by overlaps in ranges of other widespread species for the subjacent and superjacent Zones. The Zone, in Oppel's words (1856–58), is ". . . marked in any place by a number of species that are constant for it . . .," or, in other words, those with widespread distribution and overlapping ranges. If quantitative data for relative abundances are presented, they allow paleoecologic analysis as well, and thus facies occurrences may be distinguished from those significant of chronologic events.

Zones of the Jurassic and Miocene

Figure 1 illustrates six of the Jurassic Zones and some of the fossil species on which they are based. The Zone of *Arietites bucklandi* is based on the joint occurrences of two species with differing ranges in time, here called "species J" and "G," of the *Arnioceras semicostatum* Zone by overlap in the range of "species G" and "Q," and so on. The zonal fossils are characteristic but are not necessarily confined in their occurrence to the Zone.

Based on Arkell in James R. Beerbower, *Search for the Past: An Introduction to Paleontology*, 2nd ed., © 1968; by permission of Prentice-Hall, Inc.

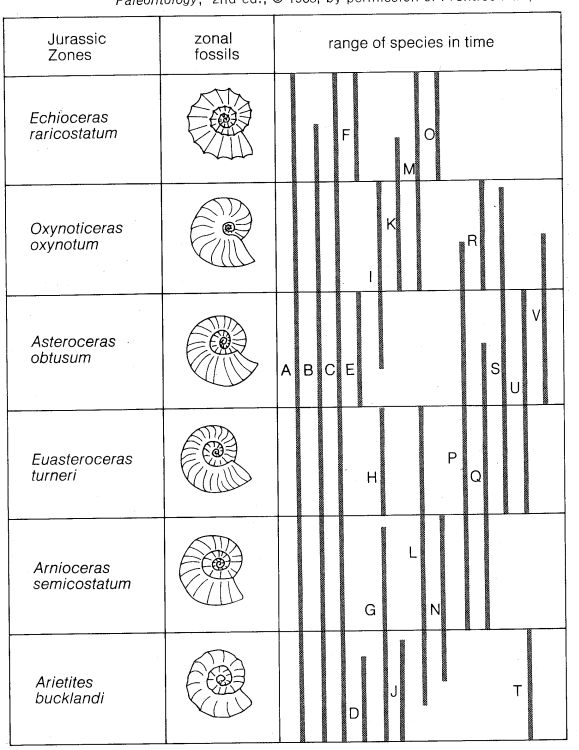


Figure 1: *Jurassic fossil zones and zonal fossils.* These six ammonite species never overlap in stratigraphic distribution and therefore provide a guide for correlation wherever they occur. The ranges of hypothetical species (indicated by letters) within the Jurassic are also shown.

Many successful applications of zonal subdivisions have been made since Oppel's time. Perhaps the most rigorous and innovative, as well as successful, has been the work of the American paleontologist and stratigrapher Robert M. Kleinpell. Kleinpell's six stages and their zonal subdivisions for the California Miocene series (7,000,000 to 26,000,000 years in age) were based upon collections of fossils deliberately made to represent the majority of faunal facies within the strata. His check lists served not only to present distributional data for chronological evaluation but also relative abundances of species. These latter were used in paleoecological evaluation. The final synthesis of data presents a comprehensve picture of the paleogeography of the Miocene and indicates major transgressive cycles (advance and retreat of seas) and orogenic events (mountain building and volcanism), suggesting their possible relation to cycles and events of similar ages in areas remote from California.

It must be noted that a vast gap exists between theory and practice with regard to stages and Zones. D'Orbigny's concept of stages delineated over the whole Earth would hold if well-documented species ranges existed for each succession of epochal strata over the world. Oppel's work extended Zones enormously and transferred them from mere local records of succession to correlational planes of much wider (theoretically universal) application. One of the more significant subsequent contributions to Oppel's method of zonal determination has been the realization that geographic limits do exist within which a Zone is valid. Outside these limits, it may not be recognized. The limitation results from the fact that most species are confined in their distribution to a particular region, termed a biogeographic province; or if they have migrated beyond this province, then their range in time is frequently so great as to not be practical for the discrimination of Zones. The Zone in the sense intended by Oppel is considered to be biostratigraphic rather than time-stratigraphic by some authorities and is called concurrent range zone.

Introduction of radiometric dating. The development of radiometric dating of rocks has been accomplished only in recent years, and a radiometric time scale is still in its infancy. Because most radiometric dates are obtained from intrusive igneous rocks (those crystallizing from silicate melts within the Earth), which often cannot be related with stratigraphic precision, large gaps occur in the dates, and the accuracy of others is open to question. The eras, periods, epochs, and some of the stages and Zones of geological time have been given absolute values in years by radiometric dating. Within the most recent geological period, the Quaternary (the last 2,500,000 years), it has been possible to date many carbon-bearing organic substances that are not older than approximately 40,000 years by the carbon-14 method. Within this time interval many major advances and retreats of glaciers, climatic changes, important dates in human history, the times of migration of animal and plant species, and past volcanic eruptions have been dated by the carbon-14 method.

STRATIGRAPHIC NOMENCLATURE IN THEORY AND PRACTICE

Terminology and standards. Since the first half of the 19th century, the work of constructing a standard time-stratigraphic and geologic time chronology has been vigorously pursued by geologists in many parts of the world. Without an organized body to suggest uniform standards, there was great divergence in terminology used by different stratigraphers. By the last quarter of the 19th century it was evident that greater uniformity would be desirable, and the First International Geological Congress was held in 1878 in Paris. The primary purpose was to study "uniformity in geological reports with respect to nomenclature and map symbols." The International Commission appointed during the First Congress made its report at the Second Congress, held at Bologna in 1881. The classification adopted at the congress was a dual hierarchy of chronologic and stratigraphic terms.

Stratigraphic commissions and codes

Table 2: Classification of Stratigraphic Units

time–stratigraphic units	geological time units	rock–stratigraphic units	biostratigraphic units (no hierarchial order intended)	geological climate units (for use in the Quaternary)
—	eon	supergroup	assemblage zone	glaciation
—	era	group	subzone	stade
system	period	subgroup	zonule	interstade
series	epoch	formation	range zone	interglaciation
stage	age	member, tongue, lentil bed	local range zone	
			concurrent range zone	

Source: American Commission on Stratigraphic Nomenclature (1961).

The Eighth Congress (1900) in Paris adopted a somewhat revised code in which the subdivision of age is called phase, and the term Zone is substituted for a subdivision of stage. No subdivision of Zone is indicated.

Usage in North America has tended to diverge somewhat from that of Europe. A general committee was appointed in 1930 by the Association of American State Geologists to study the question of a more general uniformity in stratigraphic and geologic nomenclature. The stratigraphic code that resulted from the committee's deliberations was presented in 1933. It differs from those of the International Commissions in replacing the stage and Zone by local rock unit names.

An American Commission on Stratigraphic Nomenclature was established in 1946. It adopted a code involving the idea of multiple categories of stratigraphic units, a device that is of great utility. Table 2 indicates this fourfold classification into rock-stratigraphic, soil-stratigraphic (no formal nomenclature), biostratigraphic, and time-stratigraphic units. Two additional categories that are considered to be closely related to stratigraphic units are included: namely, geological time units and geological climate units that have a unique usage in Quaternary stratigraphy (the last 2,500,000 years).

Formations, groups, members, and beds

Rock-stratigraphic units. The rock-stratigraphic classification is based upon lithologic changes in the stratigraphic record. Boundaries between units are placed either at sharply bounded differences in lithology or are fixed arbitrarily in gradational intervals. The fundamental unit is the formation, which is characterized as being lithologically homogenous, "prevailingly but not necessarily tabular," and permitting mapping at the surface or tracing in the subsurface. Its homogeneity might consist of a uniform lithology (sandstone for instance), or it might be composed of members or beds of varying lithology (*e.g.*, alternating sandstone, shales, limestones, and conglomerates). Its deposition may have occupied a large or small amount of geologic time, and the layers need not have been continuously deposited but may include breaks in the time-stratigraphic succession. Both sedimentary and extrusive igneous-layered rocks may properly constitute formations—for instance, a limestone that is interbedded (alternates vertically) with a volcanic ash or tuff layer. Only those metamorphic rock bodies in which original textures or structures allow the geologist to recognize mappable units may be considered stratigraphic. Intrusive igneous bodies that are discriminated by mineralogic or textural characteristics or chemical composition may be classed as formations or groups but not treated in the time-stratigraphic sense.

Formal assemblages of several associated formations are called groups; several formations within groups that are designatable may be defined as subgroups. Supergroups are formal assemblages of two or more related groups or groups and formations.

Rock-stratigraphic categories of next inferior magnitude to formations are the members, and units inferior to them in magnitude are beds. Such inferior categories are used to designate distinctive rock-stratigraphic subsidiary units within larger bodies. They may or may not be restricted geographically, but are frequently of local significance only as cartographic units.

Names for rock-stratigraphic units are compounded from a binomial consisting of a geographic name derived from characteristic exposures in combination with either the lithologic term, if appropriate, or the word formation, if made up of more than a single rock type. Examples are Cowlitz Formation, Stillwater Member of Cowlitz Formation, Roop Sandstone, or Pereira Shale. Names applied to beds are rarely given formal recognition in the body of stratigraphic nomenclature, and they may apply to coal beds, oil sands, beds of significance in rock quarries, aquifers, or conspicuous fossil horizons.

Soil-stratigraphic units

Also recognized are soil-stratigraphic units, which are rock-stratigraphic in nature, though no formal nomenclature has been proposed for them. Such units are composed of residual soils (soils formed mostly in place from underlying rock-stratigraphic units) exhibiting sufficiently distinctive features and stratigraphic relations to permit them to be constantly recognized and mapped in the manner of a stratigraphic unit. Although a soil-stratigraphic unit may coincidentally constitute a pedologic (soil) unit also, it might be distinct from others by comprising one or more pedologic units or parts. Boundaries of soil-stratigraphic units are placed at sharp differences in lithology or within gradational intervals. Formal names are given by rules governing rock-stratigraphic nomenclature but should not conflict with either the latter or with pedologic names.

Time-stratigraphic units. The time-stratigraphic unit differs from a rock-stratigraphic unit because time plays only a coincidental role in the deposition of the latter and none in its definition. Lithology, on the other hand, although playing only a coincidental role in time-stratigraphic boundaries, is all important for rock-stratigraphic designations.

Systems, series, and stages

The Commission recognizes the system as the fundamental unit for the worldwide time-stratigraphic classification of rocks of the Paleozoic, Mesozoic, and Cenozoic eras. The systems of Precambrian time (*q.v.*) are not yet sufficiently defined or widely accepted. The series is next in rank below the system, and is defined in the same way as system but of lesser magnitude. A stage, according to the Commission, is the time-stratigraphic unit of lower rank than series, and is based on a succession of biostratigraphic zones.

Historically many of the early geologists selected boundaries that fell at hiatuses (breaks or gaps) in the record because these frequently marked drastic changes in lithology or faunal content or some pronounced physical break. The percentage of living species used by Lyell for time-stratigraphic subdivision of the Cenozoic was previously noted. The system was the first time-stratigraphic unit to be named. This naming was the historical accident wrought by Arduino's introduction (1759) of the term "Tertiary," which is the only term of those from the earliest classifications of rock successions to survive into modern times.

Faunal evolution and biostratigraphic units. For much of the 19th and early 20th centuries, many geologists believed that diastrophism (the deforming and building processes of geology that result in continent or mountain making) was the ultimate basis for dividing geologic history. The theory holds that diastrophic movements are cyclic and synchronous on a worldwide scale. Biologic events such as fossil evolution were thought to occur in conformity with stratigraphic systems and series because like them they were closely influenced by diastrophic events. The concept seemed to offer a simple explanation for the occurrence of those times of apparent rapid evolution in the geologic past when accelerated rates of morphologic change resulted in increased complexity or anatomical specialization. Many more new species and genera were developed in a shorter span of time during inter-

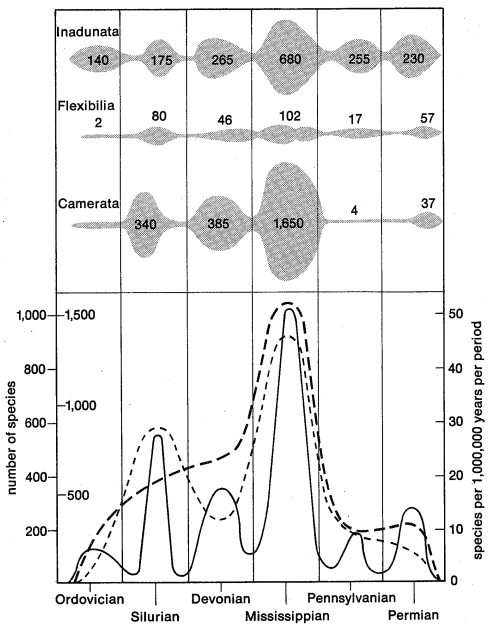


Figure 2: *Paleozoic crinoids of various ages.*
(Top) Geologic distribution of crinoid subclasses; figures indicate numbers of described species. (Bottom) Curves showing distribution of Paleozoic crinoids: solid curve (calibrated outside to left) is based on early, middle, and late parts of periods; long-dashed curve (calibrated inside to left) is based on whole periods; short-dashed curve (calibrated to right) is based on numbers of species per million years.

From (top) R.C. Moore, C. Lalicker and A.G. Fischer, *Invertebrate Fossils* (1952); (bottom) J.M. Weller, *The Course of Evolution* (1969); both by permission of McGraw-Hill Book Co.

vals termed explosive or eruptive. Whether such evolution was really rapid or only seems so because of (1) missing parts of paleontologic history, (2) unusually detailed study of some groups that resulted in recognition of many more species and genera in certain groups than in others, or (3) had actually been more rapid than in more conservative lineages is a subject of some dispute. Diastrophism has often been cited as the major causative agent for evolutionary changes in groups of organisms. If it did indeed play such a role, it affected different groups in different ways at different times. The evolution of crinoids (sea lilies) and trilobites, for instance, does not follow the same pattern. The crinoids reached a peak in described numbers of genera (one measure of rapidity of evolution) in the Middle Mississippian Period (325,000,000 to 345,000,000 years ago; see Figure 2). Diversity is also greatest at this time in the three subclasses of the group. Trilobites, on the other hand, are most diverse in numbers of species in the Cambrian (500,000,000 to 570,000,000 years ago) and Ordovician (430,000,000 to 500,000,000 years ago). Figure 3 illustrates the varying importance of different fossil groups throughout geologic time. The most useful groups stratigraphically are those with a complex morphology, widespread geographic distribution, fairly common occurrence, and rapid evolution. It is apparent that the crinoid maximum does not correspond with their time of greatest stratigraphic usefulness, whereas the trilobite maximum does. Some of the difference in stratigraphic usefulness of the two groups is related to a possibly more restricted environment of development for crinoids than for trilobites, resulting in greater representation in the stratigraphic record for the latter. Many so-called intervals of explosive evolution reflect preserved lithologic facies, accidents of discovery, concentration of research, or other factors; some others may reflect diastrophism. Several important stocks of invertebrate animals (trilobites, brachiopods, corals) show extinctions or marked evolutionary change toward the close of the Paleozoic Era (Pennsylvanian [280,000,000 to 325,000,000 years ago] and Permian [225,000,000 to 280,000,000 years ago] periods) when one of the large widespread diastrophic events of the geologic record took place. On the other hand, the nautiloid evolutionary peak seems to have occurred in Ordovician times, the crinoids in Mississippian time, and the ammonites in Jurassic time; yet none of these periods is regarded as one of outstanding diastrophic activity. Figure 1 illustrates the use of a rapidly evolving invertebrate group (ammonites) during its peak period (Jurassic) for classification. Although many species of ammonites were present in the strata, each one's time of existence as a species was variable: some were long ranging in time, and others were of relatively short duration. The most refined correlations of strata are made by selecting short ranging species or combinations of first and last occurrences of mutually occurring species. It is evident that evolution has not been uniform in all fossil groups, and that it has not coincided with stratigraphic events reflecting diastrophism. Each geologic period has some groups whose evolutionary and distributional attributes make them better suited than others for classification of geologic time. It is probable that diastrophism has played an important though not dominant role in evolution by affecting lineages of species where tectonic activity has been prevalent. Any change in the physical and chemical environment will introduce conditions favourable to some organisms but not to others. Because some species cannot adapt to such changes or be stimulated to evolve eventually into types that can, they are forced to either migrate or perish. It is evident that diastrophism cannot be invariably or regularly related to fossil evolution, and from this standpoint it also cannot be used as the "ultimate basis" for correlation or subdivision of the geologic record.

Evolution of crinoids and trilobites

Modified after Teichert in James R. Beerbower, *Search for the Past: An Introduction to Paleontology*, 2nd ed., © 1968; by permission of Prentice-Hall, Inc.

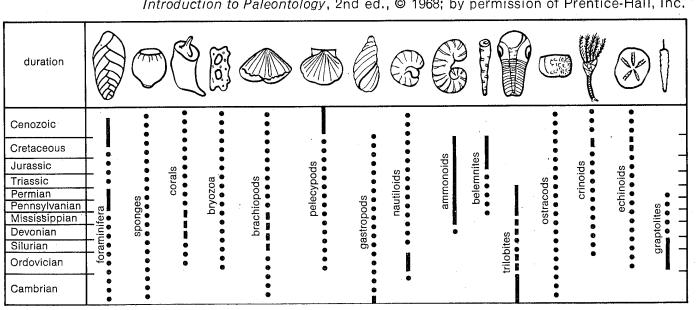


Figure 3: *Stratigraphic significance of fossil groups.*
The solid line indicates the group is important for worldwide zoning and correlation; the dashed line means that it was important in regional correlation; the dotted line signifies only occasional or rare use as zonal fossils.

Assemblage zone and biozone

A biostratigraphic unit is one defined as a body of rock strata characterized by its content of fossils contemporaneous with the deposition of the strata. The assemblage zone is a descriptive unit defined as a bed or group of beds characterized by an assemblage of organisms, one of which is chosen to give its name to the unit, although it need not be confined to this unit or found in every part of it. It is evident that the unit is defined by its directly observable fossil content and requires no interpretation (as does a Zone in the sense of Oppel) for its recognition. Fossils may occur jointly because they were part of an ecologically adapted community (*i.e.*, faunule) or as a result of transport that has brought some or all organisms from geographic sites other than that of their final deposition. Not all fossils in such an assemblage, or possibly none, are of value in interpretation of time, paleoecology, or biogeography. Three successive assemblage zones at three geographically separate sections are depicted in Figure 4. The top part of the figure illustrates the predominant fossils in each and their distribution in the strata; the middle portion depicts the lithologies in each section and the time lines between the sections, based upon an interpretation of the range of brachiopod and graptolite species in each. The lower diagram illustrates the distribution of the five range zones. None of the zones, based on either assemblage or range, need be mutually exclusive, or contiguous, nor need they maintain the constant lateral relationships that an Oppelian Zone or other time-stratigraphic unit would display. Detailed collecting of fossils from the graptolite-bearing shales of

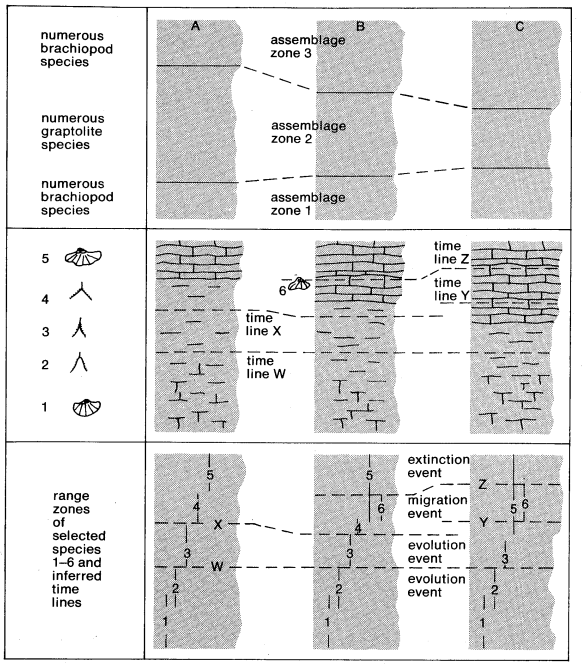


Figure 4: Assemblage zones, selected range zones, and time lines inferred with varying degrees of confidence, in three stratigraphic sections (see text).
From Don L. Eicher, *Geologic Time* (1968); Prentice-Hall, Inc., Englewood Cliffs, New Jersey

assemblage zone 2 shows a vertical gradation of species that appears to represent an evolutionary line or phyletic lineage. The evolutionary event of appearance of a new species or its extinction is a unique event in time, and therefore for each new species and its extinction within these strata, one can infer a time-stratigraphic unit for which many stratigraphers utilize the term biozone. This refers to the deposits formed during the life span of a taxon, which can be any rank, such as species, genus, family, order, or higher rank, in a hierarchial system. Some geologists have rejected the term biozone, declaring that it has been used in the time-stratigraphic sense for the strata deposited during the life span of a taxon and for the total time span of a taxon. It has been recommended that the term be retained in its time-stratigraphic sense and redefined to include all the rocks deposited anywhere during the entire life span of the taxon.

Teilzone, peak zone, and concurrent range zone

The German term teilzone was introduced for the local time of existence of a taxon. Arkell proposed the substitution of teilchron for this term and redefined teilzone in conformity with his definition of biozone as the rocks deposited during the teilchron of any taxon. A geographically widespread taxon might, with such a definition, have an infinite number of teilchrons and teilzones, and unless agreement were reached as to geographic limitations, the terms would be of limited utility. For this reason, and because natural aggregates of organisms tend to occupy areas of land or water termed provinces as floras or faunas, the American stratigrapher V.S. Mallory proposed that the local extent of the teilchron and teilzone be constituted by a biogeographic province. The Commission does not separately classify the time value of a range zone from its biostratigraphic value. Many stratigraphers, however, would consider a range zone as an observational unit in strata, the biozone or teilzone as interpretive and time-stratigraphic, and the biochron or teilchron as interpretive and geological time terms. Figure 5 illustrates the various factors that influence biostratigraphic occurrences of fossils. These are: (1) evolution (first appearance of a taxon), (2) extinction (last appearance of a taxon), (3) nondeposition, (4) erosional removal, (5) environmental exclusion, and (6) migration.

The term subzone is recognized by stratigraphers as a subdivision of zone, and the term zonule as the smallest recognized subdivision of a zone. Zonules need not be vertically contiguous, but can be recognized only when a zone has been subdivided into subzones. Unfortunately, the term duplicates a biostratigraphic term for the preserved remains of a fossil community in strata, zonule.

Peak zones are zones characterized by an exceptional abundance of a taxon, such as an abundance of species reflecting a near optimum or optimum environmental condition. The terms epibole, acme zone, and flood zone have been commonly used for somewhat similar phenomena that express environmental optima. Characteristically, because such environmental conditions may change in time, migrate laterally, and frequently be of sporadic occurrence, the peak zone has environmental rather than temporal significance; locally, if its distribution is well understood, it may be useful in the latter sense.

A concurrent range zone is a biostratigraphic unit defined by the overlapping ranges of specified taxa from one or more of which it takes its name. This zone has a type section that serves as a standard of reference. The specified taxa are only those that form a unique association, however, because of an overlap in their ranges. Normally, only a few of the species that occur jointly in a sample will be useful for establishing such a zone. Some will be too long ranging in time, others may represent the peak zone of time-recurrent epiboles, others may represent some other paleoecologic control on distribution, and still others may be of too short range to define zonal boundaries. Many stratigraphers consider concurrent range zones to be time-stratigraphic, because the boundaries are based on such time-related events as evolution or provincial migrations in first and last appearances of taxa in strata. These zones are based on designated rock successions through which the vertical ranges of the contained fossils have been determined.

Modified after Teichert in James R. Beerbower, *Search for the Past: An Introduction to Paleontology*, 2nd ed., © 1968; by permission of Prentice-Hall, Inc.

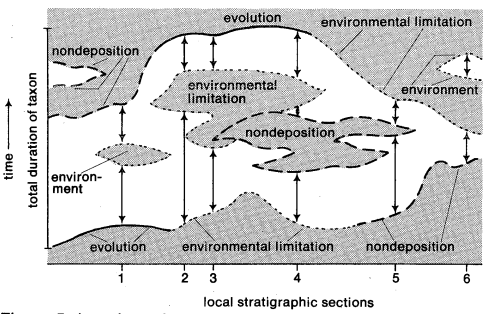


Figure 5: Local stratigraphic sections showing the factors limiting the stratigraphic range of taxon. The total potential stratigraphic range is set by evolution (lower limit) and by evolution or extinction (upper limit). In any particular local section, nondeposition (or erosion) and environmental exclusion further limit the range. The ranges in any two local sections, therefore, may or may not indicate an overlap in time.

Biostratigraphic boundaries differ from rock-stratigraphic in fundamental concept because fossils define the one, rock type the other. It may happen that a lithologic change also is reflected in responses of the fossil community to a change in the depositional environment, but the boundaries may also cross or lie at quite different horizons. Biostratigraphic units are bounded physically by the strata characterized by a certain fossil or fossils and extend no further than the limits of these fossils in the strata. Commonly, biostratigraphic data provide the best means for defining time-stratigraphic boundaries.

Geological time units. Geological time units are nonmaterial units that represent specific segments of time distinguished on the basis of the rock succession. Time terms are useful in geology, particularly for correlation, and in historical geology, despite the fact that a fossil or a rock succession can be related to a certain time interval only by inferring its position within a rock succession. Boundaries of such units depend on specified sections of the type area of the systems, a time-stratigraphic unit. A period was the time interval spanned by the deposition of the system. The boundaries of time units are probably imprecise because paleontologic and radiometric correla-

Period, epoch, and age

tions lack precision. The epoch is the time of deposition of its time-stratigraphic equivalent, the series, and the age is the time equivalent of the stage. The post-Precambrian systems are grouped together into three subdivisions that lack a time-stratigraphic name in the Commission's recommendations, but that are called erathems by many stratigraphers or terrains, a name widely employed in Europe. The time equivalents of these latter are termed eras. George Halcott Chadwick in 1930 proposed the use of the term Cryptozoic Eon for Precambrian time, and Phanerozoic Eon for the post-Precambrian eras. The term eon is recognized by the Commission as being a combination of several eras.

Stratigraphic boundary problems. *Lower Paleozoic boundaries.* The great significance of the Cryptozoic-Phanerozoic boundary (essentially, the base of the Cambrian, which marked the beginning of the principal fossil record) was crudely recognized by Werner's recognition in the 18th century of the "Transition" as being distinct from the *Urgebirge* and *Flötzgebirge*. Of the many people who subsequently studied Werner's Transition rocks, the British geologists Adam Sedgwick and Sir Roderick Impey Murchison were paramount. In 1831 they jointly began field work on thick successions of Transition rocks in Wales. Sedgwick recognized a stratal sequence of rocks in North Wales that he named the Cambrian System (500,000,000 to 570,000,000 years old), while Murchison defined the superpositionally higher Silurian System (395,000,000 to 430,000,000 years old). The formal names were proposed in 1835, as was a recommendation that "Transition" be abolished in favour of Cambrian Silurian. Sedgwick's studies emphasized structure and lithology, and although he described the rocks in detail, he included none of the fossils. Murchison, on the other hand, clearly used Smith's principle of faunal succession and implied that the Silurian System could be recognized elsewhere by its characteristic fossils. In 1842 Murchison's further studies indicated to him that the oldest fossiliferous rocks were Silurian and that fossils from Sedgwick's Cambrian were the same as those in his Lower Silurian. Since Sedgwick had failed to define the fossils as well as the rocks in the system, and Murchison by now had claimed all fossiliferous rocks in both his Silurian and the pre-Silurian of Wales, he was hard put to defend the existence of a Cambrian. Not until the middle of the 19th century were fossils from Cambrian rocks finally described, and the Murchison-Sedgwick controversy, on which the fate of the Cambrian system hung, was not resolved until 1879. In that year another British scientist, Charles Lapworth, published the result of his studies of Cambrian and Silurian rocks. He recognized the existence of three distinctive faunas. One was characteristic of the Lower Silurian, one of the Upper Silurian described by Murchison, and a third consisted of the lower part of Sedgwick's Cambrian rocks. Lapworth proposed that the name Silurian be restricted to the upper strata of Murchison's Silurian with its distinctive fauna. Cambrian, he suggested, should be redefined to include the lowest fauna and strata found in Sedgwick's original type. He further proposed that the distinctive middle fauna and the strata included both in the type Lower Silurian and type Upper Cambrian be defined as a new system, the Ordovician (430,000,000 to 500,000,000 years old). And so, about a half dozen years after the deaths of the two disputants the Gordian knot of one of the early boundary disputes was finally untied.

Upper Paleozoic boundaries. In the years 1836 through 1839, prior to their controversy, Sedgwick and Murchison spent some time investigating the stratal succession next above the Silurian. Lying above Murchison's Silurian System was a distinctive red-coloured sandstone unit that was called the Old Red Sandstone. These sandstones were recognized as a widespread lithologic unit in many places in South Wales and on its border. The age of these beds was thought to be Carboniferous (280,000,000 to 345,000,000 years old). Fossil plants collected from strata in Devonshire and thought to be possibly Cambrian or Silurian in age attracted their attention, and the two scientists traveled to Devonshire. The plant fossils were finally determined to belong to Carboniferous rocks lying above some deformed graywacke (a sandstone with a muddy matrix and rock fragments) and interbedded limestones that lithologically were similar to some Cambrian strata in Wales. The limestones at Devon differed in bearing fossils, including corals, that were lacking in the Cambrian and differed from known Silurian corals. William Lonsdale, an amateur paleontologist, became interested in the problem. He had become a specialist on Carboniferous corals and was familiar with those from Silurian strata. He concluded that the corals collected by Murchison and Sedgwick were intermediate in age between those of the Silurian and Carboniferous systems. In 1839 Sedgwick and Murchison published a paper in which the name Devonian System was formally proposed for the strata lying beneath the Carboniferous and bearing fossils intermediate in characteristics between those of the Carboniferous and the Silurian. The Devonian was considered to be equivalent to the Old Red Sandstone, which bore only plant and fish remains.

Determination of the Devonian, Carboniferous, and Permian systems

Murchison learned of fossiliferous strata exposed in European Russia and accepted an invitation in 1840 from the Tsar to study them. Together with the French paleontologist M. de Verneuil and the Russian Count Keyserling, an excursion to the region was made. Murchison was able to recognize strata of a succession representing the Silurian and red sandstones containing fossil fishes similar to those contained in the Old Red Sandstone of Great Britain. Interbedded marine fossils of Devonian age were interbedded with the red beds and thus conclusively demonstrated that the Old Red Sandstone was indeed of Devonian age. On a return to Russia the following summer, Murchison and de Verneuil studied a thick deposit of fossiliferous limestone containing fossils younger than the Carboniferous but superpositionally below and grading upward into deposits of red sandstone and shale. Although these rocks contained fossils somewhat similar to those contained in the Carboniferous, they contained fossil fishes and amphibians similar to known post-Carboniferous fossils. The Magnesian Limestone of England contained such fossils. It was known to pass from limestones into a red bed facies as it was traced westward. The red bed facies was called the Lower New Red Sandstone and assumed to belong to the Triassic System. A similar relationship between the Zechstein, a fossiliferous unit in Germany (with similar vertebrate fossils), and the Triassic suggested that it was correlative with the Russian unit also. Murchison determined that these strata lying between the Carboniferous and Triassic (190,000,000 to 225,000,000 years old) belonged to a new system, the Permian (225,000,000 to 280,000,000 years old). The previous inclusion of Zechstein and Magnesian Limestone in the Triassic was shown to be in error.

The type Permian has yet another boundary problem, one of several that plague geologists today. Murchison's type Permian had a lower boundary in sandy strata that overlay a limestone. Subsequently, it was found to contain a fossil fusulinid fauna younger than any known in the type Carboniferous. The question of the appropriate inclusion of these beds within either the Permian or Carboniferous systems is not yet settled.

Boundaries of all systems have been the subject of different interpretations by geologists from their first formal naming. Physical boundaries were found to be lacking or in a different position outside the type areas. Fossil facies were discovered to differ from one geographic place to another. With more detailed knowledge of the stratigraphic succession, unconformities were found to mark many of the boundaries in some places, whereas more complete successions were present elsewhere. A proposal was made at the International Geological Conference in 1964, in New Delhi, that a reinvestigation be made of systems in general, and that subsidiary types be redesignated in sections where more continuously deposited strata could be found. Although the suggestion contains much merit, many geologists believe that boundary disputes will always exist.

Cenozoic boundaries. Cenozoic boundaries (the Cenozoic Era began 65,000,000 years ago) present some

unique problems. Lyell was careful to point out (1833) that the epochs and series subdivisions he proposed for the Tertiary (2,500,000 to 65,000,000 years ago) were not necessarily the only ones ever to be delimited within the Tertiary. He was quite aware that further analysis of its faunas might lead to more refined subdivisions, which later took place. In 1839 Lyell modified his original terminology and renamed his Newer Pliocene the Pleistocene (10,000 to 2,500,000 years ago), and restricted the Pliocene (2,500,000 to 7,000,000 years ago) to his former Older Pliocene. Shortly thereafter, investigations of glaciation in Europe, particularly in the Alps, led Edward Forbes, an English naturalist, to declare in 1846 that the glacial epoch was synonymous with the Pleistocene. In 1873 Lyell abandoned his original basis of defining Pleistocene on the fossil content and accepted the idea of the glacial epoch (*i.e.*, time of glaciation) as the criterion for delineating the Pleistocene. The climatic basis for definition has brought about its own peculiar set of problems. It is difficult to relate alpine glaciation in regions with continental glaciation of the lowland areas. It is also difficult to relate the events of glaciation with those of nonglaciated areas, and frequently difficult to relate the continental succession of rocks to the more completely preserved succession of marine fossiliferous sediments. Moreover, the advent of glaciation occurred much earlier in some places (as in more northern areas) than in others, so as a time-defining characteristic, glaciation is not appropriate on a worldwide basis.

The Pliocene–Pleistocene and Pleistocene–Recent boundaries

Somewhat similar problems beset the recognition of the Recent or Holocene Epoch (the last 10,000 years, approximately). Lyell originally defined it rather loosely as the strata deposited subsequent to the Pleistocene. If the Pleistocene ended by the retreat of the last glaciation, then the meaning of the term is evident. If, however, as many geologists believe, the present is merely one of several interglacial intervals that have occurred, then the designation Pleistocene would seem more appropriate than Recent for the strata and life of today (see below *Climatic units*).

Other terminologies and divisions. *Sequences of strata.* Cyclic or supposedly cyclic phenomena have received much attention as possible methods of classifying geological events. The strongly catastrophist viewpoint about geologic events of many 17th and 18th century geologists undoubtedly coloured 19th and early 20th century views of diastrophism on a worldwide and intermittent basis. During uplifts of land areas, there was universal retreat of the seas with consequent erosion and formation of unconformities, whereas during submergence of land areas, universal deposition of sediment occurred. One of the strongest advocates of correlating geologic events by means of diastrophism, either by reference to unconformities or related effects, was the American geologist T.C. Chamberlin, who in 1909 referred to it as the ultimate basis of correlation. Although some widespread diastrophism has undoubtedly taken place in the geologic past, there is little evidence or belief today that tectonic activity involving uplifts or subsidence of land areas with consequent withdrawal or advance of seas has been synchronous everywhere, but rather at times affected some areas a great deal, others not at all. Among modern usage of diastrophic effects in classifying stratigraphic, the term sequence should be mentioned. The sequence concept has done much to revolutionize historical geologic interpretations. The name was proposed by the American geologists L.L. Sloss, W.C. Krumbein, and E.C. Dapples (1949) and later modified by H.E. Wheeler (1960 and 1964) and L.L. Sloss (1963). It is defined by the rock-stratigraphic units that are bounded by regional unconformities. Such major unconformities as those subdividing the stratigraphic record into sequences may be recognized over far greater geographic areas than any of the component individual rock-stratigraphic units. Because the sequence concept emphasizes not only the preserved record between the unconformity but also the physically nonrecorded parts that constitute the lacuna, it makes it possible to relate all the events in the geologic history of an area to the stratigraphic record. In the past most geological interpretations ignored or did not reflect the role of hiatus and degradation; yet, the missing units record diastrophic events at least equal to and sometimes of greater importance than depositional events.

Effects of diastrophism

Subdivision of the Precambrian. Cryptozoic time (prior to 570,000,000 years ago) presents very great problems for classification and subdivision. The problems arise principally because Cryptozoic rocks, although similar to Phanerozoic rocks in other respects, lack either diagnostic or common fossil remains. They can be readily classified as to their rock-stratigraphic classification but not as to the time, time-stratigraphic, or biostratigraphic classifications. So-called systems and series of some Precambrian rocks are not time-stratigraphic terms, but rock-stratigraphic. Precambrian time constitutes about 80 percent of all known geologic time. Because fossils do not allow the recognition of subdivisions of a time-stratigraphic succession of rocks, it seems probable that radiometric dating will be the only possible solution to an orderly stratigraphic classification and correlation of Cryptozoic rocks.

Precambrian systems

One of the classical attempts at subdivision of the Cryptozoic was attempted by the Canadian Geological Survey under Sir William Logan in 1863. Two systems were defined in eastern Canada. The older, the Laurentian, was composed of gneissic (metamorphosed) granitoid rocks, and the younger system of phyllitic and schistose rocks. The younger was named Huronian and was supposedly Cambrian in age. Subsequent studies in Canada and the United States have indicated that not only was the Huronian older than Cambrian, but that its classification is infinitely more complex than indicated. Other system names proposed find limited applicability within certain regions, but have not proven to be valid when extended elsewhere. Lake Superior studies resulted in two formally described systems, the older Archean and the younger Algonkian eras. Today, these subdivisions are considered to be of much greater magnitude than systems; the terms Archeozoic and Proterozoic are now preferred for these subdivisions. They are used in the sense of era terms, although not so defined. Other names in use find local acceptance, but none is truly time-stratigraphic or time terms except coincidentally.

Climatic units. The use of geological climatic units has been recommended when these can be established from inferred widespread climatic episodes of the Quaternary. The defining criteria are contained in the rocks or soil either in the form of minerals, type of weathering, or kinds of fossils. Although local stratigraphic boundaries may be isochronous (*i.e.*, represent equivalent values of time), it is unlikely that the boundaries defining geologic time units in different latitudes or at different elevations will be, because climate varies with both latitude and elevation and has so varied in the past. Glaciation and interglaciation are fundamental units of the geological climate classification. Glaciation refers to a period of intensive ice action marked by characteristic erosional or depositional activity. Interglaciation refers to the interval between two glacial advances as marked by characteristic erosion, deposition, and weathering. Stade and interstade represent subdivisions of times of glaciation and interglaciation, respectively: stade is the amount of time represented by a glacial deposit, and interstade is that represented by an interglacial deposit.

Remanent magnetism and magnetic reversals

Paleomagnetism and chronology. Paleomagnetism is the residual magnetism that remains in some kinds of rocks at the time of their formation. Such magnetism is termed remanent and may be produced in a variety of sedimentary or igneous rocks that contain one of several magnetic minerals. These minerals, such as magnetic iron oxides, align their magnetic polarity with the Earth's magnetic field at the time of the rocks' origin. Such original remanence can be measured to indicate the declination, or direction, of the Earth's ancient magnetic field and its inclination, or angle with respect to the horizon, at the collecting site. By making a large series of measurements of rocks of many ages, it has been determined that remanent magnetism of rocks found no earlier than the Late Tertiary reflect fairly well the position of the mag-

netic poles of the Earth today. As one measures magnetism of older rocks, however, one discovers that rocks of different ages seem to reflect differing positions of the magnetic poles. If the Earth's magnetic field has been a dipole (*i.e.*, similar to the present field, which is best represented by two principal magnetic poles) throughout geologic time, and if the residual magnetism found in rocks was induced at the time of their formation, it would appear that the poles have gradually changed position with respect to that of the continents. This polar wandering, as it is termed, can be most simply explained by assuming that some or all of the continents involved have changed their relative position subsequent to the formation of remanent magnetism in their rocks. Such displacements of the geomagnetic pole with respect to the present rotational axis can be measured at many times through the geological past, and can be used to construct polar-wandering curves showing the curving path of the computed position of the magnetic poles at different points in geological time. It should be possible, then, to date, at least in a general way, a rock with remanent magnetism of unknown age by determining the position of the geomagnetic pole at the time of its formation and comparing it with the polar-wandering curve. Where paleontologic data or radiometric dates are not available, such time-stratigraphic or time units would be of value, even if the accuracy were not as great as that obtained under more favourable conditions. Their value for Cryptozoic correlations, for instance, once a suitable polar-wandering curve was established, would be enormous.

Many deep-sea cores from certain oceanic areas have been correlated on the basis of comparison of polarity changes between adjacent cores. Most such correlations have been based on the additional phenomenon of magnetic reversals. It has been discovered that remanence in certain rocks is directed about 180° from the predicted direction. Although it is not completely clear why such reversals take place, a widely accepted theory explains them as occasional reversals in the direction of the electric currents that are responsible for the geomagnetic fields or that, additionally, some rocks have the property of self-reversal of their fields. In the drill cores cut from the ocean bottom a succession of alternating normal and reversed polarities sometimes occurs. When coincidence in successions is found between cores from different locations, a correlation is assumed.

Dendrochronology and coral growth. The stratigraphic and paleontologic literature also contains other useful methods of classifying or dating stratigraphic units. Methods of dating geologic events by using an absolute indication of age are extremely varied and employ physical rather than organic phenomena, but the latter sometimes provide a record of seasonal changes. The growth patterns of some plants and animals show such changes; perhaps tree rings are the best known of these. Dating by tree rings can extend back several thousand years under ideal conditions. Certain animals also respond to seasonal changes by showing lines of concentric growth or regularly added shell or bone layers. The regular concentric ridges found on many extinct Paleozoic horn corals probably represent increments of yearly growth. By counting such ridges the age of a horn coral can be determined. Modern corals possess very fine striations between the ridges, which represent regular increments of daily growth, and Paleozoic corals also exhibit such "daily" growth striations. Growth ridges on corals suggest that the Devonian year had more days than a modern year, and that a year in the Pennsylvanian Period had fewer days than the Devonian but more than the present. These results are remarkably close to other calculations that indicate occurrence of a slow decrease in the Earth's rate of rotation throughout geologic time because of friction caused by the tides.

Seasonal growth of plants and animals

Zonules and faunules. Stratigraphic successions may be subdivided by defining units on the basis of the inferred paleoecology of the fossils and the environment of sedimentary deposition. The term zonule has been defined as the subdivision of a biostratigraphic subzone, but also has long been used in the sense of an inferential paleoecologically controlled unit. As defined originally, a zonule was a group of rocks containing a faunule, or fossil community, the horizontal and vertical extent of such a zonule in strata being determined by the horizontal and vertical extent of the faunule. Such zonules were dependent for their existence on both biological and physical-chemical attributes of the environment. Under favourable ecologic conditions a zonule might recur several times in geologic time. Thus zonules are paleoecologic in significance rather than time-stratigraphic. The analysis of such zonules in fossils from stratigraphic successions is basic to the discrimination of temporal fossil range from simple occurrence in strata. So-called ecostratigraphic units, or ecozones, are vertically successive units based on paleoecologic analysis.

Quantitative correlation. An ingenious method of quantitative paleontologic correlation for two or more local successions of rocks has been described by the American geologist Alan Shaw. The method involves the detailed collection of fossils, carefully allocated by stratigraphic position to each fossil section. A check list establishes the range zone of all species for each locality and indicates the stratigraphic distance of first and last occurrences above a datum, such as the base of a formation. If the sections to be correlated are closely spaced, facies may be no problem. First and last occurrences of species may be almost synchronous, in the event of similar facies, and, of course, any synchronous events, whether or not organic, such as ash falls, would also be plotted. If time-coincident points measured from each datum are plotted on a graph in which one axis represents one section and the second represents the one to be correlated, they will form a series of points along a sloping line. Because species rarely will appear simultaneously, even in geographically closely spaced successions, because of facies, there is usually no actual coincidence of points, but rather a scatter. Only points that are actually synchronous will fall exactly along the line.

Plotting of time-coincident points

Equations can be established for the correlation line,

From A.B. Shaw, *Time in Stratigraphy*, copyright 1964; used with permission of McGraw-Hill Book Company

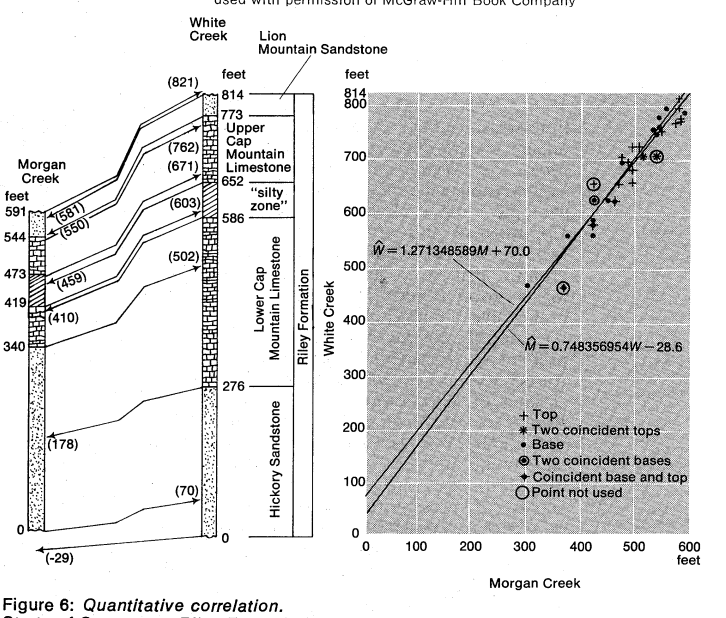


Figure 6: *Quantitative correlation.*
Strata of Cretaceous Riley Formation at White and Morgan creeks, Texas, correlated by quantitative paleontological techniques. (Left) Lines correlating rock boundaries in each section. (Right) Statistical correlation of thickness data.

thus allowing the correlative points in each section to be located. By analyzing a great number of local sections throughout a biogeographic province, for instance, the teilzones or possibly biozones of a number of species could be compiled from analysis of their local range zones. These teilzones or biozones when plotted against a composite standard section would be useful information for geologists engaged in regional stratigraphic analysis.

The large number of mathematical calculations that are necessary for computing the equations of correlation for such regional studies make access to calculating machines and to computers a great asset.

The left portion of the diagrams in Figure 6 depicts two local rock successions that have been correlated by Shaw's technique. The lines leading from one column of rocks to the adjacent column indicate time-equivalent correlations based on the correlation equation. It will be seen that the deposition of the Hickory Sandstone at Morgan Creek did not begin as early in geologic time as at White Creek. Limestone deposition began at White Creek while sandstones were still being deposited at Morgan Creek. More sediment was deposited at White Creek than at Morgan Creek during the same time interval. The graph at the right of Figure 6 illustrates the correlation lines drawn through the coincident and near-coincident points of the lithologic successions shown in the left part of Figure 6. The points plotted on the graph are based on first and last occurrences of species in the two sections. The decimal figures given for White Creek ($\hat{W}$) and Morgan Creek ($\hat{M}$) are the computed values of the equation of correlation and relate to the x and y axis of the sloping correlation lines.

This method, like that of Lyell's percentage correlation, avoids problems of faunal and lithologic facies when carried out over a biogeographic region because of the numerous observations throughout all facies of a province. Other methods for statistical correlation and for devising quantitative classification schemes for strata have been less successful as a rule. Usually such quantitative measures depend upon a sufficiently large and well enough preserved statistical sampling of paleontologic taxa to produce reliable confidence limits. Rarely, except when collecting has been intensive, are samples either large enough or well enough preserved for such purposes.

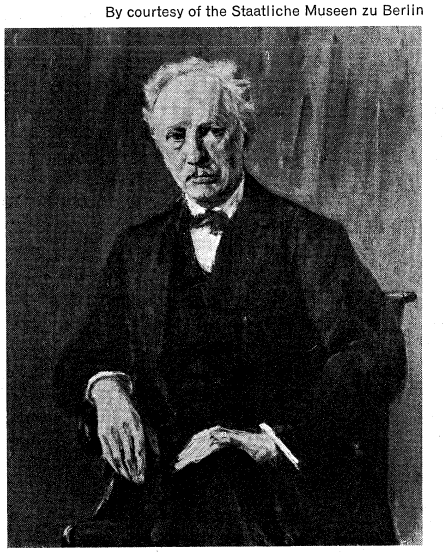
BIBLIOGRAPHY. The following works provide background material for the many topics treated in this article: F.D. ADAMS, *The Birth and Development of the Geological Sciences* (1938, reprinted 1954), a narrative review of geologists and the development of important ideas from the classic origins to the beginning of modern times; AMERICAN COMMISSION ON STRATIGRAPHIC NOMENCLATURE, "Code of Stratigraphic Nomenclature," *Bull. Am. Ass. Petrol. Geol.*, 45:645–665 (1961), the codification of biostratigraphic, time-stratigraphic, geologic time, and rock stratigraphic terms; W.B.N. BERRY, *Growth of a Prehistoric Time Scale Based on Organic Evolution* (1968), an authoritative and well-written account of the role of organic evolution in the geologic time scale, time-stratigraphy, and biostratigraphy; C.O. DUNBAR and JOHN RODGERS, *Principles of Stratigraphy* (1957), an excellent introductory presentation of stratigraphic principles in a standard college textbook, including a good description of sedimentary rocks and a comprehensive bibliography of both general and specific articles; D.L. EICHER, *Geologic Time* (1968), an elementary, but comprehensive treatment of geologic time and its classification; HENRY FAUL, *Ages of Rocks, Planets and Stars* (1966), a discussion of astronomic and radiometric dating; A.W. GRABAU, *Principles of Stratigraphy*, 2nd ed. (1924; reprinted in 2 vol., 1960), a classic and monumental work and a pioneering endeavour in presenting principles and methodology (although many concepts are outdated, the thoroughness and clarity of presentation make this an important historical reference); JAMES HUTTON, *The Theory of the Earth, with Proofs and Illustrations*, vol. 1–2 (1795, reprinted 1959), vol. 3 (1802), an important classic with initial statement and proofs of uniformitarianism; W.C. KRUMBEIN and L.L. SLOSS, *Stratigraphy and Sedimentation*, 2nd ed. (1963), a standard and much employed college textbook with excellent coverage of the majority of classical and modern phases of stratigraphy, including its historical development—also contains a complete bibliography arranged by topic for easy reference and supplementary readings with each chapter; BERNHARD KUMMEL, *History of the Earth*, 2nd ed. (1970), a college textbook of historical geology with discussions of the formations, fossils, and inferred geologic history of the world (emphasis on North America); CHARLES LYELL, *Principles of Geology*, 3 vol. (1830–33, reprinted 1969), a classic in the field of geology that laid the foundation for the science—vol. 3 contains the first subdivision of the Tertiary on the basis of percentage of living to fossil forms; V.S. MALLORY, "Biostratigraphy: A Major Basis for Paleontologic Correlation," *Proc. N. Am. Paleont. Conv. 1969*, pt. F, pp. 553–566 (1970), a comprehensive review of biostratigraphical concepts and their major refinements, with a complete bibliography of important classical works; A. OPPEL, "Die Juraformation Englands, Frankreichs und des südwestlichen Deutschlands," *Jh. Ver. Württ. Natur.* (1856–58), the classical work in biostratigraphy and time-stratigraphy that is the major foundation for all subsequent definitions of stage and zone; F.A. VON QUENSTEDT, *Der Jura* (1856–58), a classical work in which d'Orbigny's "stage" is critically evaluated, and the methodology of arriving at fossil range through species checklists is depicted; H.G. SCHENCK and S.W. MULLER, "Stratigraphic Terminology," *Bull. Geol. Soc. Am.*, 52:1419–26 (1942), a short paper, but a most important one that was influential in shaping the modern stratigraphic classification; A.B. SHAW, *Time in Stratigraphy* (1964), an exposition of statistical methods of correlation of strata; NICOLAUS STENO, *De solido intra solidum naturaliter contento dissertationis prodromus* (1669; Eng. trans., *The Prodromus of Nicholaus Steno's Dissertation Concerning a Solid Body Enclosed by Process of Nature Within a Solid*, 1916; new ed., 1968), the historically important first presentation of the principle of superposition of strata; CURT TEICHERT, "Some Biostratigraphical Concepts," *Bull. Geol. Soc. Am.*, 69:99–120 (1958), a comprehensive review of terminology and a glossary of many terms used by stratigraphers; J.M. WELLER, *Stratigraphic Principles and Practice* (1960), a well-balanced treatment of stratigraphy, its historical development, principles, and methodology—the properties of sedimentary rocks are emphasized and a bibliography arranged by chapter is included.

(V.S.M.)

Strauss, Richard

A musical iconoclast, Richard Strauss was the leading German composer-conductor in the early part of the 20th century and the last of the great Romantic composers. His reputation is based chiefly on his tone poems, such as *Don Juan* and *Also sprach Zarathustra* (*Thus Spoke Zarathustra*), and on his operas, the best known of which are *Der Rosenkavalier* (*The Knight of the Rose*), *Salome*, and *Elektra*.

By courtesy of the Staatliche Museen zu Berlin



Richard Strauss, portrait by Max Liebermann, 1918. In the National-Galerie, East Berlin.

Early life and works. Richard Georg Strauss was born on June 11, 1864, in Munich. His father, Franz Strauss, was Germany's most celebrated French-horn player, and his mother, Josephine Pschorr, was the daughter of a successful Munich brewing family. At the age of four he began piano lessons. When he was six, he took violin lessons and composed his first song, "Weihnachtslied" ("Christmas Song"). He began receiving instruction in theory and composition when he was 11. Immature works followed quickly and naturally with praise and no pressure, so that his musical development was second nature, whereas his academic learning needed effort and concentration. He was devoted to his family, which lived comfortably, though without luxuries. They were depen-

dent upon the paternal income, which was derived from playing and teaching and from conducting Munich's semiprofessional orchestra, the Wilde Gung'l, in which Richard played the violin for a few years, thus early acquiring orchestral discipline and experience.

In 1882 Strauss began a course of philosophy, aesthetics, and the history of art. His private reading on these subjects was sustained for the rest of his life. A meeting with the celebrated conductor Hans von Bülow led to Strauss's conducting (for the first time and without rehearsal) his own *Suite for 13 Winds*, Opus 4. As a result he went as Bülow's assistant to the Meiningen court orchestra in October 1885; a month later, following Bülow's departure, he found himself in complete charge. He stayed with the orchestra for only six months and quickly absorbed the basic orchestral repertory of the time. Ambitious for a more prominent engagement, he went as third conductor to Munich. Strauss's musical mentors were Richard Wagner (early works) and Johannes Brahms. At first his understanding of the later Wagner operas was discouraged by his father, but because of the influence of his philosopher friend, Alexander Ritter, Strauss overcame his father's prejudices. In 1887 in Munich he conducted his *Aus Italien* (*From Italy*), Opus 16, a symphonic fantasy based on his own vivid impressions of Italy. It was booed by many of the audience—to Strauss's great satisfaction, for he at last knew he had created an impression. In the summer of 1889 he worked as a musical assistant at the Wagner festival in Bayreuth and in October moved to Weimar, because Munich had become too constricting for him and he could not bear the jealousy of his superiors. The poor opinion of *Aus Italien* was overcome in November at the Court Theatre in Weimar, when the tone poem *Don Juan* burst triumphantly on an astounded audience. It was the start of Strauss's successful composing career. Cosima Wagner, the composer's wife, who had ruled Bayreuth since Wagner's death in 1883, had designs on Strauss—as a senior member of her staff and as a son-in-law; but he was already courting Pauline de Ahna, daughter of a Prussian general, whom he had first met in 1887.

Conductor at Meiningen and Munich

Mature life and works. In 1890 Strauss was a guest conductor with the Berlin Philharmonic Orchestra. He suffered two serious illnesses, caused by overwork, in 1891–92. He went to the Mediterranean to convalesce and discovered a strong mental and spiritual affinity with Greece. In Egypt at Christmas 1892, he finished his first opera, *Guntram*, which was totally Wagnerian in concept. It was produced in Weimar in May 1894, when Strauss announced his engagement to the leading soprano, Pauline de Ahna. They were married four months later. In October Strauss obtained a senior appointment as court conductor at Munich, where he and his wife set up house and lived for the next four years. Strauss also obtained another year's engagement with the Berlin Philharmonic Orchestra, and so, at 30, he was well established as a composer and conductor of ability and promise. As soon as they moved to Munich, Pauline Strauss ordered her husband's way of life so that time for composition always came first and was never interrupted to make room for less important matters. They loved each other dearly, although her eccentric and overbearingly tasteless behaviour had given rise to much criticism and many anecdotes. He was inclined to be lazy but grew to rely on her management of him, without which he might never have written as much as he did. At the same time, she was as devoted to him as he was utterly faithful to her. Their marriage existed more on a mental than a physical understanding, and Strauss's wedding present to Pauline, the set of songs, *Vier Lieder*, Opus 27, is a true expression of this. They had one child, Franz Alexander, born in 1897.

Marriage to Pauline de Ahna

For the next 13 years Strauss's frequent conducting tours proclaimed his remarkable stamina. He travelled mainly in Europe, but he also visited Russia, and, in America in 1904, he gave the first performance of his *Sinfonia Domestica*—like the earlier *Ein Heldenleben* (*A Hero's Life*), an autobiographical tone poem. By the turn of the century he had composed seven tone poems and 85 songs, a practical and useful way of learning how to set words with the maximum effect, which showed in his second opera, *Feuersnot* (*Fire-Famine*; Dresden, 1901), far removed in style and effect from *Guntram*.

Strauss was delighted at the receipt of an honorary Doctor of Philosophy degree from the University of Heidelberg in 1903, the idea of which greatly appealed to him; and in this year he began to work for the rights of German composers, a selfless task that he undertook voluntarily and in face of opposition from music publishers. His two operas *Salome* and *Elektra* confirmed him as Europe's most advanced composer. With his royalties from the first year's performances of *Salome*, Strauss built himself an ideal villa at Garmisch, where he hoped to live and compose from the age of 50. In 1908 he was appointed musical director of the Berlin Court Opera.

Elektra had been Strauss's first work written to a libretto by Hugo von Hofmannsthal, then Austria's most talented and interesting poet. The success of their next collaboration, *Der Rosenkavalier*, earned them both a fortune, and Strauss invested his funds in England. The alliance with Hofmannsthal was, however, a mixed blessing. Strauss, the Bavarian, aimed to write music that was universally understood and appreciated; Hofmannsthal was an intellectual and an aristocrat to whom the masses were anathema. His philosophical symbolism from *Ariadne auf Naxos* (1912) and onward was not for the general public. Had Strauss been able to control his librettist's direction or to insist on the choice of subjects for their operas, he might well have explored further than the few moments of atonality in *Elektra* and into the realm that seemed from then on denied him.

Partnership with Hofmannsthal

Josephslegende (*Legends of Joseph*) was an hour-long ballet commissioned from Strauss by the Russian impresario Sergey Diaghilev. It displayed a Strauss entirely uninterested—composing from the wealth of his astonishing musical technique, not from his heart. He conducted the ballet in Paris and in London in the summer of 1914, immediately after which war was declared between England and Germany. His fortune in England was sequestered and his plans for retiring at the age of 50 were aborted. He was thus obliged to continue with heavy conducting schedules to help re-establish his finances.

Strauss did not understand politics, nor did he consider himself to be personally involved. He refused to allow his name to be used for propaganda (although he wrote for the Kaiser), and considered the whole affair a nuisance, an unnecessary interruption of his and Hofmannsthal's work on their next opera, *Die Frau ohne Schatten* (*The Woman without a Shadow*). During the war Strauss was engaged on three further tasks: a large, descriptive work of the symphonic poem genre, the *Alpensinfonie* (*Alpine Symphony*; 1911–15), Opus 64; the founding, together with Hofmannsthal and the theatrical producer Max Reinhardt, of a Salzburg Festival Association, a far-sighted enterprise; and the revision of *Ariadne auf Naxos*, separating it from Hofmannsthal's translation of Molière's *Le Bourgeois Gentilhomme* to which it had formerly been attached. This brought the Strauss–Hofmannsthal relationship to its lowest ebb, with the poet being hurtfully rude and the composer remaining good-natured and generous.

After the armistice in 1918 Strauss was called temporarily to Berlin to direct the State Opera; then he moved to Vienna to become co-director (with Franz Schalk) of the opera there. *Die Frau ohne Schatten* failed to meet public taste in an impoverished Vienna of 1919.

Late life and works. Over the next 20 years Strauss's prestige as a composer declined along with his health. He could not achieve artistic and financial successes of the order of *Salome*, *Elektra*, or *Der Rosenkavalier* and had long since yielded his position as Europe's leading composer; but he was still enormously successful as a conductor. By the end of the 1920s he had reduced his movements on the podium to a minimum, although his control was still complete. He was tall, with a large, domed and bald forehead, pale blue eyes, and a seemingly aloof and haughty manner that belied his true, kindly nature.

Decline in Strauss's influence

The operas of the interwar years, with the possible exception of *Intermezzo* (1924), lacked the authority of their predecessors. This was an extension of the earlier *Heldenleben—Sinfonia Domestica* idea, in that it presented Strauss and Pauline as they really were: through Strauss's eyes. It also showed him attempting an altogether new kind of sung speech and precisely judged orchestral texture and vocal balance, all clearly described in an essay in the score. The work was to his own libretto, but he returned to Hofmannsthal for *Die ägyptische Helena* (*The Egyptian Helen*; 1928) and *Arabella* (1933). Work on the last opera revived the friendship and rapport between them; unhappily, Hofmannsthal died suddenly after completing the libretto of the latter work. Although he may have led Strauss away from more suitable texts, together they achieved a great deal, and it is significant to note that all Strauss's operas after 1928 contained elements of Hofmannsthal's invention.

Collaboration with Zweig

In 1932 another Austrian poet, Stefan Zweig, became Strauss's librettist, and their collaboration on *Die schweigsame Frau* (*The Silent Woman;* 1935) progressed well. By 1933 the Nazi Party was in command of Germany. Toscanini, then a leading conductor in Europe, strongly disapproved of the Nazis and refused to conduct Wagner's *Parsifal* at Bayreuth. Strauss replaced him and also took over a Berlin concert from Bruno Walter, who as a Jew was unacceptable to the Nazis. In this instance Strauss gave his fee to the orchestral orphans' fund. He found his own actions high-principled and defensible for the saving of artistic enterprises; others saw them differently. In November he was made president of the Reichsmusikkammer (Chamber of State Music), which he accepted only because he felt that by taking the position he could exercise an influence that would be advantageous to musicians.

For Strauss there were only two kinds of humans—the talented and the rest. In a letter to Zweig in connection with the production of their opera in Dresden in 1935, he propounded these views at the expense of the Nazis. The letter was intercepted, Strauss was forbidden to work with Zweig, who was Jewish, and further performances of *Die schweigsame Frau* were banned. Strauss was not permitted to conduct anywhere in Germany or in Salzburg, and for a year he was in a political vacuum. Zweig had managed to escape to South America, and this event affected Strauss profoundly. He was not only again without a librettist, but the significance of the fact that his daughter-in-law was Jewish was now brought forcefully to him. He was forced to resign his office but was permitted to conduct his new *Olympische Hymne* at the opening of the Olympic Games in Berlin, in July 1936. In November he went to London, with the Dresden State Opera, where he received the Gold Medal of England's Royal Philharmonic Society.

Munich had been giving model performances of his operas for some years, and they now produced two one-act operas: *Friedenstag* (*Day of Peace*) and *Daphne* (both 1938) to librettos by his latest, distinguished librettist, Joseph Gregor. His 75th birthday celebrations in 1939 were not hindered, despite the fact that he had fallen from Nazi grace. World War II marked a return for Strauss to the bitterness of 1914–18. Then it had been the loss of his fortune; now it was loss of respect and real danger to his family. He did everything—including self-effacement before the local Nazi official—to protect them, and he succeeded. Strauss was forbidden to go abroad for imperative Swiss cures and spent a miserable four years between his villa at Garmisch and his house in Vienna. He was shocked by the destruction of his four most beloved opera houses (Berlin, Munich, Dresden, and Vienna) and never really recovered from these events.

Between 1942 and 1948 he composed a number of works in which he returned to the power and sweetness of former years. Immediately after the war was over, Strauss went to Switzerland, where he and Pauline stayed, penniless, for four years, deaf to calls to attend denazification trials. He was found guiltless and was invited to London, where he conducted at a festival of his music in 1947. He returned to Germany in 1949 and died on September 8 in his house at Garmisch-Partenkirchen, in Bavaria, leaving the richly scored, poignantly retrospective *Vier letzte Lieder* (*Four Last Songs;* 1948) as his final completed composition. He was the last Romantic composer of the 19th century, a master of orchestration, a composer of distinction, a magnificent conductor, and a patriot, faithful to the soil of Bavaria rather than to Germany's political leaders.

Strauss, having begun as an avant-garde composer in his early tone poems and operas (*Salome* and *Elektra*), became in later years a conservative, unaffected by the advances and experiments going on around him. In his last years his style became leaner, more classical in the Mozartian sense; indeed, the opera *Capriccio* (1942), *Metamorphosen*, for 23 solo strings (1944–5), and the *Four Last Songs* may be said to have achieved a perfect fusion of the Romantic German and the Neoclassical manner. His reputation has fluctuated wildly over the years, and his music is apt to provoke violent advocacy or dislike. In spite of his conservatism, he remains one of the 20th century's controversial figures.

MAJOR WORKS

OPERAS: 15 including *Salome*, op. 54 (first performed 1905), *Elektra*, op. 58 (1909), *Der Rosenkavalier*, op. 59 (1911), *Ariadne auf Naxos*, op. 60 (*Adriadne on Naxos*, 1912, rev. 1916), *Die Frau ohne Schatten*, op. 65 (*The Woman Without a Shadow*, 1919), *Intermezzo*, op. 72 (1924), *Die ägyptische Helena*, op. 75 (*The Egyptian Helen*, 1928), *Arabella*, op. 79 (1933), *Die schweigsame Frau*, op. 80 (*The Silent Woman*, 1935), *Friedenstag*, op. 81 (*Day of Peace*, 1938), *Daphne*, op. 82 (1938), *Die Liebe der Danae*, op. 83 (*The Loves of Danae*, 1952), and *Capriccio*, op. 85 (1942).

BALLETS: *Josephslegende*, op. 63 (*Legends of Joseph*, 1914); *Schlagobers*, op. 70 (1924).

INSTRUMENTAL WORKS: 7 symphonic poems, including *Don Juan*, op. 20 (composed 1889), *Till Eulenspiegels lustige Streiche*, op. 28 (1894–95), *Also sprach Zarathustra*, op. 30 (*Thus Spoke Zarathustra*, 1895–96), and *Ein Heldenleben*, op. 40 (*A Hero's Life*, 1898); *Don Quixote*, op. 35 (1897), fantastic variations based on Cervantes; *Sinfonia Domestica*, op. 53 (*Domestic Symphony*, 1903); *Alpensinfonie*, op. 64 (*Alpine Symphony*, 1911–15); 2 horn concertos (op. 11, 1882–83; 1942); *Burleske* for piano and orchestra (1885); oboe concerto (1945–46); *Metamorphosen* for 23 strings (1945–46).

SONGS: About 200, including *Vier letzte Lieder* (*Four Last Songs*, 1948), for soprano and orchestra.

BIBLIOGRAPHY

In English: NORMAN DEL MAR, *Richard Strauss*, 3 vol. (1962–72), the standard, technical source of information on the compositions; WILLIAM S. MANN, *Richard Strauss: A Critical Study of the Operas* (1964), a detailed study of the operas; ERNST KRAUSE, *Richard Strauss: Gestalt und Werk* (1963; Eng. trans., *Richard Strauss: The Man and His Work*, 1964), the most convenient single-volume book; *The Correspondence Between Richard Strauss and Hugo von Hofmannsthal, 1907–1918* (1927, reprinted 1961), an illuminating description of the way in which the two men worked—gives a clear profile of their characters and adds much to our knowledge of the 7 major works that they created; *The Correspondence Between Hans von Bülow and Richard Strauss* (1955), the only translation of any of Strauss's letters during the period 1873–93; *Betrachtungen und Erinnerungen* (1949; Eng. trans., *Recollections and Reflections*, 1953), sharp and often cynical musings by Strauss, written near the end of his life; ALAN JEFFERSON, *The Leider of Richard Strauss* (1971), an analysis in great detail of 50 songs with general information about them all.

In German: MAX STEINITZER, *Richard Strauss* (1911; 4th ed., 1927), contains much unique observation on the early works especially, but it stops 22 years before the end of Strauss's life; FRANZ TRENNER (ed.), *Richard Strauss: Dokumente seines Lebens und Schaffens* (1954); WALTER PANOFSKY, *Richard Strauss: Partitur eines Lebens* (1965), both contain new material but are not in themselves comprehensive; MUELLER VON ASOW, *Richard Strauss: Thematisches Verzeichnis* (begun in 1955 and not yet completed), contains precise and elaborate details of all the known compositions; FRANZ GRASBERGER and FRANZ HADAMOVSKY (comps.), *Richard Strauss-Ausstellung zum 100. Geburtstag* (1964), a useful checklist and source of elusive information. The main Strauss Archive is in the composer's house at Garmisch, Bavaria.

(A.R.J.)

Stravinsky, Igor

Undoubtedly one of the greatest composers of the 20th century, Igor Stravinsky dazzled successive generations of musicians with his protean ability to change and renew his musical thought. Frequently, the work of a great composer establishes an immediate currency in his native country and only gradually acquires a wider significance in the context of the world's musical heritage. The case of Stravinsky was at variance with this general rule, and his achievement must be assessed in light of the fact that at least twice in the course of his long life he changed his country, his nationality, his language of words, and his musical idiom. As a Russian, he lived in St. Petersburg and (in exile) in western Europe; as a Frenchman, he lived mainly in Paris; and as an American, he lived mainly in Hollywood. Until the end of World War I his musical idiom was closely related to Russian sources, and his music exercised a somewhat exotic appeal on audiences outside Russia. The Neoclassical works of his middle period were more obviously rooted in traditional western European musical procedures. The serial idiom (*i.e.*, music featuring rows of tones) that he adopted during the last 15 or so years of his life produced shortish works of comparatively dense musical content, which may take some time to establish themselves in the international repertory. His working life as a composer covered about two-thirds of a century; and the originality and the strength of his compositions were maintained throughout this period.

By courtesy of the Gernsheim Collection, Humanities Research Center, The University of Texas at Austin; photograph, Felix H. Man



Stravinsky, 1929.

Russian period. Igor Fyodorovich Stravinsky was born at Oranienbaum (now Lomonosov) near St. Petersburg (now Leningrad) on June 17 (June 5, old style), 1882. His father, Fyodor Ignatyevich Stravinsky, was a bass singer of great distinction, who had made a successful operatic career for himself, first at Kiev and later in St. Petersburg. Igor was the third of a family of four boys. As a child he loved music. He was accustomed to hearing his father practicing his operatic roles at home. From an early age he attended opera and ballet performances at the neighbouring Mariinsky Theatre in St. Petersburg. Regular piano lessons started when he was nine, and a little later he received training in harmony and counterpoint. Despite his obvious bent for music, his parents refused to allow him to pursue a musical career. Sent to St. Petersburg University to study criminal law and legal philosophy, he was graduated in 1905.

Meanwhile, his musical interests had turned to composition. In the summer of 1902, a few months before his father's death, he obtained an introduction to the Russian composer Nikolay Rimsky-Korsakov through one of Rimsky-Korsakov's sons, who was also a student at the university. Rimsky-Korsakov was sufficiently interested in the young man's early attempts at composition (of which an interesting example has survived in manuscript, a piano *Scherzo* dated 1902) to advise him not to enter the conservatory for conventional academic training but to pursue his studies privately. This advice was followed. A year later Rimsky-Korsakov agreed to tutor him privately, mainly in instrumentation. The arrangement continued for about three years (1903–06). It was Stravinsky's habit to discuss his compositions with his mentor as they were planned and written. Rimsky-Korsakov arranged for several of these, including the *Symphony in E Flat Major* (1905–07), to be performed at private or public concerts in St. Petersburg. The last composition to be presented in this way was *Fireworks* (1908), a brief symphonic poem that Stravinsky intended as a wedding present for Rimsky-Korsakov's daughter. Rimsky-Korsakov died in the summer of 1908, before it could be performed. After his death the mourning pupil wrote a funeral dirge (1908) in memory of his master. It was performed in St. Petersburg the following season; the score is lost.

Influence of Rimsky-Korsakov

Stravinsky left the university in 1905. The following year he married his first cousin, Catherine Nossenko. They lived in two rooms of his family's apartment in St. Petersburg and spent their summers in the country at Ustilug in Volhynia, an area of the Ukraine. A son (Theodore) was born to them in 1907 and a daughter (Ludmila) a year later.

Collaboration with Diaghilev and the Ballets Russes. When *Fireworks* and an earlier orchestral piece, *Scherzo Fantastique* (1907–08), were performed in St. Petersburg on February 6, 1909, they were heard by the impresario Sergey Diaghilev (*q.v.*), who was then busy making preliminary arrangements for the summer season of his Ballets Russes to be held in Paris. He was so favourably impressed by Stravinsky's promise as a composer that he invited him to join his small group of artistic collaborators. For the 1909 ballet season Stravinsky was invited to orchestrate various pieces of ballet music, including two piano numbers by Chopin for *Les Sylphides*. For the 1910 season, Diaghilev commissioned from him a completely new ballet score, *The Firebird*.

The next few years were a period of intensely close collaboration between the two men, for Diaghilev was anxious that all of Stravinsky's major new works should be mounted by his company. After the decisive success of *The Firebird* at the Paris Opéra (June 25, 1910), Stravinsky started to write a *Konzertstück* ("concertpiece") for piano and orchestra; but, yielding to Diaghilev's arguments, he agreed to adapt the music he had already written to fit a new ballet scenario, and this work, *Petrushka*, received its first performance during the Ballets Russes 1911 season in Paris. Before this, Stravinsky had had the idea of writing a kind of primitive spring symphony to be called *Great Sacrifice;* here, too, Diaghilev persuaded him that it should be cast in the form of a ballet. The composition of *The Rite of Spring* (as it was finally called) was spread over two years (1911–13). The music of this dynamic, profoundly disturbing score created a major scandal when the ballet had its first performance at the Théâtre des Champs Élysées, Paris (May 29, 1913). Stravinsky then reverted to the task of finishing a short opera based on the Danish writer Hans Andersen's tale "The Nightingale," which he had started in 1908–09 but which had been interrupted by the commission to write *The Firebird*. This opera was now requested by the Moscow Free Theatre; but when that new venture suddenly collapsed, Diaghilev saw his chance and, taking over the work, arranged for it to be produced as part of the Ballets Russes seasons in Paris and London in the summer of 1914.

Composition of *The Rite of Spring*

That summer also marked the conception of a new ballet cantata to be called *The Wedding*, which Stravinsky decided to base on Russian peasant themes and customs. Had World War I not intervened, this new score might have been ready for production by the Ballets Russes in 1915 or 1916. Although the actual composition was completed by 1917, the final form of the instrumentation was not decided on until 1923.

Stravinsky's close connection with Diaghilev and the Ballets Russes meant that during the five years 1910–14 he spent a considerable part of his time outside Russia. He was usually in Paris for the company's summer sea-

sons and occasionally followed them on tour to Rome, Berlin, Vienna, Budapest, and London. He still managed to spend part of each summer in his Russian country home at Ustilug, the quiet atmosphere of which he found conducive to composition. His visits to St. Petersburg, however, became increasingly rare. The health of his family was precarious; and a tendency to tuberculosis made the climate of Switzerland attractive. His second son (Soulima) was born at Lausanne, Switzerland, in 1910, his second daughter (Milena) at Leysin in 1914. Parts of *The Rite of Spring* and *The Nightingale* were written at Clarens, Switzerland, in the intervening years, and the war years were spent entirely in Switzerland, first at Clarens and later at Morges.

Years of exile. These were years of isolation that ultimately led to years of exile, for as the war progressed, Stravinsky became cut off, not only from Russia but also from the Ballets Russes—the European engagements of which had to be abandoned and replaced by an American tour—and from his music publishers who had their headquarters in Berlin. He tried to overcome these difficulties in various ways. He found a local publisher in Geneva for many of his wartime compositions. In collaboration with the Swiss novelist Charles Ferdinand Ramuz he created *The Soldier's Tale* (1918), an entertainment "to be read, played, and danced" which was intended to be played on tour by a small travelling theatre. The project collapsed after a successful first performance in Lausanne, owing to a sudden epidemic of Spanish influenza.

French period. As soon as the war was over, Stravinsky decided to move from Switzerland and settle in France; during the next 20 years (1920–39) he lived in various places there—Biarritz, Nice, Voreppe, and Paris. These years were marked by an important change in his music—the abandonment of the Russian features of his earlier style and the adoption of a Neoclassical idiom. Such a radical change cost him a great effort and only after several years of what he called "samplings, experiments, amalgamations" did he find his way to new works such as *Oedipus Rex* (1927) and the *Symphony of Psalms* (1930), capable of holding their own against such masterpieces of the Russian period as *The Rite of Spring* and *The Wedding*.

Adoption of Neoclassicism

In the early postwar years, Stravinsky's ties with Diaghilev and the Ballets Russes were renewed. They were now on a much looser basis, however, than in the prewar years, for he saw that in the case of an itinerant company, without a firm base, there could be no guarantee of permanence. The only new ballet commissioned by Diaghilev from Stravinsky was *Pulcinella* (1920), the score of which consisted of music by Giovanni Battista Pergolesi that was arranged by Stravinsky and adapted to a Neapolitan commedia dell'arte scenario. *Apollo Musagetes* (1928) was the last new ballet of his to be mounted by the Ballets Russes. The following year (1929) Diaghilev died, and his ballet company melted away almost overnight.

By this time Stravinsky was no longer so dependent on Diaghilev and the Ballets Russes as he had been previously. The need to build up his income as a result of the loss of his private property in Russia had led to his embarking on subsidiary careers as concert pianist and conductor. Some of his new works, such as the *Concerto for Piano and Wind Instruments* (1923–24), *Sonata* for piano (1924), *Serenade in A Major* for piano (1925), *Capriccio* for piano and orchestra (1929), and *Concerto for Two Solo Pianos* (1935), were written primarily for himself as soloist. A considerable part of each year was devoted to touring, either as solo performer or conductor. Most of these tours were in Europe. He also made three visits to North America in 1925, 1935, and 1937 and one to South America in 1936.

Stravinsky's composition of ballet scores did not come to an end with Diaghilev's death. In the late 1920s, the Russian dancer Ida Rubinstein assembled a company of her own and commissioned two ballet scores from Stravinsky—*The Fairy's Kiss* (1928) and *Persephone* (1934). The former was based on a selection of the piano and vocal music of the Romantic Russian composer Peter Ilich Tchaikovsky. The latter was a setting of a poem by the 20th-century French writer André Gide as a melodrama with song and speech, dance, and mime. For the then recently founded American Ballet he wrote *The Card Party* (1937).

The years 1938 and 1939 were marked by a succession of family bereavements. In the autumn of 1938 his elder daughter died of tuberculosis. The deaths of his wife and mother followed in March and in June 1939.

American period. The outbreak of World War II in 1939 led to a completely new orientation of his life. An invitation from Harvard University to deliver the Charles Eliot Norton Lectures during the 1939–40 academic year gave him the opportunity to leave Europe and settle in the United States. Early in 1940 he remarried. His second wife was the artist Vera de Bosset, whom he had known for many years. When the lectures at Harvard on the poetics of music had been given, the Stravinskys travelled to California and acquired a house in Hollywood, where they were to live for more than a quarter of a century.

During the war years, two important symphonic works were composed—the *Symphony in C Major* (1938–40) and *Symphony in Three Movements* (1942–45). Whereas the *Symphonies of Wind Instruments* (1920) display an original and ingenious treatment of typical musical material from his Russian period on symphonic lines, the *Symphony in C Major* represents a summation of Neoclassical principles in symphonic form, and the *Symphony in Three Movements* successfully combines the essential features of the concerto with those of the symphony.

Symphonic works

The immediate postwar years were devoted to an attempt to regain the copyright on his early scores, which, having been originally published while he was a Russian citizen, were unprotected outside Russia because his native country had not adopted the Berne Convention. He achieved his objective by issuing revised versions of the scores concerned. These revisions amounted to little more than corrections, except in the cases of *Petrushka* (revised 1947) and the *Symphonies of Wind Instruments* (revised 1947), in which each work was given a completely new instrumentation.

During the years 1948–51 he worked on *The Rake's Progress*, a full-length, Neoclassical opera with a libretto by the Anglo-American poet W.H. Auden and the American writer Chester Kallman. When it was finished, Stravinsky returned to Europe for the first time since 1939 and conducted the first performance of the work at the Teatro la Fenice, Venice.

While he was working on this score, he invited a young American musician, Robert Craft, to help him with certain musical tasks at his home in Hollywood. The visit was a success, and Craft eventually became established in the Stravinsky household as friend, consultant, and musical assistant. This was a particularly important moment in Stravinsky's musical development because after the composition of *The Rake's Progress* he felt he had outgrown "the special incubator" in which the works of his Neoclassical period had gestated. Craft's natural liking for serial music stimulated Stravinsky into examining carefully the music of the Austrian composer Anton von Webern and also some of the works of the Austrians Arnold Schoenberg and Alban Berg, with which he was not very familiar at that time. His interest was aroused and he found his own idiom becoming profoundly affected. To begin with, his serial essays were cautious experiments carried out within a framework of tonal music, featuring note rows and procedures involving inversion, retrogression, and retrograde inversion. The *Canticum Sacrum* (1955) and the ballet *Agon* (1953–57) were larger scale works, in which the music travelled from a modal and tonal beginning (*i.e.*, with harmonies derived from traditional scales and modes) into a fully serial core, reverting at the end to the modal and tonal music of the beginning. The first of his fully serial compositions was *Threni* (1958); this was followed by *Movements* (1959), *Variations* (1964), and *Requiem Canticles* (1966), which occupy just as honoured a place

Serial works

in his output as the masterpieces of his Russian and Neoclassical periods. It is a characteristic of these serial works that in comparison they are much briefer than the composer's tonal works but seem to have a much greater musical specific gravity.

After the *Requiem Canticles*, ill health caused a slowing down of Stravinsky's compositional activity, but even as late as 1970 he was working on instrumental transcriptions of some of Bach's preludes and fugues. He died in New York on April 6, 1971, and was buried in Venice on the island of San Michele.

Contribution to music. Stravinsky's special contribution to music in the 20th century was wide and varied. He never seemed to start a composition with preconceived ideas but always examined his raw material with a fresh and critical ear. He was not prepared to accept established practice about development but preferred to subject his musical material to a personal system of tests. This had particularly interesting results whenever time, metre, and dynamics were concerned. He explored with zest the asymmetrical patterns of compound metres and, by using devices of prolongation and elision, broke down the tradition of symmetrical phrasing. He was meticulous about degrees of articulation and emphasis. He restored to music the sense of a healthy, unwavering pulse, and this helped to make so many of his compositions suitable for dancing. His music made a "clean" sound—there was no filling in merely for the sake of filling in—and after his symphonic poem *The Song of the Nightingale* (1917), his orchestral practice became mainly a question of using concertante groups of instruments with plenty of breathing space around them. He never worked to an instrumental formula, but every work, regardless of idiom, had a different instrumental specification and a different sound.

MAJOR WORKS

Stage works

OPERAS: *The Nightingale* (1914); *Mavra* (1922); *Oedipus Rex*, opera-oratorio (1927); *The Rake's Progress* (1951).

BALLETS: *The Firebird* (1910); *Petrushka* (1911); *The Rite of Spring* (1913); *Pulcinella* (1920); *The Wedding* (1923); *Apollo Musagetes* (1928); *The Fairy's Kiss* (1928); *Persephone* (1934); *The Card Party* (1937); *Orpheus* (1947); *Agon* (1957).

OTHER STAGE WORKS: *Renard*, "a burlesque in song and dance" (1916); *The Soldier's Tale*, "to be read, played, and danced" (1918); *The Flood*, "a musical play" (1962).

Vocal music

CHORAL WORKS: *The King of the Stars* (1911); *Symphony of Psalms* (1930); *Babel* (1944); *Mass* (1948); *Cantata* (1952); *Canticum Sacrum* (1955); *Threni* (1958); *Requiem Canticles* (1966).

OTHER VOCAL WORKS: *A Sermon, a Narrative and a Prayer* (1961); *Abraham and Isaac* (1963); various songs.

Instrumental music

ORCHESTRAL WORKS: *Symphony in E Flat Major* (1907); *Fireworks* (1908); *Symphonies of Wind Instruments* (1920); *Symphony in C Major* (1940); *Danses Concertantes* (1942); *Circus Polka* (1942); *Ode* (1943); *Scènes de Ballet* (1944); *Symphony in Three Movements* (1945); *Variations* (1964).

CONCERTOS: *Concerto for Piano and Wind Instruments* (1924); *Capriccio* for piano and orchestra (1929); *Violin Concerto in D Major* (1931); *Dumbarton Oaks* for 16 wind instruments (1938); *Ebony Concerto* for clarinet and jazz band (1945); *Concerto for Strings in D Major* (1946); *Movements* for piano and orchestra (1959).

CHAMBER WORKS: *Three Pieces for String Quartet* (1914); *Three Pieces for Clarinet Solo* (1919); *Concertino for String Quartet* (1920); *Octet for Wind Instruments* (1923); *Duo Concertant* for violin and piano (1932); *Elegy* for solo viola (1944); *Septet* (1952); *Double Canon* for string quartet (1959).

PIANO: *Piano Rag-Music* (1919); *Sonata* for piano (1924); *Serenade* (1925); *Concerto for Two Solo Pianos* (1935); *Sonata for Two Pianos* (1944).

Writings

Chroniques de ma vie, 2 vol. (1935; *Chronicle of My Life*, 1936), written in collaboration with Walter Nouvel; *Poétique musicale* (1942; *Poetics of Music*, 1947), in collaboration with Roland Manuel; *Conversations with Igor Stravinsky* (1959); *Memories and Commentaries* (1960); *Expositions and Developments* (1962); *Dialogues and a Diary* (1963); *Themes and Episodes* (1966); *Retrospectives and Conclusions* (1969). The last six works were written in collaboration with Robert Craft.

BIBLIOGRAPHY. *Igor Stravinsky: A Complete Catalogue of His Published Works* was issued in 1957 (new ed. with addenda, 1962). Manuscripts of Stravinsky's compositions may be found in many public institutions throughout the world; but the most comprehensive collection to date is in the Library of Congress, Washington, D.C. The *Sketches* (1911–13) for *The Rite of Spring* were published in facsimile reproduction in 1969.

From the large array of biographical and critical material the following books may be recommended: EDWIN CORLE (ed.), *Igor Stravinsky* (1949), a symposium with essays and articles by a number of distinguished contributors; PAUL HENRY LANG (ed.), *Stravinsky: A New Appraisal of His Work* (1963), critiques by six eminent musicologists dealing with various aspects of his style; MINNA LEDERMAN (ed.), *Stravinsky in the Theatre* (1949), essays and articles by various contributors on his music for the stage, with a useful bibliography; NICOLAS NABOKOV, *Igor Strawinsky* (1964), a short monograph (in German) written by a personal friend; CHARLES F. RAMUZ, *Souvenirs sur Igor Strawinsky* (1929), vivid reminiscences written by a close friend and collaborator covering the years 1915–20; ANDRE SCHAEFFNER, *Strawinsky* (1931), one of the best of the early studies (in French) of his music, particularly well illustrated; ALEXANDRE TANSMAN, *Igor Stravinsky* (1948; Eng. trans., 1949), a study of his music and aesthetics by a musician who was in close touch with him in the 1940s; ROMAN VLAD, *Strawinsky* (1958; Eng. trans., 1960), detailed analyses of most of his works up to 1958; ERIC WALTER WHITE, *Stravinsky: The Composer and His Works* (1966), a biography and a fully documented register of all his compositions up to 1965, with numerous appendixes, including a catalog of manuscripts (1904–52) in his possession, and a bibliography; ROBERT CRAFT, *Stravinsky: Chronicle of a Friendship, 1948–1971* (1972), a fascinating account by the composer's assistant detailing their activities over a period of 23 years; LILLIAN LIBMAN, *And Music at the Close: Stravinsky's Last Years* (1972), a highly personal portrait of Stravinsky's last years (1959–71) as seen by a close associate.

(E.W.W.)

Stresemann, Gustav

Gustav Stresemann, chancellor of the Weimar Republic of Germany in 1923 and foreign minister from 1923 to 1929, was the man chiefly responsible for the restoration of his country's international status after World War I. The re-entry of the German *Reich* into the family of nations after the defeat of 1918, the beginning of a reconciliation with France in 1925, the continuing settlement with the Soviet Union, and the admission of Germany into the League of Nations were largely the work of Stresemann. A Prussian German monarchist who was among the proponents of an annexationist peace during World War I, Stresemann, after the defeat of the empire, gradually became an advocate of a European settlement, no doubt in the hope that his policy would restore Germany to a position of power.

Youth and education. Stresemann was born in Berlin on May 10, 1878, the son of Ernst Stresemann, an innkeeper and beer dealer. He was the only one of five children able to attend high school and ultimately the university. He grew up in a petit bourgeois environment in which his parents were able to contribute little toward the formation of his intellect. From early childhood he displayed a strong "inclination towards solitude," as he put it as a high school student, as well as a tendency to be melancholy and to daydream.

At school he displayed an unusual gift, an "almost unhealthy partiality" for history, especially modern history. He was particularly interested in the lives of great personalities, particularly Napoleon and Goethe, both of whom fascinated him throughout his life and inspired several of his literary studies. His intense preoccupation with historical figures may well have compensated the young Stresemann for the lack of intellectual stimulation in his environment.

It was from his parents, however, that Stresemann assumed the opinions and ideals of the left-liberal middle class. Attending the universities of Berlin and Leipzig

from 1897 to 1900, he first studied literature and history but changed to economics, probably to improve his professional prospects rather than because he recognized his practical gifts.

By courtesy of the Staatsbibliothek, Berlin



Stresemann.

During his student years Stresemann belonged to a relatively progressive fraternity and retained a lifelong attachment to the sentimental glories of student life. In 1900 he received his doctorate with a dissertation entitled "The Growth of the Berlin Bottled-Beer Industry." The subject of his study, which was based on his knowledge of his father's business and dealt with the decline of a sector of small business as a result of competition from giant industry, was characteristic of his origins as well as of his point of view. It was used against him in future political struggles by the right-wing opposition.

As a liberal Prussian Protestant, Stresemann became a typical representative of the chauvinistic spirit prevailing in imperial Germany. He believed in the spiritual, military, and economic superiority of the German empire over other nations and frequently gave literary expression to this conviction. His political idealism manifested itself in a sentimental enthusiasm for the heroic liberalism of 1848, as well as in a romantic style of speech. A youthful carelessness and emotional exuberance remained lifelong characteristics.

Political rise. After completing his studies, Stresemann began his professional career in a trade association. Unusually rapid success in commerce gave him a springboard into politics. As an administrative assistant in the German Chocolate Makers' Association from 1901 to 1904, he gained a reputation as an accomplished organizer and negotiator. Already, in 1902, he had succeeded in founding the Saxon Manufacturers' Association. As its legal representative until 1911, he occupied an important position in economic life at the age of 25.

Marriage

His wife, whom he married in 1903, was the sister of a fraternity brother and the daughter of the Berlin industrialist Adolf Kleefeld. Of Jewish extraction, the charming and elegant Frau Stresemann played a leading role in Berlin society of the 1920s. The couple had two sons.

After initially sympathizing with the ideas of the Protestant social reformer Friedrich Naumann and collaborating with his National Social Union, Stresemann joined the rightist National Liberal Party in 1903. Strongly represented in Saxony at the time, the party became Stresemann's political home. Often involved in conflicts over his support of social-welfare measures with the right wing of his party (which was dominated by representatives of heavy industry), he attracted general notice at his first appearance at a party congress in 1906.

As a Dresden city councillor from 1906 to 1912 and editor of the Dresden magazine *Sächsische Industrie* ("Saxon Industry"), Stresemann became a well-known writer on economics and an expert on municipal affairs. Recognizing the importance of the press in influencing public opinion, he took advantage of it to support his aims.

Election to the Reichstag

He was elected in 1907 to the Reichstag (Parliament) as a National Liberal from the Annaberg district in the Saxon metal-mining country, and thus he succeeded in gaining a foothold in national politics. At 28 he was the youngest deputy in the Reichstag. The party chairman, Ernst Bassermann, helped to advance his political career, and he was soon considered Bassermann's "crown prince." Stresemann was primarily interested in economic policy both as a journalist and a deputy. He energetically defended the interests of the commercial middle class, but his advocacy of extended social-welfare legislation embroiled him in a conflict with the representatives of his party's right wing, which in 1912 prevented his re-election to the National Liberal Party executive committee. After losing his seat in the new Reichstag elections in the same year, he travelled with other business leaders to the United States to study economic conditions.

By this time Stresemann, who had moved to Berlin, was already one of the best-known leaders of German economic life. He occupied leading positions in a number of trade associations including the German-American Economic Association, established at his suggestion.

Stresemann's many offices brought him financial independence. He was known for his organizational gifts, knew how to handle people, and was well aware of the power he wielded. As a member of the Pan-German Deutscher Kolonialverein (German Colonial League) and an advocate of a strong naval construction program, he supported the imperialist goals of German policy carried out under the aegis of Alfred von Tirpitz and Bernhard, Fürst von Bülow. Tirpitz had served from 1897 as state secretary of the Imperial Naval Office, in which post he created the German battle fleet, and Bülow was chancellor of Germany from 1900 to 1909.

Policies during World War I. Stresemann was a part of the great majority of Germans who, in the firm belief that Germany was conducting a purely defensive war, greeted the outbreak of World War I with enthusiasm. Because of his poor health he was exempted from military service. His political hour struck, however, in December 1914, when he was returned to the Reichstag in a special election in Hanover.

During the war, Stresemann emerged as one of the most vociferous exponents of Pan-Germanism and as a champion of Germany's extensive claims on Polish and Russian territory in the east and on French and Belgian territory in the west. He virtually took over leadership of his party's Reichstag faction from Ernst Bassermann, whom military service and illness kept away from Berlin much of the time. During these years Stresemann moved increasingly to the right.

From 1916 he worked closely with the German Army Supreme Command under Field Marshal Paul von Hindenburg and General Erich Ludendorff and became their parliamentary mouthpiece. He thus advocated unrestricted U-boat warfare and opposed the policy of Chancellor Theobald von Bethmann Hollweg, who held to a moderate course and did not allow himself to be committed to expansionist war aims.

Role in dismissal of Chancellor Bethmann Hollweg

Stresemann played a leading role in Bethmann Hollweg's overthrow in July 1917 but failed to bring back to power the former chancellor Bülow, whom he admired. After Bassermann's death in the same month, Stresemann succeeded him as leader of the party's Reichstag faction, becoming chairman of the entire party later in the same year. Despite radical differences within the National Liberal ranks, Stresemann was able to prevent a party split between the Reichstag faction and its more conservative counterpart in the Prussian House of Deputies over the Prussian three-class suffrage system, in which a citizen's vote was weighted according to the value of his property. Hoping to strengthen the monarchy, Stresemann advocated abolition of the voting system. On the other hand, he allowed himself to be deceived about the seriousness of the military situation of the *Reich* and its allies until the Supreme Command admitted defeat at the end of September 1918.

Conversion into a "realistic republican." The defeat, the collapse of the monarchy on November 9, 1918, and Emperor William II's flight to Holland were cruel blows to Stresemann. Nonetheless, he quickly accommodated himself to the realities of a republican Germany. When, however, the newly formed left-liberal German Democratic Party, led by Friedrich Naumann and the renowned sociologist Max Weber, refused to admit him to its higher councils, Stresemann founded his own party, the German People's Party (DVP). A right-liberal grouping of educated and propertied elements, it sought to rally the right-wing supporters of the former National Liberal Party. Stresemann, fundamentally a monarchist and an opponent of the Weimar Republic, assumed an ambiguous "wait-and-see" attitude during the rightist Kapp Putsch of March 1920. When the putsch was suppressed, however, he prepared to cooperate politically with the republic. He tried to persuade the democratic parties that the DVP was qualified to participate in a coalition and pressed for a position in the government. For the time being, however, he was still counted among the "national opposition" to the Weimar coalition—the Social Democrats, the German Democratic Party, and the Centre Party.

A member of the German National Constituent Assembly in Weimar in 1919–20, Stresemann was among the opponents of the new German constitution. He also opposed the Treaty of Versailles and was to devote his political life to its revision. From 1920 until his death Stresemann was a Reichstag deputy and chairman of the DVP, and in August 1923 he became chancellor of the *Reich* at the head of a government of the "Great Coalition," comprising representatives of the Social Democrats, of the Centre, and of the German Democrats, as well as of his People's Party.

As chancellor of Germany

As chancellor from August 13 to November 23, 1923, during the national crisis over the Allied occupation of the Ruhr, and as foreign minister from August 1923 to his death in 1929, Stresemann exercised decisive influence over the fate of the Weimar Republic, and he became a statesman of European stature. His first decision as chancellor was to abandon the policy of passive resistance in the Ruhr, which in January 1923 had been occupied by French and Belgian troops to enforce payment of German war reparations. This policy had accelerated inflation and was precipitating Germany's financial collapse.

On the domestic scene, he sought to steer his way among opposing domestic forces. While proceeding harshly against Communist-influenced state governments in Thuringia and Saxony, he displayed a lenient attitude toward revolutionary attempts of the radical right, such as Adolf Hitler's putsch of November 9, 1923, in Munich. At the height of the internal political crisis of November 1923 there was danger that the occupied territory west of the Rhine (occupied by the Allies) might withdraw from the *Reich*. Only the stabilization of the currency in the middle of November—the last significant achievement of Stresemann's government—restored domestic order and created the basis for economic recovery.

Years as foreign minister. Overthrown by a vote of no confidence, the Cabinet resigned in November 1923. Stresemann took over the post of foreign minister in the new government and held it, unchallenged until his death, in coalition governments of varying composition under three chancellors ranging from the left to centre. His policy was aimed at securing a reconciliation with the victorious Western powers, especially France, for Germany had already renewed ties with Russia through the Treaty of Rapallo in 1922. By meeting the reparation payments, for the reduction of which he fought as stubbornly as he did for removal of French troops from the territory west of the Rhine, he hoped to gain a favourable position for his negotiations with the victorious Allies. His enduring aim was to obtain equal rights for Germany and to restore it to its former position among the nations of Europe.

Principally, however, this meant a revision of Germany's eastern border of 1919, which would require Poland to return Danzig, the Polish Corridor, and Upper Silesia, as well as the annexation of Austria. Realistically appraising Germany's central position in Europe and exploiting Anglo-French and Anglo-Soviet tensions, Stresemann tried to achieve his goals through negotiation, but his seesaw policy between East and West was strongly criticized by many contemporary critics. Yet Stresemann retained his optimistic outlook, which, if anything, was strengthened by his poor health. This optimism, however, which he often carried to extremes, especially in his speeches, frequently led him to underestimate opposition both at home and abroad.

Stresemann's successes in dealing with the Allied powers during those years can be marked out in stages. In 1924 the U.S.-proposed Dawes Plan was signed, providing for reduction in payment of reparations and stabilization of German finances. It was followed by the Pact of Locarno in 1925, which included acceptance of the new Franco-German border, agreements to arbitrate disputes with other nations, and immunity from new sanctions by the victors of World War I. In 1926 the first Rhineland zone was evacuated by the Allies, Germany was admitted to the League of Nations, and the Berlin Treaty with the Soviet Union, an agreement providing for mutual neutrality, was signed. In 1928 the Kellogg–Briand Pact outlawing war was signed by Germany. Stresemann did not live to see the complete evacuation of French troops from the Rhineland and the completion of the new settlement reducing German reparations through the Young Plan (also a U.S. proposal) in 1929, although he had conducted the negotiations when already marked by death. He had reached the zenith of his career in 1926; at the end of that year, together with the French foreign minister Aristide Briand, he was awarded the Nobel Peace Prize for his policy of reconciliation and negotiation.

Rapprochement with Allies

Any summary of Stresemann's diplomatic successes should not obscure the fact that he had to devote an extraordinary amount of effort to combatting strong domestic opposition that arose, above all, from his own party. Stresemann, who took the importance of the press into consideration, used publicity and propaganda to promote his policy but, by making premature statements, often aroused political hopes that could not be realized. After his spectacular secret meeting in 1926 with Aristide Briand, which gave rise to exaggerated hopes, Franco-German rapprochement came to a standstill. In the last two years of his life, which were marked by illness, Stresemann became increasingly dissatisfied at his failure to further his foreign policy, especially after his party dwindled, and large sections of it went over to the extreme right. He himself contemplated formation of a new party of the liberal centre. The domestic struggle in particular weakened his already precarious health, and he died unexpectedly, after two strokes, on October 3, 1929, at the age of 51.

Assessment. By virtue of his six years of service as foreign minister and the esteem he enjoyed, particularly abroad, Stresemann made an essential contribution to securing the Weimar Republic's stability and survival for a few years. Because of domestic undercurrents and opposition, he succeeded at the cost of extraordinary personal effort. He was the target of malicious defamation by the "national" right opposition in its struggle against democracy and the Weimar Republic, and his nationalist past offered convenient points of attack from the left.

On his death, the republic, which honoured him with a state funeral, lost one of its few statesmen. In his personal as well as his political development he embodied the uncertainty of the period of transition from monarchy to republic. Yet he was unable to integrate his own party, over which he jealously maintained leadership, into the Weimar state and thus failed, as he wrote in March 1929, to form the bridge "between the old and the new Germany."

As an advocate of a "policy of national realism," as opposed to a "pacifist policy of resignation," he was by no means a champion of European unification. He supported its objectives, however, since he could thus more easily obtain the urgently sought revision of the Treaty of Versailles.

Stresemann's image is still controversial. Historically,

Continuing controversy over Stresemann

he was first pronounced a hero after 1945, when he was viewed as a champion of a united Europe. This view was succeeded in the 1950s by an increasingly critical evaluation, especially during the disclosure of his voluminous literary estate, which was at first exclusively at the disposal of U.S. historians. He was then portrayed as a flexible and opportunistic politician of nationalistic sympathies who shrewdly adjusted his aims to meet the needs of the time; and it was said that Stresemann had not become a democrat out of conviction but rather that he had raised "finessing," which he mentioned in 1925 in a well-known letter to the former crown prince, to the level of a principle. His volatile character and sentimental attachment to uniforms and tradition were also emphasized. The Communists, meanwhile, regarded him as a representative of monopoly capitalism and a forerunner of Hitler.

One of his more recent biographers has convincingly characterized Stresemann as a "pragmatic conservative" who remained flexible in his choice of political means while pursuing his national aims of restoration of German wealth and power and the continuation of German traditional social and economic order. Others have emphasized the European aspect of the German "patriot" Stresemann, a viewpoint from which German historical research has in the meantime departed. Yet, his political changeability notwithstanding, Stresemann is counted among the few statesmen of his time, as well as among the "great Germans."

BIBLIOGRAPHY. FELIX E. HIRSCH, *Gustav Stresemann: Patriot und Europäer* (1964), is the best biography to date, concise and interesting. See also ANNELISE THIMME, *Gustav Stresemann: Eine politische Biographie zur Geschichte der Weimarer Republik* (1957); DONALD WARREN, JR., *The Red Kingdom of Saxony: Lobbying Ground for Gustav Stresemann, 1901–1909* (1964), basic for Stresemann's rise as a corporate lawyer; MARVIN L. EDWARDS, *Stresemann and the Greater Germany, 1914–1918* (1963), pertains to Stresemann's annexationist objectives in World War I; ROLAND THIMME, *Stresemann und die Deutsche Volkspartei, 1923–1925* (1961), an adequate treatment of an important subject; FELIX E. HIRSCH, "Stresemann in Historical Perspective," *Review of Politics*, 15:360–377 (1953), an important evaluation by a contemporary; HENRY L. BRETTON, *Stresemann and the Revision of Versailles* (1953), the beginning of research on Stresemann; HANS W. GATZKE, *Stresemann and the Rearmament of Germany* (1954), an excellent treatment of a single topic; HENRY A. TURNER, *Stresemann and the Politics of the Weimar Republic* (1963), the best general treatment; and MARTIN WALSDORFF, *Westorientierung und Ostpolitik: Stresemanns Russlandpolitik in der Locarno-Ära* (1971), a carefully considered evaluation with a detailed bibliography, and *Gustav-Stresemann-Bibliographie* (1972).

(R.Mo.)

Strigiformes

Owls constitute a homogeneous order (Strigiformes) of primarily nocturnal birds of prey, nearly worldwide in distribution. Their secretive habits, quiet flight, and weird calls have made them the objects of superstition and even fear in many parts of the world. The same characteristics have made the scientific study of them more difficult, with the result that the ecology, behaviour, and taxonomy of many species remain poorly understood.

This article is divided into the following sections:

General features
Natural history
  Ecology
  Behaviour
  Reproduction and development
  Locomotion
Form and function
Paleontology and classification
  Fossil history
  Distinguishing taxonomic features
  Annotated classification
  Critical appraisal

GENERAL FEATURES

The size range of owls is about the same as that of their diurnal counterparts, the hawks (exclusive of the eagles and vultures), with lengths of around 13–70 centimetres and wingspreads around 30–200 centimetres. There are more owl species at the lower end of the size range. Owls apparently feed exclusively on animals. The most common prey is rodents; the smaller species, however, eat insects. All owls have the same general appearance: large head; hooked beak; large, forward-looking eyes; short tail; rather long, rounded wings; and large feet with sharp talons. Diversity occurs in size, in the presence or absence of decorative "ear" tufts, and in the extent of the facial ruff.

Owls are found on all continents except Antarctica and on most oceanic islands. Some, such as the barn owl (*Tyto alba*) and the short-eared owl (*Asio flammeus*), are among the most widely distributed of bird species; others, such as the Palau owl (*Pyrroglaux podargina*) and the Seychelles owl (*Otus insularis*), are endemic island species with small populations. Owls often attain higher population densities than hawks and have survived better in areas of human activity. Their nocturnal habits and inconspicuous daytime behaviour provide them some protection from shooting. The greatest densities are attained by small, territorial, insectivorous species, with pairs spaced about 200 metres apart in suitable woodland.

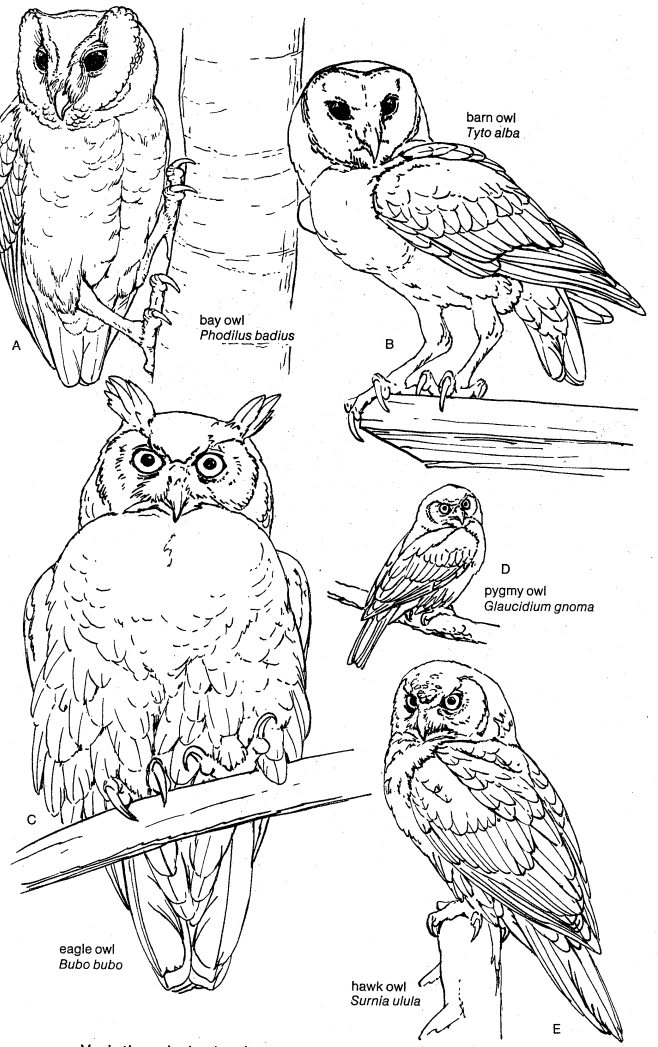
NATURAL HISTORY

Ecology. Owls utilize virtually all habitats, from grassland and tundra to dense woodland and rain forest. The distribution and density of most species apparently are limited by the availability of suitable nesting sites, rather than by the number of potential prey animals. In general the type of prey taken is governed by the size of the owl and by the relative abundance of potential prey species. Owls that hunt over grassland, such as the barn owl and short-eared owl, hunt by sustained flight, dropping into the grass to catch rodents. Many woodland owls secure prey by dropping from perches at the edges of forest openings. The Southeast Asian hawk-owl (*Ninox scutulata*) sallies from a perch to take flying insects. The whiskered owl (*Otus trichopsis*) takes flying insects in foliage. Owls of two genera (*Ketupa* and *Scotopelia*) are adapted for taking fish. Specialized forms of feeding behaviour are sometimes observed. The elf owl (*Micrathene whitneyi*) has been seen hovering before blossoms, scaring insects into flight with its wings and then catching them with its beak. A bay owl (*Phodilus badius*) has been observed stationed within a bat cave, catching bats as they issued forth at dusk. Although each species normally feeds on the type of prey for which it is adapted, a variety of owls may depend on a single prey animal when it becomes exceptionally abundant. Prey is generally swallowed whole, and indigestible material, such as feathers, fur, and bones, are regurgitated in the form of a compact pellet.

Behaviour. Sound communication is important to owls, especially in mating and territorial defense. With their camouflage, immobile daytime stance, and silent flight, owls may be as difficult for each other to see as for natural enemies and human observers. Their sounds used include snaps of the bill, claps of the wings in flight, and a variety of vocalizations, with pitch, timbre, and rhythm unique to each species and with pitch, at least, differing between sexes (the female higher). Although less melodious than the calls of some birds, the vocalizations of many owls are "songs" in the biological sense and may even be musical to the human ear. The song varies from deep hoots in some large species to chirps, whistles, or warblings in many small owls. When the nest is threatened, nestling burrowing owls (*Speotyto cunicularia*) emit a buzzing call that resembles the warning buzz of the rattlesnake, a frequent inhabitant of rodent burrows.

Vocalizations

In the North American screech owl, a duet that seems to reinforce the pair bond starts with a special song by the male, eventually answered in kind by the female, often from a distance. After ten to 15 minutes of antiphonal (answering) singing, during which the two approach each other, the pair switches to a duet with a second song pattern, during which they meet. In the early spring this may be followed by precopulatory calls and posturing,



Variations in body plan among owls.
Drawing by (A) R. Keane and (B,C,D,E) D. Malick

then mounting. Other calls of the screech owl include: a note uttered by the female to stimulate the young to reveal their location after they have left the nest; a food soliciting call by the young; after the breeding season, barking calls, accompanied by bill-snapping, which indicate that the young are being ejected from the territory, or the adjustment of territorial boundaries.

In many smaller species that do not normally sing duets, the male may sing all night from a single perch. The nocturnal routine of most owls involves peaks of activity at dusk and dawn. The owl leaves its secluded roost about dusk and moves to a perch overlooking the hunting area. There is a brief period of song, followed by about half an hour of foraging, then a longer period of song. Most of the darker hours of the night are spent inactively, with a period of alternating singing and hunting just before dawn.

Posturing by singing owls indicates that they communicate by sight as well as by sound. The male horned owl bows deeply with each song and raises the tail over the back. The wood owls (*Strix* species) engage in bowing, bobbing, and dancing, especially when courting. A defense display given by most large owls when threatened or when defending the nest involves increasing apparent body size by half spreading the wings and rotating them forward. The body feathers are raised and the fearsome appearance enhanced by snapping the bill and rocking the body from side to side. When seeking to avoid attention at its daytime roost (especially when being attacked by small birds), the owl compresses the plumage, elevates the ear tufts, and half closes the eyes. Combined with the bark-like colour pattern of the plumage, this behaviour provides the appearance of a broken branch.

Reproduction and development. Most owls nest in natural cavities in trees or cliffs or in woodpecker holes. The barn owls and the Eurasian little owls (*Athene* species) frequently use cavities in buildings. Some of the larger owls utilize old hawk or crow nests. Grassland and tundra owls nest on the ground, sometimes on an elevated hummock, and the burrowing owl digs a nest chamber in a rodent burrow. Frequently the nest cavity provides a daytime roost for one or both of the pair during the non-breeding months. Most owls add no nesting material to the site, but the fur and feathers of accumulated prey remains and regurgitated pellets may provide some cushion for the eggs. When an open nest is used, leaves, grass, or other soft material may be added as a lining. The great gray owl (*Strix nebulosa*) occasionally constructs its own platform nest in a tree. In desert areas the smaller owls rely primarily on holes made by woodpeckers in large cacti. Intense competition has been observed among nesting birds, including owls, for occupancy of a limited number of nest sites. The invasion of the saguaro desert by the European starling (*Sturnus vulgaris*) has had a serious effect on small owls and other saguaro birds. The aggressive and abundant starlings occupy cavities before the other species have returned from wintering grounds and successfully defend the holes against native species. Nesting sites

Egg laying is timed so that the young become independent of their parents at a time when populations of prey species are greatest. At northern latitudes, many owls nest a month or two earlier in the spring than hawks of the same regions. An incubating owl is frequently covered by several inches of snow. Barn owls have been found nesting in every month of the year, even at the northern edge of their range, but the peak of nesting is in the spring. Owls lay more eggs than most diurnal raptors, with clutches of up to twelve (in the snowy owl, *Nyctea scandiaca*). In years of lemming abundance, snowy owls attain higher nesting densities, nest earlier, lay more eggs, and have higher fledging rates than when lemmings are scarce. The eggs of owls are more nearly spherical than those of any other bird group, the long diameter averaging only about 1.2 times the short diameter. They are laid at one or, more often, two day intervals, and incubation begins when the first is laid. The result is that the oldest and youngest nestlings of a large brood may be two or three weeks apart. If the prey populations are inadequate for the adult owls to support the entire large brood, the younger nestlings starve, while the more aggressive older ones are able to maintain normal growth rates and are strong at fledging.

While in the nest, young owls grow two successive coats of white natal down. In the smaller species, the down is replaced by the immature juvenal plumage of softer, lacier texture than that of the adult. At this age, about two-thirds of the way to fledging, the young owls may leave the nest and spend the day several metres from it. In the screech owl, the juvenal plumage has fine barring, unlike the streaking of the adult. The first flight feathers, which appear about the same time, are like those of the adult but are more pointed. In the fall, only a few weeks after its acquisition, this plumage is replaced in a complete molt that introduces a plumage identical to that of the adult. In the larger owls that have been extensively studied (several species of *Bubo*), the juvenile plumage resembles the adult except that in *B. nipalensis* and *B. sumatranus* it is mostly white, with a different pattern of black marking from the adult. Plumage succession

Locomotion. Most owls use their feet only for perching and grasping prey. The burrowing owls, being terrestrial, can run rapidly over the ground and rarely perch in vegetation. The typical perch of other owls is a horizontal branch in a tree. The bay owl habitually perches on the vertical trunk of a tree, with the legs reaching sideways, one foot above the other. The body is held in a vertical position as though the owl were perched on a horizontal limb.

The flight of owls is a steady flapping on a straight path, ending in a short upward glide to the perch. Hunting, usually done from a perch, rarely involves extensive flight, requiring only a short burst of speed to surprise the

prey on the ground. Forest owls rarely fly above the canopy of the foliage, but the Southeast Asian hawk-owl has been observed high in the air, flapping—or, on one occasion, soaring—in circles through swarms of bats, apparently without catching any. A few grassland and tundra owls hunt in sustained flight and have proportionately more wing area than their forest-dwelling relatives. The short-eared owl flaps slowly, the large area of the wings causing the light body to bob up and down, and sails for a brief period of time with the wings held in a high V over the back.

FORM AND FUNCTION

All owls share the same general body plan. The wings are long and rounded, the tail short. The legs and toes are of medium length and exceptionally strong for the size of the bird. Each toe is provided with a needle-sharp, curved talon. The outer toe is reversed in perching and is normally directed outward or backward in taking prey, providing the maximum possible toe spread.

The head is broad, to accommodate the exceptionally large eyes. The owl eye is elongated forward and encased in a tube made up of joined bony elements. The eye is virtually immovable, rigidly encased in the bony socket. Remarkable flexibility of the neck compensates for the immobility of the eyes; an owl can turn its head more than 180° in either direction and can thus look directly backward. The vision is binocular, and depth perception is often enhanced by moving the head away from the central plane. Owls that are mostly active at night have only rods in the retina, resulting in an absence of colour vision but a great increase in visual acuity and light sensitivity. Contrary to popular opinion, owls are not blind in strong light. The pupils, which operate independently, can be greatly reduced, protecting the sensitive retina and providing better daytime vision than that found in man.

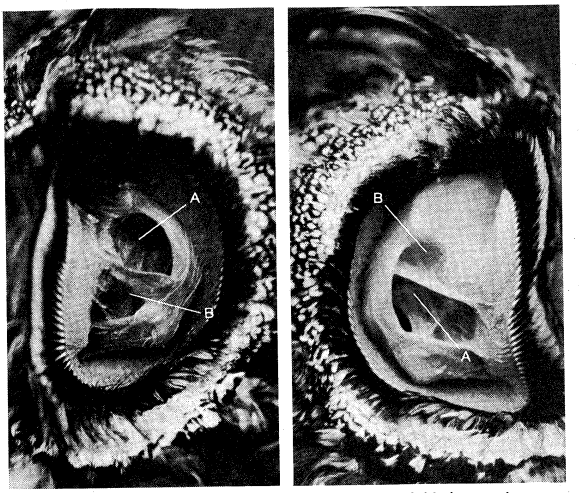
Hearing

The ear is large and is surrounded by a ruff of papery feathers that serve to concentrate the sound. The feathers covering the ear opening are lacy and permeable to sound. A movable flap (operculum) on the anterior margin of the opening may function as a baffle to focus sounds. Some owls can locate and capture prey in total darkness, relying on their ability to localize the rustle of a mouse in leaves and to fly to that spot. In many owls the relative position of the ear opening is asymmetrical, being above a so-called blind cavity on one side of the head and below it on the other. The asymmetry is thought to be related to the sensitivity of each ear to sounds of various frequencies and provides the owl with the ability to localize sound sources in two planes simultaneously.

Many owls have erectile tufts of feathers ("ears" or "horns") above the eyes. In some cases the tufts give the owls a feline appearance and might frighten potential predators, but it is more likely that the erect "ears" serve to break the rounded outline of the head, adding to the concealment gained from colour and pattern.

The plumage of owls is soft, dense, and loose. A thick layer of down provides northern owls with insulation against cold. The upper surfaces of the flight feathers of most species are provided with a nap that makes the flight perfectly noiseless, allowing the owl to hear prey without interference of flight sound.

Owls vary in coloration from white through many shades of tan, gray, brown or rufous to deep brown. A few are solidly coloured, but most are cryptically patterned, with streaks, bars, or spots, often resulting in the bird being almost invisible against bark. The concealing bark pattern is well exemplified in the small owls of the genus *Otus*. The soft brown, rufous, buff, or gray ground colour of each breast feather is adorned by a blackish bar, shaft streak, or combination of both, sometimes outlined in white or rufous. In some widespread species, such as the Eurasian scops owl (*O. scops*) and the screech owl, geographic variation is so great that the more divergent races are more different than some full species. In the far north there is only a faint pattern on a whitish background; in humid, temperate forests a bold pattern on a sooty background; in desert areas a medium to fine pattern on pale gray; in arid tropics a fine pattern on rufous; and in humid tropics a coarse pattern on fulvous. Size variation is also present, with northern birds weighing about twice as much as those of southern races. The horned owl (*Bubo virginianus*) exhibits similar variation.



Asymmetry of the ear canal in the long-eared owl (Asio otus). In the left ear (left), the opening of the ear canal (A) is above the transverse fold and the blind cavity (B), below. In the right ear, the situation is reversed.
Heinz Meng

Colour dimorphism

Screech, scops, and whiskered owls may be either gray or rufous, apparently controlled by a single gene. Such dimorphism in ground colour is found in only certain populations of each species: the southern in the whiskered owl and the eastern in the North American screech owl and the Eurasian scops owl. In each case, interbreeding between the uniformly coloured monomorphic and sexually distinct dimorphic populations is limited. J.T. Marshall, Jr., regards the dimorphic strains as "incipient species." The mechanism by which natural selection maintains colour dimorphism is poorly understood. Moreover, little is known about the type and degree of predation on small owls. The red phase of the screech owl may have survival value in the predominantly deciduous forest of eastern North America, blending with the foliage of summer and autumn, in which reds and browns are strongly represented.

PALEONTOLOGY AND CLASSIFICATION

Fossil history. The earliest recognizable fossil strigiform is *Protostrix* from the Eocene of Wyoming. A number of genera from the later Tertiary have been assigned to Recent strigiform families. Some early owls reached far greater size than their modern descendents. A giant barn owl, about twice the size of the modern *Tyto alba*, inhabited Puerto Rico in the Pleistocene. Another large owl, *Ornimegalonyx oteroi* from the Pleistocene of Cuba, apparently was flightless. Both owls must have exceeded the modern eagle owls in size.

Distinguishing taxonomic features. The owls form a homogeneous group readily distinguished from all other orders by their general body plan, soft plumage, and osteological peculiarities. To determine relationships within the Strigiformes, taxonomists utilize the osteology of the skull and the sternum, relative specializations of the eye and ear, and the development of the facial disk. Within some genera the taxonomy is very complex, relying on general proportions, behaviour, voice, and ectoparasites (Mallophaga or feather lice).

Annotated classification.

Order Strigiformes. Nocturnal predatory birds with hooked beaks; nostril marginal to fleshy cere; broad skull, facial disk, with forwardly directed eyes; reversible outer toe; talons for securing prey. One fossil and 3 extant families.

Family Protostrigidae. Fossil. Known only from lower Eocene of Wyoming.

Family Tytonidae (barn owls and grass owls). Lower Miocene to present. Medium-sized owls (30–54 cm long) with heart-shaped facial disk completely encircling face; bill relatively long and slender; legs rather long, middle claw

with comb. One pair of sternal notches, clavicles separate and not fused to sternum. One genus (*Tyto*), about 9 species. Worldwide, except in polar and subpolar regions and on some oceanic islands.

Family Phodilidae (bay owls). Miocene and Recent. Like a small *Tyto*, but clavicles fused to sternum, which has 2 pairs of notches. Facial disk differs from *Tyto* in forming 2 movable lappets that can be focussed at sound, like mammalian ears. Disk interrupted at forehead. Wings short, broad. Legs and toes shorter and thicker than in *Tyto*. Voice, syrinx, behaviour, and ecology differ from both Tytonidae and Strigidae. One Recent genus (*Phodilus*), with 1 species widespread in Southeast Asia, and another (known only from a single specimen) in central Africa.

Family Strigidae (typical owls). Upper Eocene to Recent. Clavicles fused to sternum, which has 2 pairs of notches. Facial disk consists of a semicircle on each side. Middle claw without comb. About 22 Recent genera, 125 species. Length about 13–70 cm. Worldwide, except Antarctica and some oceanic islands.

Critical appraisal. Most recent taxonomists have divided the Recent Strigiformes into two families, Tytonidae and Strigidae. The bay owl (*Phodilus*) has been placed in the Tytonidae by some authorities and in the Strigidae by others. F.E. Beddard, in 1898, suggested that the distinctive characteristics of the bay owl justified separation in a monotypic family, Phodilidae, a view recently endorsed by J.T. Marshall, Jr., after observations in the field.

Within the Strigidae, most current taxonomic problems are concerned with the placement of certain genera within the family and with the specific status of some populations of complex genera, such as *Otus*. Some of these problems await the development of new information on species isolating mechanisms, especially vocal behaviour.

BIBLIOGRAPHY. O.L. AUSTIN, *Birds of the World*, ed. by H.S. ZIM (1961), presents an excellent general account of owls, with fine colour illustrations that include some rarely seen species. A.C. BENT, "Life Histories of North American Birds of Prey, Part II," *Bull. U.S. Natn. Mus. 170* (1938; reissued in 1961 as vol. 2 of series of same title), contains complete accounts of the natural history of North American owls. J.J. and F.C. CRAIGHEAD, *Hawks, Owls and Wildlife* (1956, reissued in pa. 1969), is a classic in the ecology of raptorial birds. M.L. GROSSMAN and J. HAMLET, *Birds of Prey of the World* (1964), provides range maps for each species, good photographs, and a reading list by geographical region and by genus. G.P. DEMENTIEV *et al.*, *Birds of the Soviet Union*, vol. 1 (Eng. trans. 1966; orig. pub. in Russian, 1951), contains an extensive and well illustrated section on Eurasian owls. E.C.S. BAKER, *Fauna of British India (Birds)*, vol. 4 (1927), includes an excellent discussion of the anatomical traits used in the classification of owls. Of a more specialized nature: R.S. PAYNE and W.H. DRURY, JR., "Marksman of the Darkness," *Nat. Hist.*, vol. 67, pp. 316–323 (June 1958), describes the discovery of the hearing ability of the barn owl. D.J. LIGON, "The Biology of the Elf Owl, *Micrathene Whitneyi*," *Misc. Publs. Mus. Zool. Univ. Mich. No. 136* (Feb. 1968), is a detailed study of the life history of the smallest owl, compared with other owls. A. ROSS, "Ecological Aspects of the Food Habits of Insectivorous Screech-owls," *Proc. West. Fdn. Vertebr. Zool.* vol. 1, no. 6, pp. 301–344 (1969), is a definitive correlation between body size and prey size, one of the few such papers in which stomach contents are correctly identified. L.W. WALKER, "Nocturnal Observations of Elf Owls," *Condor*, vol. 45, no. 5, pp. 165–167 (Sept.–Oct. 1943), is a model of excellence in photography and concise scientific writing. B.E. SMYTHIES, *The Birds of Burma*, 2nd ed. rev. (1953), and *The Birds of Borneo* (1960), provide the best and most reliable information on the habits and calls of Asian owls.

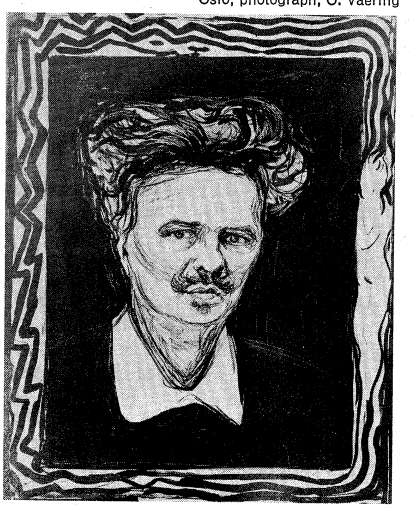
(J.T.M.)

Strindberg, August

Although he led a tortured, half-mad existence, August Strindberg, the greatest Swedish playwright, wrote a series of novels, short stories, and plays that marked the beginning of modern Swedish literature and exercised a profound influence on European and American drama.

He was born in Stockholm on January 22, 1849, the son of Carl Oskar Strindberg, a bankrupt aristocrat who worked as a steamship agent, and of a former waitress. His childhood was marred by emotional insecurity, poverty, his grandmother's religious fanaticism, and neglect, as he relates in his remarkable autobiography *The Son of a Servant* (1886). He studied intermittently at the University of Uppsala, preparing in turn for the ministry and a career in medicine but never taking a degree. To earn his living, he worked as a free-lance journalist in Stockholm, as well as at other jobs that he almost invariably lost. Meanwhile he struggled to complete his first important work, the historical drama *Mäster Olof* (published in 1872), on the theme of the Swedish Reformation, influenced by Shakespeare and by Henrik Ibsen's *Brand*. The Royal Theatre's rejection of *Mäster Olof* deepened his pessimism and sharpened his contempt for official institutions and traditions. For several years he continued revising the play—later recognized as the first modern Swedish drama—thus delaying his development as a dramatist of contemporary problems.

By courtesy of the Munch-Museet, Oslo; photograph, O. Vaering



Strindberg, lithograph by Edvard Munch, 1896. In the Munch-Museet, Oslo.

In 1874 he became a librarian at the Royal Library, and in 1875 he met the Finno-Swedish Siri von Essen, then the unhappy wife of an officer of the guards; two years later they married. Their intense but ultimately disastrous relationship ended in divorce in 1891, when Strindberg, to his great grief, lost the custody of their four children. At first, however, marriage stimulated his writing, and in 1879 he published his first novel, *The Red Room*, a satirical account of abuses and frauds in Stockholm society: this was something new in Swedish fiction and made its author nationally famous.

First plays of social criticism

He also wrote more plays, of which *Lucky Peter's Travels* (1881) contains the most biting social criticism. In 1883, the year after he published *Det nya riket* ("The New Kingdom"), a withering satire on contemporary Sweden, Strindberg left Stockholm with his family and for six years moved restlessly about the Continent. Although he was then approaching a state of complete mental breakdown, he produced a great number of plays, novels, and stories. The publication in 1884 of the first volume of his collected stories, *Married*, led to a prosecution for blasphemy. He was acquitted, but the case affected his mind, and he imagined himself persecuted, even by Siri.

He returned to drama with new intensity, and the conflict between the sexes inspired some of the outstanding works written at this time, such as *The Father*, *Miss Julie*, and *The Creditors*. All of these were written in total revolt against contemporary social conventions. In these bold and concentrated works, he combined the techniques of dramatic naturalism—including unaffected dialogue, stark rather than luxurious scenery, and the use of stage props as symbols—with his own conception of psychology, thereby inaugurating a new movement in European drama. *The People of Hemsö*, a vigorous novel about the Stockholm skerries (rocky islands), always one of Strindberg's happiest sources of inspiration, was also produced during this intensively creative phase.

The years after his return to Sweden in 1889 were lonely and unhappy. Even though revered as a famous writer who had become the voice of modern Sweden, he was by now an alcoholic unable to find steady employment. In 1892 he went abroad again, to Berlin. His second marriage, to a young Austrian journalist, Frida Uhl, followed in 1893; they finally parted in Paris in 1895.

Period of religious crisis

A period of literary sterility, emotional and physical stress, and considerable mental instability culminated in a kind of religious conversion, the crisis that he described in *Inferno*. During these years Strindberg devoted considerable time to experiments in alchemy and to the study of theosophy.

His new faith, coloured by mysticism, re-created him as a writer. The immediate result was the drama in three parts, *To Damascus*, in which he depicts himself as "the Stranger," a wanderer seeking spiritual peace and finding it with another character, "the Lady," who resembles both Frida Uhl and Siri von Essen.

By this time Strindberg had again returned to Sweden, settling first in Lund and then, in 1899, in Stockholm, where he lived until his death. The summers he often spent among his beloved skerries. His view that life is ruled by the "Powers," punitive but righteous, was reflected in a series of historical plays that he began in 1889. Of these, *Gustav Vasa* is the best, masterly in its firmness of construction, characterization, and its vigorous dialogue. In 1901 he married the young Norwegian actress Harriet Bosse; in 1904 they parted, and again Strindberg lost the child, his fifth.

Yet his last marriage, this "spring in winter," as he called it, inspired, among other works, the plays *The Dance of Death* and *A Dream Play*, as well as the charming autobiography *Ensam* ("Alone") and some lyrical poems. Renewed bitterness after his parting from his last wife provoked the grotesquely satirical novel *Svarta Fanor* ("Black Banners," 1907), which attacked the vices and follies of Stockholm's literary coteries, as Strindberg saw them. *Kammarspel* ("Chamber Plays"), written for the little Intima Theatre, which Strindberg ran for a time with a young producer, August Falck, embody further developments of his dramatic technique: of these, *The Ghost Sonata* is the most fantastic, anticipating much in later European drama. His last play, *The Great Highway*, a symbolic presentation of his own life, appeared in 1909.

To the end, Strindberg debated current social and political ideas (returning to the radical views of his youth) in polemical articles, while his philosophy was expounded in the aphoristic *Zones of the Spirit* (1907–12). He died of cancer on May 14, 1912, in Stockholm, ignored in death, as in life, by the Swedish Academy but mourned by his countrymen as their greatest writer. On Swedish life and letters he has exercised a lasting influence and is admired for his originality, his extraordinary vitality, and his powerful imagination, which enabled him to transform autobiographic material into dramatic dialogue of exceptional brilliance.

Assessment

The pregnant, colloquial style of Strindberg's early novels and, especially, of his short stories, brought about a long-overdue regeneration of Swedish prose style, and *The Son of a Servant* gave perhaps the strongest impulse since Jean-Jacques Rousseau's *Confessions* to the publication of discreditable self-revelations. His greatest influence, however, was exerted in the theatre, through his critical writings (such as the introduction to *Miss Julie*), his plays, and the production devices that their staging dictated. The continuous, brutal action and phonographic dialogue of *Miss Julie* and other plays written between 1887 and 1893 reached the *ne plus ultra* of naturalistic drama. With the later phantasmagoric plays, such as *To Damascus, A Dream Play*, and *The Ghost Sonata*, Strindberg led that section of the revolt against stage realism that issued in the Expressionist drama, which was developed mainly in Germany after 1912 and which influenced such modern playwrights as Sean O'Casey, Elmer Rice, Eugene O'Neill, Luigi Pirandello, and Pär Lagerkvist.

MAJOR WORKS

PLAYS: *Den fredlöse* (first performed 1871, published 1881; *The Outlaw*, 1912); *Mäster Olof* (poetic version published 1878, performed 1890; *Master Olof*, 1915); *Lycko-Pers resa* (published 1881, performed 1883; *Lucky Peter's Travels*, 1930); *Fadren* (published and performed 1887; *The Father*, 1899); *Kamraterna* (published 1888, performed 1905; *Comrades*, 1912); *Fröken Julie* (published 1888, performed 1889; *Countess Julie*, 1912; *Miss Julie*, 1918); *Fordringsägare* (first published in Danish trans., 1888; *Creditörer*, published and performed 1890; *Creditors*, 1914); *Paria* (performed 1889, published 1890; *Pariah*, 1913); *Till Damascus* (parts 1 and 2 published 1898, performed 1900; part 3 published 1904, performed 1916; *To Damascus*, 1913); *Folkungasagan* (published 1899, performed 1901; *The Saga of the Folkungs*, 1931); *Gustav Vasa* and *Erik XIV* (published and performed 1899; Eng. trans. 1931); *Påsk* (published and performed 1901; *Easter*, 1912 and 1929); *Dödsdansen* (published 1901, performed 1905; *The Dance of Death*, 1912); *Kronbruden* (published 1902, performed 1906); *Drömspelet* (published 1902, performed 1907; *A Dream Play*, 1929); *Brända tomten* (performed and published 1907; *After the Fire*, 1913); *Spöksonaten* (published 1907, performed 1908; *The Spook* [or *Ghost*] *Sonata*, 1916); *Stora landsvägen* (published 1909, performed 1910; *The Great Highway*, 1954).

NOVELS: *Röda rummet* (1879; *The Red Room*, 1913); *Hemsöborna* (1887; *The People of Hemsö*, 1959); *I havsbandet* (1890; *By the Open Sea*, 1913).

SHORT STORIES: *Giftas*, 2 vol. (1884–85; *Married*, 1913); *Fagervik och skamsund* (1902; *Fair Haven and Foul Strand*, 1913); *Sagor* (1903; *Tales*, 1930).

AUTOBIOGRAPHY: *Le Plaidoyer d'un fou* (first published in French, 1888; enlarged, 1893; *The Confessions of a Fool*, 1912); *Inferno* (1897; Eng. trans., 1912); *Ensam* (1903).

TRANSLATIONS: There is no English-language edition of the complete works of Strindberg, which fill 55 volumes in Swedish. Individual works are, however, readily available. Recommended modern translations include *Twelve Plays* by Elizabeth Sprigge (1963); and *The Plays of Strindberg* by Michael Meyer, which began publication in 1964.

BIBLIOGRAPHY. There is no definitive biography or full up-to-date bibliography.

Editions: *Samlade Skrifter*, ed. by JOHN LANDQUIST, 55 vol. (1911–21); several collections of letters (*Brev*, ed. by TORSTEN EKLUND, 1948–).

Criticism and biography: ERIK HEDEN, *Strindberg en ledtråd vid studiet av hans verk* (1921); MARTIN LAMM, *Strindbergs dramer*, 2 vol. (1924–26), *August Strindberg*, 2 vol. (1940–42; Eng. trans., 1971); TORSTEN EKLUND, *Tjänstekvinnans son* (1948); GUNNAR OLLEN, *Strindbergs dramatik* (1948); GUNNAR BRANDELL, *Strindbergs Infernokris* (1950); S. AHLSTROM and TORSTEN EKLUND (eds.), *August Strindberg*, 2 vol. (1959–61); JOAN BULMAN, *Strindberg and Shakespeare* (1933, reprinted 1971); ELIZABETH SPRIGGE, *The Strange Life of August Strindberg* (1949); MAURICE GRAVIER, *Strindberg et le théâtre moderne* (1949); B.M. MORTENSEN and B.W. DOWNS, *Strindberg: An Introduction to His Life and Works* (1949); F.L. LUCAS, *The Drama of Ibsen and Strindberg* (1962), inimical to Strindberg; ELIE POULENARD, *August Strindberg, romancier et nouvelliste* (1962); WALTER JOHNSON, *Strindberg and the Historical Drama* (1963); C.R. SMEDMARK (ed.), *Essays on Strindberg* (1966); E.O. JOHANNESSON, *The Novels of August Strindberg* (1968); GORAN LINDSTROM, "Strindberg Studies 1915–1962," *Scandinavica*, 2:27–50 (1963).

(B.Mo.)

Stringed Instruments

This article covers the family of musical instruments known as stringed instruments, or chordophones, and includes their structure and history and their comparative, symbolic, and acoustical qualities. All chordophones produce sound by the vibration of strings, which may be made of vegetable fibre, metal, animal gut, or plastic. In nearly all stringed instruments the sound of the vibrating string is amplified by the use of a resonating chamber or soundboard. The string may be struck, plucked, or rubbed (bowed); in each case the effect is to displace the string from its normal position of rest and to cause it to vibrate in complex patterns.

Since most stringed instruments are made from wood or other easily perishable materials, their history before written documentation is almost unknown, and contemporary knowledge of "early" instruments is limited to the ancient cultures of the Far East, India, the Mediterranean, and Mesopotamia; but even for these places histori-

ans must depend largely on iconographic sources rather than surviving specimens.

Chordophones seem to have spread rapidly from one society to another across the length and breadth of Eurasia by means of great population shifts, invasions and counterinvasions, trade, and, presumably, through sheer cultural curiosity. In the Middle Ages the Crusades stimulated Europe in the adoption of a whole set of new instruments; similarly, the Chinese accepted many new instruments from their Central Asian invaders. Indeed, the only world area that did not echo to the sound of strings was pre-Columbian America.

No system of classification can adequately categorize the interactions of natural material, craftsmanship, and man's exuberant imagination that produced an endless variety of stringed instruments. The most widely accepted system of classification is one devised in 1914 by the writer-musicologists Erich Moritz von Hornbostel and Curt Sachs, a method based on the type of material that is set into vibration to produce the original sound. Thus, stringed instruments are identified as chordophones; that is to say, instruments in which the sound is produced by the vibration of chords, or strings. This main category is then further divided into four subtypes arranged according to the manner in which the strings are positioned in relation to the body of the instrument. Within these categories, the descriptive nomenclature of an instrument is given in terms of parts of the human body: for example, the belly (front), back, sides, and neck. Instruments are not necessarily related only to others in the same classification. Transformations continually occur, and "hybrids," according to the Sachs–Hornbostel system, may in fact represent altogether viable subtypes of their own. The instrument maker, who usually has been a forgotten link in the musical chain, has always represented a blend of conservatism with the ability to quickly seize on and use a new constructional technique, a new tool, or a new material. The instrument maker's contribution to both the history of music and the history of musical instruments has been enormous and little appreciated.

The role of the instrument maker

This article is divided into the following sections:

The production of sound
Types of instruments
  Lutes
  Zithers
  Lyres
  Harps
The music of stringed instruments
  Solo uses
  For accompaniment
  Ensembles
Social and cultural associations
Stringed instruments as objects of art
The development of stringed instruments
  The harp
  The lyre
  The zither
  The lute
  Bowed instruments

THE PRODUCTION OF SOUND

The ear, because of its own structure, adds to and subtracts from the outside sound. It is, for instance, relatively insensitive to low-frequency sound pressure but is extremely sensitive to fine degrees of pitch change. At the same time it can accept a great number of pitch and tuning systems. On a worldwide basis, there is a large and varied number of tonal systems. The oldest known of these in the West is the so-called Pythagorean system; others include mean-tone tuning, just intonation, and the even-tempered system, methods of tuning calculation that vary slightly in the exact size they assign to the intervals within an octave. All of these systems represent theoretical mathematical concepts to some degree, and their origins must be sought in arcane numerological systems rather than in practical musicianship. Thus, "tuning" and "playing in tune" do not necessarily refer to the same thing; players and tuners make constant adjustments to any basic mathematically determined framework according to their judgment and experience. Again, a given "scientific" tuning system outlines scales and modes; the instrumentalist who plays an instrument with great pitch flexibility (the violin, for instance) spends much time in the spaces between the notes assigned in the given scale. Again, the Japanese zither, koto, can be tuned according to a number of fixed systems, and, even so, its player produces many microtonal variations on these fixed pitches by manipulation of the strings. Even in such a technological society as that of the modern West, the well-tempered scale is honoured more in the breach than in the observance: a piano tuner would not think of tuning altogether according to its dictates and uses a so-called stretched tuning, in which he imperceptibly sharpens (raises) pitches as he ascends, thus making the highest notes relatively sharper than the lowest ones. Again, investigation has disclosed that string players tend to play in the Pythagorean rather than the well-tempered system.

Inconsistencies, then, are inherent in all tuning systems; makers of fretted lutes (the guitar or the Greek lute, *laghouta*, for example) operate according to a combination of ear and rule of thumb when they insert frets (note-position markers—*e.g.*, of gut or wire) in the fingerboard. Such instruments are fretted according to the "rule of the eighteenth," in which the first fret is placed at one-eighteenth of the distance from the top to the bottom of the string, the second, one-eighteenth of the distance from the first fret to the bottom, and so on. Even if this method produced an acoustically perfect scale (which it does not), the player would not be able to reproduce this exactly, for as he presses the string to the fingerboard, the string is stretched and is thus slightly lengthened. That is why the act of stopping a string at its exact centre gives a note slightly sharper than the expected octave above the open string. Despite all of this, the search for an acoustically perfect tuning system goes on.

Though constructional methods differ widely from one area and instrument to another, there are a limited number of basic problems to be overcome by the maker of stringed instruments. The very principle that makes it possible for chordophones to sound is string tension; at the same time, tension is destructive to the instrument since it tends literally to pull it apart. So the body of an instrument must be made of strong material; it must be reinforced, and at the same time it must not be so rigid that it cannot easily resonate; *i.e.*, produce a supplementary vibration intensifying that of the string. The challenge of reconciling these opposite needs is the central one for the chordophone maker. Climate, too, has a marked effect on musical instruments: humidity expands a wooden instrument, and dryness contracts it. Of these factors, dryness is the most harmful, since the contraction of the wood actually pulls the instrument apart. Much energy has been expended over the centuries in investigations of various varnishes, shellacs, glues, and sealers. Many makers prefer to make their instruments in dry conditions, for the expansion caused by humidity is unlikely to prove as harmful as the contraction caused by dryness.

Effect of climate on instruments

However common the practice may have been in the past, wooden instruments are not now usually gouged out of one piece of wood (though the Japanese lute, *biwa*, portions of the koto, and often the Puerto Rican lute, *cuatro*, are thus fashioned); it is too painstaking and time-consuming for the maker. Instead, the instrument is built up of many pieces of wood glued together; the shaping of curved pieces is accomplished by gouging and planing (as in the belly of the violin) or by heating and pressing in a frame (the sides of the violin or guitar). Soundboards, the most important part of the resonance system of stringed instruments, are carefully planed to close tolerances. Mass-production methods are unsuitable for the production of high-quality chordophones because no two pieces of wood are precisely alike in their acoustical qualities; each piece of wood requires special judgment and treatment. Ideally, therefore, stringed instruments of the highest quality must be individually

Making of modern pianos

made. Piano manufacture is a partial exception to this rule, but even in a piano factory, individual treatment and craftsmanship are allowed full sway. The modern piano is a product of several different factories. The cast-iron frames are made by specialized foundries—the steel strings, the keyboards, and the actions by specialized firms. Each of these processes requires experienced craftsmen; and the work of assembly, polishing, tuning, and tone regulation calls for hours of individual attention to each instrument.

The construction and maintenance of stringed instruments has been complicated over the centuries by a continual rise in standard pitch, requiring strings to be tightened. Older instruments (such as a Stradivarius violin) have been subjected to additional physical strain and therefore needed heavier bass-bars (braces under the belly).

As already stated, the three possible methods of sound production on a stringed instrument are plucking, striking, and bowing. A string vibrates in a complex way: the entire string vibrates in one segment (producing the fundamental pitch), and various segments at the same time vibrate independently to produce overtones. The resulting sound is weak indeed unless the instrument is provided with a resonator to amplify the sound. The shape of the resonator varies greatly. It has been influenced by the materials, tools, and technology available in the society, the symbolic meaning of the shape, and the sound desired by the culture. The last factor seems to be governed by the first three: that is to say, the prescribed shape of the resonator affects the overtone structure of the instrument, producing a certain timbre, which the society in question then defines as attractive sounding.

Importance of the shape of the resonator

One of the clearest illustrations of the basic importance of the shape of the resonator to a musical instrument is the African mouth bow (a musical bow in which the player inserts the bow in his or her mouth). By varying the size and shape of the oral cavity while striking or plucking the single, unfingered string, the player produces a clearly perceptible, if quiet, melody that exists only because the changes in the mouth emphasize various overtones. On stringed instruments with permanently fixed resonators, the size, dimensions, shape of apertures, thickness, and bracing of the resonating surfaces largely determine which overtones will be emphasized and, thus, what the instrument will sound like. On a good violin, for example, the resonances of the body of air enclosed in the body of the instrument and of the belly should be close in pitch to the two strings A and D, thus amplifying and colouring these pitches and their overtones. The sound quality of a stringed instrument is also influenced by the thickness and material of the strings; primarily, however, it is the size and shape of the resonating body and especially the material, density, and thickness of the soundboard that determine the sound of an instrument. A well-known Spanish guitar maker, in a successful attempt to prove the importance of the belly of the guitar, once constructed an instrument—an excellent one—from the acoustically dead material papier-mâché, except for a carefully chosen and wrought wooden soundboard. Makers, then, devote a large part of their skill and knowledge to the choice of material for the soundboard; the maker of wood-bellied instruments prefers old wood because it is dry and well seasoned.

The Puerto Rican *cuatro* maker uses wood obtained from old houses; some guitar makers find the soundboards of discarded pianos unusually suitable for their purposes; makers of the classical Chinese zither, ch'in, preferred old coffins.

The timbre of a struck or plucked stringed instrument is also affected by the manner of setting the string into motion. A string plucked with a sharp point (the player's fingernail or a plastic plectrum) emphasizes the higher overtones, thus creating a "bright" tone quality. By contrast, a soft pad, such as that on a piano hammer, emphasizes the fundamental pitch. The relative hardness of the hammer on the piano is thus of critical importance to the sound of the instrument, and the final process in piano manufacture is voicing. A skilled worker adjusts the timbre of the instrument by the simple expedient of pricking the felt hammers with needles until he achieves a unified quality throughout the range of the instrument. The tone of an instrument is markedly affected by the place where the string is struck. The permanently fixed striking place on keyboard instruments has to be chosen with a combination of concern for the timbre and the mechanical requirements of the instrument. On all other stringed instruments the player varies the tone quality by choosing to pluck, strike, or bow at various places along the length of the string.

"Voicing" a piano

Another way in which musicians and musical instrument makers influence the sound of their instruments is by the use of sympathetically vibrating strings. On the piano, for example, when the so-called damper pedal is raised, thus leaving the strings free to vibrate, the act of striking one note causes all closely related pitches to vibrate in sympathy, thus modifying the loudness and tone of the struck note. This effect (which is encountered also on the zither and harp) is not a central feature of these instruments, but there are a number of Eurasian chordophones on which the principle is of fundamental importance. The North Indian plucked instruments, sarod and sitar, each possess about a dozen sympathetic strings tuned according to the notes of the mode being played. The Indian fiddle, sarangi, has from 35 to 40 sympathetic strings; the Norwegian Hardanger fiddle (*Hardangerfele*) has four sympathetic strings, the viola d'amore six or more. Sympathetic strings are generally made of thin brass or steel, and their vibration reinforces the upper harmonics, thus producing a bright, silvery sound.

TYPES OF INSTRUMENTS

Lutes. Probably the most widely distributed type of stringed instrument in the world is the lute (the word is used here to designate the family and not solely the lute of Renaissance Europe). The characteristic structure consists of an enclosed sound chamber, or resonator, with strings passing over all or part of it, and a neck along which the strings are stretched. The player moves his fingers up and down the neck, thus shortening the strings and producing various pitches.

In the lute the part of the resonating chamber over which the strings pass is called the belly and the other side of the resonator is called the back. The portion between the back and belly is the side, or rib. A lute may be plucked with the fingers or a plectrum or may be bowed, but the means of sound production do not affect the essential morphological identity of plucked, struck, and bowed lutes.

Lutes may be subdivided into those with skin and those with wood bellies; in most Eurasian cultures examples of both types exist side by side. In Japan the wood-bellied lute is the *biwa*, the skin-bellied one the samisen; in the Near East the wood-bellied one is the *ʿūd* and the skin-bellied, the tār; in the United States it is the guitar and the banjo, respectively. The two different varieties of lute are distinct in sound and structure; and methods of construction, timbre, history, and symbolic associations differ markedly. A second subdivision concerns the shape of the instrument; for instance, the lute proper has a round back, the guitar a flat one. These two basic constructional principles (the shape of the resonator and whether the belly is constructed from wood or skin) constantly interact, producing all sorts of combinations.

The string vibrations of the lute are transferred to the resonating chamber by a bridge; the resonator magnifies them and transmits them to the air. Makers lavish great attention on the choice and fashioning of material for the belly: if it is of wood, it must be selected and aged with much care and planed to a prescribed thickness; if it is of skin, it must be made only from certain materials. (The belly of the Japanese samisen is preferably made from the skin of a female cat; the wooden belly of the Puerto Rican *cuatro* is best constructed from wood of the female *jagrumo* tree.)

Lute strings

The lute strings themselves have traditionally been made of animal intestines (gut) or metal or silk, though nylon has now replaced these older materials all over the

world. Whatever the material, each string must be of equal thickness throughout its length. Some lutes have only a single string, but the great majority have three, four, or more. Very often there are sets, or courses, of two strings to a pitch, so that an instrument that produces four pitches with open strings actually has eight strings arranged in pairs.

In the tunings of lutes, though fourths and fifths (intervals the size of four and five tones of a Western seven-note scale, as C–F and C–G) predominate in many places, any given instrument is likely to be tuned differently from one location, piece, or player to another. Functionally more important is the question of whether a lute is fretted or unfretted. It is clearly easy to change from one pitch to another on an unfretted instrument by sliding the finger along the string, but it is also possible to do this on an instrument with extremely high frets (Japanese *biwa*, Indian vina) by pressing—hence stretching—the string into the cavity between two frets. Instruments with low frets (the guitar, banjo, European lute, viola da gamba) are in use mainly in Europe, where a limited and clearly defined tonal system is in use and where significant microtonal changes in pitch are not necessary.

The fiddle can be distinguished from other lutes only by the manner in which it is played—with a horsehair bow. The practice of rubbing the strings with this implement is of uncertain age and origin, but it seems to have appeared almost simultaneously (9th–10th centuries AD) in China, Java, the Arab world, Byzantium, and Europe. As with the other lutes, there is a fundamental division between skin and wood-bellied instruments. (The former are far more common on the fiddle than the latter, which occur mainly in Europe.) Musically more significant, however, is the division between the stick fiddle, in which the player's finger does not actually press the string to a fingerboard (but rather slides up and down the string itself), and the fiddle with a fingerboard (for example, the violin). On fiddles without fingerboard (including the Chinese *erh-hu*, the Javanese and Arab *rabāb*, various African fiddles, and the east Indian sarangi), the player's left hand is able to move with extreme flexibility up and down the string, thus making the subtlest kind of inflection possible.

As with other lutes, fiddles may have only one string (the Tuareg *imẓhad*) or as many as 40 (the sarangi); on the latter, sympathetic strings are not directly touched or sounded by the player but vibrate sympathetically when other strings are set into motion, thus giving a fuller resonance. Examples, in addition to the sarangi, include the Norwegian *Hardangerfele*, the Swedish nyckelharpa, and the viola d'amore.

The fiddle bow itself generally is constructed so that the player can tighten or loosen the hair at will; on most stringed instruments the player is able to make immediate changes by manipulating his hand on the bow hair while playing, thus producing various tone qualities. The modern Tourte bow of the violin has a screw mechanism that cannot be changed while playing. Most bows, again, are actually bow shaped, but the Tourte bow is made in a compound curve to which considerable tension can be applied, making it possible to apply much pressure to the strings.

The fiddle bow

The bowing principle has been applied to non-lutes from time to time: the old Icelandic *fidla* was a bowed zither; the Scandinavian *talharpa* was a bowed lyre.

Zithers. Instruments of the zither family, in which the strings lie parallel to and are of the same length as the resonator, are widely distributed in Eurasia, the Americas, and Africa. There are two important subdivisions of this category. The so-called long-zither family is found only in the Far East; its characteristic resonating chamber is slightly convex, as much as six feet (180 centimetres) long, and about a foot (30 centimetres) wide; there is a varying number of strings frequently provided with movable bridges. These instruments, of which the best known example is the koto, seem to devolve ultimately from zithers made directly from lengths of bamboo. The bamboo prototypes are said to be idiochordic because their strings, part of the bamboo itself, are worked loose from the tough surface of the tube, to which they remain attached at either end. The maker then inserts small bridges at the extremes of the strings. (Various modifications and transformations of this principle exist, as the bamboo-tube *valiha* of Madagascar and Malaya, in which wire strings replace the idiochordic ones). All long-bodied, curved-surfaced Far Eastern of koto type may owe something to this idiochordic principle. In Eastern tradition the most ancient zither is the seven-stringed ch'in, which seems to have originated in the Shang dynasty (*c.* 1400 BC). The Japanese *wagon* and koto, the Korean *kayagum*, and the Chinese *cheng* fit into this general category. All these ancient instruments are rich in symbolic associations.

The other important subdivision of the zither family is the flat zither; in Africa it is made either from a hollowed plank over which strings are fastened (board zither) or from individual narrow canes lashed together, each having one idiochordic string (raft zither). The box zither, which seems to be native to the Middle East, is a rectangular- or more often trapezoid-shaped hollow box the strings of which are either struck with light hammers or plucked. Examples of the former are the Persian *santūr* and its modern Chinese derivative, the *yang ch'in* ("foreign zither"), the cimbalom of East Central Europe, and the piano (which is a sort of cimbalom with keyboard). The most prominent plucked box zither is the Arab *qānūn* and its various derivatives, including the harpsichord (a plucked zither controlled by a keyboard). In Europe a variety of plucked zithers developed having a fretted fingerboard under one or a few of the strings.

The flat zither

Struck zithers are occasionally termed dulcimers, unfretted plucked ones psalteries, after European instruments using those names.

Lyres. The lyre family, though it was of great importance in the ancient centres of Babylonia, Egypt, and Greece, is now found only in a few areas of East Africa. A lyre is made from an oval, round, or rectangular sound chamber (usually skin bellied); from this resonator two arms protrude; they are joined at the top by a crosspiece; the strings extend from this crosspiece over the belly, with which they are connected by a bridge. These strings are not normally stopped but are allowed to vibrate throughout their entire length when plucked by the performer.

In Ethiopia, where the lyre remains in extensive use, there are two distinctive types, as there were in ancient Greece. The ten-string *begenna* (which corresponds to the ancient Greek kithara) is a large, heavy, rectangular-shaped instrument that is considered by the Christian Ethiopians to be a God-given instrument that came to them from King David; it is used, of course, for sacred music. The smaller lyre, *krar* (the ancient Greek *lyra*), is emphatically secular in its use and connotations; indeed, Ethiopians think of it as the instrument of Satan. The construction of this five-stringed instrument illustrates the sort of change that is of wide occurrence in contemporary instrument making everywhere, for though the traditional *krar* was made from wood, the resonator of the present-day instrument is made of an easily available metal pan. The *krar* can be played in two ways: in the first (called muting) the left hand mutes the unwanted strings while the right hand strums with a plectrum; in the second, the fingers of the left hand pluck while the right hand plucks a drone on tonic strings tuned to the tonic, or focal note, of the melody.

The *begenna* and the *krar*

The lyres of medieval Western Europe (4th–12th centuries) had from five to seven strings and, to judge from iconographic evidence, were played in a way that closely resembled the muting technique of Ethiopia. Northern European shapes are considerably varied, but both the rectangular kithara-like shape and the rounded *lyra* shape apparently existed.

Harps. The harp family exhibits an extraordinary variety of constructions, but in all harps the strings are of unequal lengths and are fastened at either end to a frame that lies in the same plane as the strings. The least complex, and prototypical, harp type of instrument is the musical bow, shaped very much like a hunter's bow. The

single string is tapped or struck, and the pitch can be varied by varying the tension of the string or by using the player's mouth as a resonator and varying its size and shape, thus emphasizing different harmonics. It is a favourite instrument in equatorial Africa.

The arched, or bow-shaped, harp, known in Egypt as early as 3000–4000 BC, is a multistringed version of this musical bow; its player kneels or stands, supporting the harp on the shoulder. Harps of this type may be found in Western and Central Africa, where they are often provided with elaborate anthropomorphic carvings and skin-covered resonators. The bow harp, then, is a traditional African form that has been in use in that continent for at least 5,000 years.

On an angle harp the bowlike support is replaced with two crosspieces at right angles to one another; the strings are stretched between these at an angle of 45°. This type seems to have originated in Assyria, though occasionally it is found in Egypt and Greece.

The closed, or frame, harp is characteristic of both medieval and modern Europe; the shape of its frame is more or less triangular, the frame being strengthened by a pillar that encloses the strings in a kind of tripartite structure. It is to this category that the modern orchestral harp of Europe and the old Irish and Scottish harps belong. In all of these instruments the crosspiece held nearest the player is a hollow resonating chamber. The so-called Brian Boru harp (14th century), now at Trinity College, Dublin, is 32 inches (81 centimetres) high, with 36 brass strings; the sound box is carved from a single piece of willow, and the harp was plucked by the fingernails.

THE MUSIC OF STRINGED INSTRUMENTS

Solo uses. Music for a solo instrument is often, though not always, used to accompany dancing or is derived from dance music. In Europe and America, the violin is widely used to play dance music; for instance, the Norwegian Hardanger fiddle player performs rhythmically complex polyphonic music (*i.e.*, composed of several simultaneous melodic strands) to accompany the *halling, gangar,* or *springar.* In Scotland, Ireland, and rural North America, the violin is favoured for accompanying country dances. It is known that a single fiddler played for dancing among the upper classes in medieval Europe and also probably among the common people, with whom the fiddle was popular. In all areas, the fiddler has an additional repertory not for dancing, which shows off his compositional and improvisational ability as well as his virtuosity. A skillful medieval fiddler constantly improvised new pieces. It is no coincidence that the violin solo sonatas and partitas of Bach, like their country cousins, are polyphonic compositions having a liberal sprinkling of stylized dance movements.

A second genre of solo piece is the descriptive composition. In such diverse places as Ireland, West Africa, and the United States are found fiddle pieces picturing the hunt, in which the player delights in the depiction of such realistic sounds as the baying of the hounds, the sound of hooves, and the groans of the exhausted animal. In China there is a genre of highly realistic descriptive compositions for *p'i-p'a*—including some that depict famous battles from start to finish. Among numerous descriptive pieces in Renaissance and Baroque Europe, a well-known example is a sonata for viola da gamba, by the French virtuoso gamba player Marin Marais (1656–1728), that graphically and in detail describes a gallstone operation in the days before anesthetics.

Solo music of the Middle Ages and Renaissance

In the European Middle Ages and Renaissance, an important type of composition for solo stringed instruments was the transcription of polyphonic songs; in the medieval period fiddle and hurdy-gurdy players frequently rendered elaborated versions of well-known strophic songs. In Northern Europe, songs intended to accompany dance could also be played on the fiddle, and in Spain the *vihuela de arco* filled a similar function. In the Renaissance, a string player was expected to have the ability to perform highly elaborated versions of pre-existent melodies on the viol, lute, or keyboard; and many manuals provided the player with exercises designed to teach him to elaborate the basic melody in a stylistically correct manner. Similar techniques continued to be practiced during the Baroque period and even into the 19th century.

For accompaniment. All over the world, stringed instruments have as one of their central functions the accompaniment of bards, who sing tales of heroes and otherwise reflect the preoccupations of their societies. The example of Homer and his lyre is of course the best known, and even today in the Balkans, bards, self-accompanied on the fiddles, *gusla,* or *lira,* play preludes and interludes while they sing tales of heroes in the wars against the Turks. The harpers of Celtic Ireland and Scotland devoted their skills to the accompaniment of learned bards; the *azmari* of Ethiopia sings lengthy historical epics and strophic love songs to his own accompaniment on the fiddle or lyre. In Japan, blind *biwa* players chant a narrative style of music known as *katarimono,* and here again the *biwa* is used only between verses, as it performs interludes and commentaries. A similar technique is in use among the minstrels of North Africa: the lute, *gunbrī,* is played only between verses of the story, as a descriptive comment. Again, the 19th-century European art song for solo voice and piano belongs to this genre.

Ensembles. Musical ensembles everywhere have their own internal social structure, mirroring that of their society at large in the degree of authoritarianism, the type of leadership, amount of freedom available to the individual players, and so on. The audience for a given ensemble also tends to be socially stratified. Large, authoritatively ruled ensembles tend to be found in societies that have a complex bureaucratic and pyramidal social order. Such groups are most typically to be found clustered around the royal or princely courts of China, Japan, and Korea, of Java and Bali, of North Africa, and in the court-derived music of Europe and America. In all of these instances a more or less fixed and archaic repertory remains in use. In Japan, Korea, and, to an extent, Southeast Asia, the court music has derived from an archaic (and no longer extant) court music of China. The two types of stringed instruments in use in Japan are the zithers, koto and *wagon,* and the lute, *biwa.* The koto has 13 strings and the *biwa* four.

The gamelan orchestras of Indonesia employ but two chordophones in ensembles, which are otherwise dominated by struck metallophones, or metal instruments, such as tuned gong sets and xylophone-like instruments with metal bars. The bowed *rebab* (probably entered the orchestra from the Near East, where it was called *rabāb,* ancestor of the European rebec, when Java was converted from the Hindu religion to Islām). It plays in sliding fashion around the fixed pitches of the orchestra; its player is thought of by Javanese as the "rajah" of the gamelan, with the drummer as his "prime-minister."

In Islāmic countries and in Europe, stringed instruments predominate numerically in the orchestra; the classical ensembles of the Maghrib, which mainly perform the ancient Andalusian suites, typically consist of about one dozen musicians. The instrumentalists include a number of violin or viola players, a *rabāb* player, lute players, drummers, and sometimes flutists, cellists, and a pianist. All of the string players play the melody together but not in unison, for they are expected to vary it individually with ornaments and improvised passages. A true unison, therefore, is neither expected nor desirable. The ensemble is normally led by an *ʿūd* player who indicates the pitch, mode, mood, the changes of tempo, and so on.

Orchestral stringed instruments

In Europe as in North Africa, the greater number of stringed instruments in use in the orchestra are bowed; Europe, however, is distinctive in that these bowed chordophones exist and are played in families of graduated size (violin, viola, cello, bass), and this phenomenon derives from the extent to which European musicians have been preoccupied with harmony and multipart composition. In the development of the European orchestra two great—but gradual—changes have occurred over the centuries: first, stringed instruments have increased dramatically in number in the orchestra; and second, higher and higher notes and hence finger positions have been de-

manded of string players. Socially and musically, the role of the player in an early 18th-century symphony (in which there would be no more than two or three players to each part) is very different from the anonymous function allowed the violinist in a symphony by Berlioz, say, or Mahler. The symphonic orchestra seems to represent the merger of two functionally and aesthetically different groups that medieval Europe probably derived from Arab civilization. One of these, used for outdoor ceremonial music, consisted originally of shawms (oboes), drums, and trumpets; the other, for indoor (chamber) use, was made up of stringed instruments, quiet percussion instruments, and flutes. The modern orchestra preserves vestiges of these two ensembles in its seating arrangements, orchestration, and leadership.

The European concept of chamber music is nowadays thought of as being entirely distinct from orchestral music; in the 17th and early 18th centuries, however, the distinction would have been difficult to draw, since there was no clear difference either in the music, the size, or the sociological position of the two types of ensembles. The symphonic orchestra arose because of middle class demand for public concerts and from the desire of composers after the French Revolution to provide democratic music for the masses; thus, mass audiences demanded large orchestras, and the striking multiplication of instrumentalists affected the strings above all others.

Modern Japan cherishes a number of traditional chamber-music ensembles; of these, the commonest one consists of the koto, the samisen, and the flute, *shakuhachi*. The *jiuta* is a common form played by this group; here, the koto plays the principal melody, and the samisen and *shakuhachi* simultaneously produce a variant of it. The Japanese picturesquely describe this ensemble by saying that the koto is the bone, the samisen the flesh, and the *shakuhachi* the skin.

Basic elements of Indian ensembles

The basic ensemble in South Indian music consists of a drone instrument (usually the tamboura), a tuned drum (mṛdaṅga), a violin, and a singer. The group is often augmented to include a number of other instruments, but the basic elements must be present. The violinist is an accompanist to the singer; he must be able to play in unison with him in fixed pieces and to follow him in immensely complicated improvisational sections, maintaining all the while the correct raga (melodic mode) and *tāla* (rhythmic mode). An analogous group in North India might include the sitar or sarod as solo instrument, with the accompaniment of the drums, *ṭablā*, and a drone. Often the fiddle, sarangi (with three main strings and 35–40 sympathetic ones), accompanies a singer.

Rather similar principles apply in the Eastern Mediterranean: a chamber-music ensemble must include instruments—usually strings—to play the melody and to improvise in the introductory sections known as *taqsīm* and percussion instruments to deal with the rhythmic modes, known as *īqāʿāt*. To these, in modern Greek music, the element of harmony has been added so that a solo instrument (clarinet or violin) plays the highly ornamented melody and another instrument (guitar, or *laghouta*) plays the chordal background and maintains the reiterated rhythmic pattern. Similar ensembles are found in the Americas and in Africa.

In Puerto Rican traditional music a guitar plays the harmony, the five double-stringed lute, *cuatro*, plays the melody with the singer, and the scraper, guiro, and drums, *timbales*, produce the rhythmic part. Among the Berbers of North Africa, groups of itinerant professional musicians typically play one or two lutes, *gunbrī*, a one-stringed fiddle (rendering a highly adorned melody), and percussion. In Senegal, lutes similar to the *gunbrī* are played by the musicians known as *griots*, often in ensembles of three or four flutes, drums, and rattles. The so-called Bluegrass style of Appalachian North America largely conforms to this Mediterranean and African pattern; the ensemble typically includes only bowed and plucked strings and singers.

Rock groups also adhere to these general principles: they also represent a blend of African and European traditions. The typical rock ensembles use plucked electric stringed instruments, electric keyboard instruments, voice, and drums. The electric guitar plays the melody, and the electric bass renders the bass part. In the American blend of Mediterranean, African, European, and indigenous tradition, the African element includes the percussion instruments and the skin-bellied lute (banjo); the Mediterranean element includes the wood-bellied guitar and *cuatro*. The bass parts are derived from Europe, the scraper from the American Indians, and the singing style from various of the four traditions.

SOCIAL AND CULTURAL ASSOCIATIONS

Throughout the world, particular instruments are associated with particular functions, social levels, areas, ages, and sexes. Certain instruments preserve the same function all over the world: the bard (a singer-poet) recounts tales of heroes, recites the genealogies of the great, and preserves and disseminates the precious traditions of a given society; this bard everywhere accompanies himself on a stringed instrument. An early example would be the blind Homer, who, presumably accompanied by his lyre, sang of gods and heroes. Similar traditions survive even today in Greece and the Balkans, but the instrument of the modern "singer of tales" is the fiddle. In Ethiopia, where so many traditions remind one of Greece, the *azmari* accompany themselves with a lyre or fiddle. In North Africa and West Africa *griots* sing and play skin-bellied lutes called *gunbrī* or *halam;* in Morocco they play and sing alone in the squares of the towns for coppers; in Senegal (where they perform in groups) they are employed by the wealthy. The Japanese *biwa* (often, again, played by blind bards) was used to accompany epic narratives, the texts of which usually concern the adventures and battles of the samurai: the *biwa* is used also to accompany Buddhist chanting. In medieval Europe, minstrels, self-accompanied on the *vielle* (fiddle), sang of the heroism of such as Charlemagne.

Very often, stringed instruments are identified with a certain social class, either in the sense that the player himself occupies a particular caste position or in that the instrument is the prerogative of a particular rank in society.

A well-known example is that of the Irish and Scottish harp of the periods during which the Celtic kingdoms flourished; harpers, who were highly respected members of society, were employed in the households of only the very wealthy and the high nobility. By the same token, harpers were regarded with great suspicion by English officials in Ireland, for they were suspected of spreading sedition. In sharp contrast is the attitude of respectable members of medieval Arab society toward lute players; their status was low indeed (certain judges even held that the lute was an unlawful possession which could be destroyed at will). Paradoxically, the *ʿūd* player's skill and imagination was highly valued by the courts. These attitudes are old indeed in the Mediterranean; Aristotle regarded professional kithara players as low caste and of questionable morals. In his view (and this is still prevalent in the West) the highborn should learn to play as an accomplishment but not professionally. In the 17th century, members of the English upper class were swept with a craze for playing consort music for ensembles of various sizes of viola da gamba ("leg viols," held, cello-fashion, between the knees). It was a common—and much valued—accomplishment to be able to play a part in a viol consort, and, as Samuel Pepys recorded in his diary, servants were often chosen because of their ability as violists.

The violin family in Europe has had two very different social connotations. In the string quartet and orchestra the instrument has been associated with the entertainment of the upper class and the bourgeoisie, and this represents the continuity of the medieval tradition under which the fiddle was highly regarded by the upper class. Alfonso X, the Wise, King of León and Castile, was one of many rulers who supported a group of fiddlers at his court.

In opposition to this, the violin has long had great popularity among the folk. In almost every country in Europe

The aural technique of the country fiddler

and the Americas the rural fiddler has provided the music for dances and entertainments. This country fiddler is distinguished from his urban colleague by his nonstandardized technique, playing position, and tunings and by his learning process, which is totally aural. His virtuosity may be of a high order. The pleasures of the common folk have long been associated with the fiddle, though the medieval fiddler usually played the bagpipes, sang, juggled, and told jokes as well.

The role of musical instruments in religious and healing practice is complex: at one time or another Judaism and Islām and Christianity have been generally hostile to the use of instruments in places of worship. The Church Fathers made frequent allegorical and symbolic reference to instruments, and it is common to find depictions of angels, cherubs, or of King David playing on the very instruments that were not permitted to sound in Christian worship. The Hindu attitude toward stringed instruments is very different. Though Vedic chanting is the oldest known music of India, the Ṛgveda nevertheless mentions the use of such chordophones as the *karkari* and a number of vinas.

Caste and symbolic associations of the ch'ins

The Chinese scholars' lute, ch'in, had definite caste and symbolic associations. No scholar would have been without one, for it restrained licentiousness and purified the heart; it compelled man to return to his divine origin and restrain his passions. Thus, the ch'in could not be played when drunk or after sexual intercourse; it could not be played for a courtesan, a merchant, or a barbarian. The ch'in was often played by Buddhist monks; complex cosmological and symbolic meanings existed for every aspect of its shape, constructional material, tone, and playing technique. The ch'in, therefore, was involved with healing, religion, and philosophy; and of these three, its therapeutic function was perhaps the most pervasive; this instrument harmonizes the blood circulation and regulates breathing; it leads to health and long life. It cures melancholy and can, it is claimed, even make a devoted player immortal.

The function of stringed instruments in the healing process is very often an indirect one: a healing specialist may sing his powerful songs to the accompaniment of a lute (the *gunbrī* of the North African Gnāwa) or a fiddle (Central Asia). Beyond this relatively prosaic association, stringed instruments have symbolic meanings that transcend their technological, sound-producing function. The horse-headed fiddle of the Mongols is directly a part of the Mongol system of belief in the horse as a symbol of fecundity and rebirth; the winged horse that the shaman is believed to ride in his quest for enlightenment and healing power is equivalent symbolically to the fiddle.

This close connection of stringed instruments with ideas of birth and rebirth explains why so often a particular instrument is typically played, or even listened to, by members of one sex only; in certain instances instrumentalists, whether they were Greek kitharodists or Arab *ʿūd* players, were slaves with a low moral reputation. The playing of many stringed instruments by women has been either a low-caste occupation or an accomplishment to be savoured in the confines of the home. In India during the Vedic period there seem to have been stringed instruments (such as the *kāṇḍa-vīṇā*) that were played only by women; and Chinese tradition has it that the long zither, *cheng*, was invented by a woman. In the Western continents, instrument makers seem usually to be men, since in most societies the manipulation of woodworking tools has been a male occupation. In medieval Europe, and in some places (such as Puerto Rico) today, minstrels made their own instruments. Since the making and playing of instruments have until recently been mainly in male hands, women have not had available to them this particular link to musical instruments through manufacture.

STRINGED INSTRUMENTS AS OBJECTS OF ART

There is no reason to suppose that the shape of an instrument is governed by acoustical requirements; it seems rather to be the other way around: the symbolically appropriate shape preferred by a given culture produces a particular tone quality, which then becomes the desirable one. Available materials, manufacturing techniques, and complex historical, symbolic, and artistic considerations guide the craftsman who creates musical instruments. Transformations occur as old meanings are lost: the scroll surmounting the pegbox of the violin may well be a faint visual echo of the horse's-head carving that surmounts the fiddle in its ancestral lands in Central Asia. Craftsmen are well aware that the soundboard of an instrument is responsible for the best part of its resonance; they are aware that other portions can be considerably altered without making much change in the sound quality. Instrument makers have long delighted in producing modifications of existing designs and in experimenting with new materials. In all this they are limited only by the essential technological requirement of a chordophone—ability to withstand the constant strain imposed by the tension of the strings and proper resonation.

A limited number of basic woodworking techniques are used almost the world over in the production of stringed instruments. In lute construction the entire sound box (and sometimes the neck as well) can be hewn from one block of wood. Instruments that are often made in this fashion are the Japanese *biwa*, the North African *gunbrī*, and the Puerto Rican *cuatro*. Existing side by side with this technique is another, in which the lute is made out of a kind of composite box in which any bent portions are fashioned by heating the wood and putting it into a mold. (Common examples include the violin and the guitar.) Again, the resonator can be made of pre-existent natural materials, as in the case of the Malagasy *valiha* zither (bamboo), the South Indian vina (gourd), or the lyre invented by Apollo (tortoiseshell).

Though occasionally (as in the case of many harpsichords and virginals) instruments with pictures painted upon them are encountered, this decorative technique is most unusual—probably because it requires a separate craftsman—and most instrument makers prefer to show their skill by the choice of woods, by complex inlays, and by carving.

The anthropomorphic harp

Anthropomorphic and zoomorphic shapes surmount many of the harps and lyres of ancient Mesopotamia and Egypt. The anthropomorphic harp of ancient Egypt and modern Africa represents another example of this visually symbolic approach to instrument making, which has been described as producing "sounding statues." In Europe, anthropomorphic and zoomorphic carving exists side by side. Notable examples are the Irish Brian Boru harp (at Trinity College, Dublin), which exhibits a profuse variety of animal shapes; the Norwegian *Hardanger* fiddle, surmounted with a dragon's head; and the viola d'amore, which often is surmounted with the head of a blindfolded woman. The Balkan fiddle, *gusla*, and the Central Asian horse fiddle are usually crowned with a carved horse's head, and the South Indian vina possesses a carving of a dragon. These shapes are not mere decoration or fancy; they represent important symbols in the society from which the instruments come.

In East Asia the craftsman's emphasis is on beauty of wood, excellence of finishing, and apparent simplicity of construction. Occasionally, on instruments such as the ch'in, one finds pictures painted or calligraphy inscribed. The proper material is extremely important: the belly of a good Japanese samisen should be made from the skin of a female cat; a ch'in should be made of wood taken from an old tree that had nestled next to a bubbling stream. Phoenix-bird symbolism of rebirth is abundant in the nomenclature of Far Eastern instruments, and this fact serves to point up a profoundly important truth about musical instruments everywhere: they are intimately connected in folktale, myth, and legend to local symbols of rebirth. Thus, legend relates that Apollo made the first lute from a turtle carapace; similarly, the first Arab lute was modelled after the body of a beloved male child; the Finnish culture hero, Väinämöinen, made the first zither, kantele, from the body of a giant pike; the Celts made their first legendary harp from whalebones. In each case, the symbolically significant creature is "reborn" as an instrument which "sings" as well as appears in a shape reminiscent of the creature modelled.

THE DEVELOPMENT OF STRINGED INSTRUMENTS

Little or no evidence is available concerning the chordophones of prehistoric times: the earliest iconographic evidence and the oldest surviving specimens come from Mesopotamia and Egypt, and evidence concerning instruments earlier than these can be gleaned only from myth and legend.

The harp. The harp clearly evolves from the musical bow, which must have been in existence prior to the establishment of Mesopotamian and Egyptian civilizations. Since this instrument is identical in shape with the simple hunter's bow, it has been assumed that the musical bow derives from the hunter's implement. There is, however, no evidence for this; quite the opposite: the musical bow is very often played by women, who do not usually hunt, and, further, there is absolutely no way of knowing whether or not man hunted with a bow-shaped implement before he made musical sound with a bow. The three early harps found in the burial chambers at Ur were bow-shaped instruments with 12 to 15 strings; nearly identical instruments were played in Egypt at roughly the same time. Iconographic evidence suggests that both in Egypt and Mesopotamia the harp, often played by women, was used in secular, erotic entertainments, although it had sacred uses as well.

Although the harp family is known to exist in Burma and to have been in use in China and India, it is not characteristic of these areas; it is, rather, musically important in equatorial regions of Africa, in Europe, and in the Americas. In Africa it clearly is related to the Egyptian harps, and again it is associated with women.

The Celtic harp

The harp has maintained and increased its importance in Europe. It is omnipresent in folktales and legends; it is the national symbol of Ireland. The Celtic harp must have been in use as early as the 10th century, and fragments of one were found in the 7th-century Sutton Hoo burial ship unearthed at Suffolk, England. In Gaul, Ireland, Wales, and Scotland, the harp was an important and favoured symbol; it was said that there were but three things necessary to a comfortable household—a virtuous wife, a chair cushion, and a harp. By the end of the 18th century the harp had almost gone out of use in the Celtic realms, but by then the large orchestral harp had made its appearance in other places in Europe.

This basic instrument (with the addition of a set of pedals for instantaneously altering the pitch of a note) is found—often still played by a woman—in the modern symphony orchestra; but the older European harp tradition flourishes best in Mexico and South America, where the harp is a component of various folk ensembles.

The lyre. Greek legend credits the invention of the lyre to Apollo, who had stolen Zeus's cows and, in order to atone for his transgression, presented the great god with the lyre, which he had accidentally discovered when he had brushed against a turtle carapace that lay on the ground and, as he did so, heard its sinews begin to vibrate. The tale is interesting for two reasons: first, the turtle shell *was* frequently used as the resonator of the Greek *lyra;* and second, the tale makes an explicit relationship between the lyre and cattle. Similarly, in Mesopotamia the lyre was surmounted with a carved bull's head, and in modern East Africa the lyre is most frequently encountered in cattle cultures.

A famous lyre from Ur (now at the University Museum, Philadelphia) is one of nine dug up at the burial ground; these and similar instruments seem to have been used both to accompany bardic recitations and for religious purposes. The "harp" that the Hebrews hung in the tree in their Babylonian captivity was actually a lyre, as was the instrument used by Homer (*phorminx*). In view of the importance of the bull in the worship ceremonies of Crete and Mycenae, it is not surprising to find lyres among the stringed instruments of these peoples. In Celtic society, depictions of lyres are found on the coins of pre-Christian Gaul. These instruments, which were U-shaped, may have come to western Europe from Southwest Asia with groups of Indo-European peoples who spread across Europe. Other types of lyres were found in Europe, too, and it is possible that these variously shaped but still related instruments might be analogous to the various Indo-European languages in that they are basically closely related but quite different in detail.

In medieval Germany and Scandinavia long, narrow lyres with four to seven strings were played, and similar lyres (the Finnish *jouhikantele*, the Swedish *talharpa*, the Welsh crwth) were played with a bow in parts of Europe until recent years. Today the lyre flourishes only in Ethiopia, in the Sudan, among the fishermen of the Persian Gulf, and in restricted areas of East Africa.

The zither. Several different types of instruments are classified as zithers; they are used today in all continents. The long zithers of China, Japan, and Korea are venerable indeed. Their curved surface and long, narrow shape betrays their affinity with the prototypical idiochordic bamboo zithers of the Pacific, Southeast Asia, and southeast Africa. The importance of bamboo to music in the Orient is literally legendary; in Java, music is thought to have been first produced by the accidental admission of air into a bamboo tube. In China, the source of Far Eastern civilization, musical instruments are classified according to their constructional material; one of the eight substances in the system is bamboo, which the Chinese relate to the direction East, the season Spring, and the phenomenon Mountain. The Chinese zither *cheng* includes the radical meaning "bamboo" in its ideograph. The ideographs of the older zithers, ch'in and *se*, are more difficult to interpret; but the narrow, curved shapes of the instruments themselves betray their affinity with the ancient principle. Zithers of this type are known to have existed in the Shang period (*c.* 1766–*c.* 1122 BC); the *cheng*, with either 13 or 16 strings, came into use during the Ch'in period (221–206 BC). By the 18th century, interest in this instrument had waned in China, but at an early date it was introduced to Japan, where, as the 13-string koto, it flourishes today. The koto, like the *cheng*, is played frequently by women.

Bamboo zithers

A relative newcomer to the spectrum of Chinese zithers is the *yang ch'in* (foreign zither), which reached China from Persia a number of centuries ago; it is the only representative of the box zither in the East. The history of this type of chordophone is obscure indeed, but two instruments of this general shape that may be very old are the African raft and board zithers. The raft zither is constructed on the idiochordic principle, but it uses a number of canes about half an inch in diameter; each of these has one string raised out of its own surface, and all of the canes are then lashed together. The board zither is made from a hollowed-out board over which a number of strings are attached. This latter instrument is found only in restricted areas of East Africa; it is possible that its principle of construction was carried to the Middle East by traders. Medieval Arab authors (including Ibn Khaldūn) mention the plucked trapezoidal zither, *qānūn* (derived from Latin *canon*, "rule"). Nowadays, the instrument has a range of three chromatic octaves with three strings to each pitch, and a complex system of levers by which its many strings may be finely and quickly retuned to the various Arab scales. Closely related is the Persian-derived *santūr*, another trapezoidal zither that is struck by two light hammers. Versions of this zither are found in China (*yang ch'in*), Greece (*santouri*), and Eastern Europe (cimbalom). These two zithers are the prototypes for the later keyboard instruments of western Europe: the *qānūn*, which is played with two plectrums, became, with the addition of a keyboard, the harpsichord; the cimbalom, with the addition of a keyboard, became the piano. In the Middle Ages, the keyboard was attached to a number of instruments, including the lute, the hurdy-gurdy, and the various Scandinavian keyed fiddles, of which the Swedish nyckelharpa survives. The experiment was truly successful, however, only on the clavichord, harpsichord, and later the piano; on the fiddle, it always remained of peripheral importance.

The lute. The first surviving evidence of the existence of the lute comes from Mesopotamia and Egypt. One of the earliest Babylonian delineations (*c.* 2500 BC) shows a shepherd with a long-necked, small-bodied lute; this in-

Earliest known lutes

strument, which was doubtless skin bellied, had a rounded back that might well have been made from a turtle carapace (as in modern North Africa) or a gourd (as in Puerto Rico). A similar instrument is also found in Egypt, and from specimens surviving from ancient times it is known that the stick forming the neck passed through the body of the instrument, ending at the sound hole, where it served as a fastener for the ends of the strings. This instrument, like the harp, survives in West Africa and in North Africa (as the *gunbrī* or *halam*).

The Greeks of old seem to have made rather sparing use of the lute, but it was commonly encountered in Rome. Though the skin-bellied lute predominated in the ancient Mediterranean, the wood-bellied lute has been more typical of Europe; indeed (with the exception of the African-derived banjo) more recently, all currently used European lutes and fiddles have wood bellies. The most important model for the European lute was the classical Arab lute, *al-ʿūd* ("the wood"), which has been the most important instrument of classical Arab music for 1,000 years. Not only the instrument itself, with its characteristically curved back, but also its name entered Europe, probably through a number of different gateways, including Sicily, Andalusia, and Palestine with the crusaders.

Though European and Arab lutes are very similar in general construction, the Arab lute is not fretted as is the European lute, and the European lute little by little acquired additional courses of strings during the Renaissance, when it became a standard instrument for solo playing and the accompaniment of singers. The many courses of strings caused numerous problems: "When a lute-player is 80 years old," wrote the 18th-century German theorist Johann Mattheson sarcastically, "he has spent 60 years tuning his instruments."

The many lutelike instruments of modern India (tambura, vina, sitar, and sarod) are, technically speaking, lute-zithers, since their fingerboards are hollow; in appearance and playing technique, however, they are indistinguishable from lutes. They seem to represent a special, and quite recent, subfamily of chordophones. In the Far East, again, the lute family exists only in a comparatively recent manifestation, but both in wood-bellied form (Chinese *p'i-p'a*, Japanese *biwa*) and skin-bellied version (Chinese *san-hsien*, Japanese samisen). In Japan, these instruments existed in the year 752, and the Imperial Treasury at Nara preserves three *biwas* used at a great concert held in that year. The *biwa–p'i-p'a* family can be traced ultimately to Persia, where, as the *barbat*, it influenced the music of Afghanistan and Turkestan on its way to China and Japan. The skin-bellied lute, *san-hsien*, can be traced in China only to the 13th century; from there it was taken to the Ryukyu Islands and thence to Japan (mid-16th century), where it still plays an important musical role as the samisen.

Lutes of the Far East

Bowed instruments. The principle of bowing is nearly always applied to stringed instruments of the lute class, though one occasionally finds it used with zithers or lyres. It is difficult, if not impossible, to make a clear-cut distinction between plucking with a plectrum and bowing, since plucking sometimes involves rubbing the string. But bowing, defined as the use of the almost universally encountered horsehair bow, can be traced as far back as the Islāmic civilization of the 10th century. The stroked *rabāb* evidently possessed a long neck and up to four strings. There is no iconographic evidence of such instruments before the 13th century in Islām itself, though an illustration (*c.* 930) from Christian Spain delineates a bowed instrument with three strings. In Byzantium, too, the bowed *lira* existed by this time, and it seems likely that the principle of bowing originated among the horse cultures of Central Asia, whence it spread quickly through Islām and the East, so that by 1000 it had almost simultaneously reached China, Java, North Africa, the Near East and Balkans, and Europe. Modern scholarship has shown that a word meaning "horse" is given to the bridge of the fiddle throughout Eurasia, from Japan (*koma*) to western Norway (*hest*). In China the fiddle is acknowledged to have come from the "barbarians" of Central Asia; in those days it was in China an instrument of the people; later, at the time of Kublai Khan, it began to be used in court music. In Africa, the fiddle seems to have entered the continent with Islām.

In Europe the bow appears first in Spain and shortly thereafter in Italy, and it is suspected that it entered through Muslim Spain and Sicily, though it is possible that other fiddles entered Europe through the Balkans, Hungary, and Scandinavia. Even in early illustrations, evidence is found of the uniquely European method of holding the fiddle against the shoulder. (Everywhere else in the world, the fiddle is held vertically.) At the same time, other European fiddles were held vertically, in the Oriental manner. Early European fiddles were made (by their players) in only two parts: the belly was a thin piece of spruce or fir wood, and the back was hollowed out of one piece of hardwood. These medieval fiddles divide into two types, one of which has a clearly discernible neck; the other possesses a neck that merges imperceptibly into the body; second type, exemplified by the rebec, is equivalent to the North African *rabāb* and the Byzantine–Greek *lira*. As the centuries passed, Europe continued to have two distinct types of fiddles: one, relatively square shaped, held in the arms, became known as the viola da braccia ("arm viol") family and evolved into the violin; the other, with sloping shoulders and held between the knees, was the viola da gamba ("leg viol") group. The gambas, which were important and elegant instruments during the Renaissance, lost ground to the louder (and originally less aristocratic) braccia-violin family.

The bowed instrument in Europe

BIBLIOGRAPHY. The standard history of instruments is CURT SACHS, *The History of Musical Instruments* (1940). A recent book that traces bowing on a global scale is WERNER BACHMANN, *Die Anfänge des Streichinstrumentent spiels*, 2nd ed. (1966; Eng. trans., *The Origins of Bowing*, 1969). Articles on the symbolic meaning of musical instruments are THEODORE GRAME, "Bamboo and Music: A New Approach to Organology," *Ethnomusicology*, 6:8–14 (1962), "The Symbolism of the *ʿŪd*," *Asian Music*, 2:25–34 (1972), and with GENICHI TSUGE, "Steed Symbolism on Musical Instruments," *Musical Quarterly*, 63:57–66 (1972). Two standard books on African instruments are PERCIVAL R. KIRBY, *The Musical Instruments of the Native Races of South Africa* (1934); and MARGARET TROWELL and K.P. WACHSMANN, *Tribal Crafts of Uganda* (1953). Instruments of Mesopotamia are discussed in H.G. FARMER, "The Music of Ancient Egypt," and "The Music of Ancient Mesopotamia," in EGON WELLESZ (ed.), *The New Oxford History of Music*, vol. 1 (1957); and F.W. GALPIN, *The Music of the Sumerians and Their Immediate Successors, the Babylonians and Assyrians* (1937). Articles on various African instruments are collected in H.G. FARMER, *Studies in Oriental Musical Instruments* (1931). An excellent work on the history and symbolism of the Chinese ch'in is ROBERT VAN GULIK, *The Lore of the Chinese Lute*, rev. ed. (1969). NICHOLAS BESSARABOFF, *Ancient European Musical Instruments* (1941), is a full and complete treatment of European instruments through the 18th century; HORTENSE PANUM, *Middelalderens Strengeinstrumenter og deres forløbere i oldtiden* (1915; Eng. trans., *The Stringed Instruments of the Middle Ages*, 1939), is an older but still useful work on European chordophones and their analogues; IRVING SLOANE, *Classical Guitar Construction* (1966), is an illustrated practical account of the process of guitar making. Two important Scandinavian instruments are fully treated in OTTO ANDERSSON, *The Bowed Harp* (1930); and JAN LING, *Nyckelharpan* (1967), with summary in English; Celtic harps are discussed in JOAN RIMMER, *The Irish Harp* (1969); and FRANCIS COLLINSON, *The Traditional and National Music of Scotland* (1966). *Ellenika Laika Mousika Organa* (1965), is an illustrated catalog of modern Greek instruments; and S. KRISHNASWAMI, *Musical Instruments of India* (1965), is a similar work on the instruments of the Indian subcontinent. Two standard works on Japanese music and instruments are SHIGEO KISHIBE, *The Traditional Music of Japan* (1966); and WILLIAM P. MALM, *Japanese Music and Musical Instruments* (1959). The standard work on Javanese instruments is JAAP KUNST, *Music in Java*, 2nd rev. ed., 2 vol. (1949), while a specialized study of the Javanese fiddle is found in M. HOOD, "The Javanese Rebab," in *Music, Libraries and Instruments* (1961).

(T.Gr.)

Submarine

A submarine is a naval vessel capable of operation underwater. Submarines first became a major factor in naval

warfare during World War I (1914–1918), when Germany employed them as commerce destroyers; they played a similar role on a larger scale in World War II (1939–1945), in both the Atlantic (by Germany) and the Pacific (by the United States). In the 1960s the nuclear-powered submarine, capable of remaining underwater for months at a time and of firing long-range nuclear missiles without surfacing, had come to be regarded by most military strategists as the most important of all strategic weapons. Nuclear submarines armed with torpedoes as well as antiship and antisubmarine missiles were in the 1970s considered a key element of naval warfare.

HISTORY

For centuries man has attempted to descend into the depths for scientific observation, for salvage, to reap the animal and mineral riches of the oceans, and for the purpose of attacking enemy ships in time of war. Herodotus, Aristotle, and Pliny the Elder all mentioned attempts to build diving bells or other such devices. A 13th-century French manuscript, *La Vrai Histoire d'Alexandre*, described a fictitious underwater venture by Alexander the Great (356–323 BC) in a glass barrel. Among the many inventions associated with the genius of Leonardo da Vinci was a device for underwater exploration.

The first submarines. The first serious discussion of a "submarine"—a craft designed to be navigated underwater—appeared in 1578 from the pen of William Bourne, a British mathematician and writer on naval subjects. Bourne published *Inventions or Devises*, in which the "18th Devise" was a completely enclosed boat that could be submerged and rowed underwater. It consisted of a wooden frame covered with waterproof leather; it was to be submerged by reducing its volume by contracting the sides through the use of hand vises. Bourne did not actually construct his boat, and Cornelis Brebbel (or Cornelius van Drebel), a Dutch inventor, is usually credited with building the first submarine. Between 1620 and 1624 he successfully manoeuvred his craft at depths of from 12 to 15 feet (four to five metres) beneath the surface during repeated trials in the Thames River, England. King James I is said to have gone aboard the craft for a short ride. Drebbel's submarine resembled that proposed by Bourne in that its outer hull consisted of greased leather over a wooden frame; oars extended through the sides and, sealed with tight-fitting leather flaps, provided a means of propulsion both on the surface and underwater. Drebbel's first craft was followed by two larger ones built on the same principle.

A number of submarine boats were conceived in the early years of the 18th century. By 1727 no fewer than 14 types had been patented in England alone. An unidentified inventor, whose work was described in the *Gentleman's Magazine* of 1747, proposed an ingenious method of submerging and returning to the surface: his submarine design had goatskin bags attached to the hull with each skin connected to an aperture in the bottom of the craft. He planned to submerge the vessel by filling the skins with water and to surface by forcing the water out of the skins with a "twisting rod." This arrangement was a forerunner of the modern submarine ballast tank.

Bushnell's "Turtle"

The submarine was first used as an offensive weapon in naval warfare during the U.S. Revolution (1775–1783). The "Turtle," a one-man craft invented by David Bushnell, a student at Yale, was built of wood in the shape of a walnut standing on end. Submerged, the craft was powered by propellers cranked by the operator (Figure 1). The plan was to have the "Turtle" make an underwater approach to a British warship, attach a charge of gunpowder by a screw device operated from within the craft, and leave before the charge was exploded by a time fuse. Piloted by Ezra Lee, an army sergeant from Lyme, Conn., the "Turtle" made a night attack on HMS "Eagle" in New York harbour. After repeated failures to force the screw through the copper sheathing on the warship's hull, Lee gave up, released the charge, and withdrew safely. The powder exploded without damaging the "Eagle," which shifted to a berth farther out in the harbour.

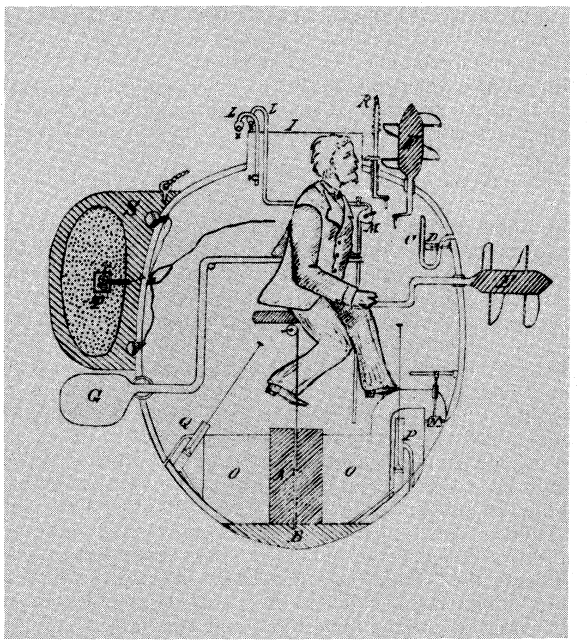


Figure 1: Bushnell's submarine torpedo boat, 1776. Drawing of a cutaway view made by Lieut. Comdr. F.M. Barber in 1885 from a description left by Bushnell.
By courtesy of the U.S. Navy

Robert Fulton, famed U.S. inventor and artist, experimented with submarines several years before his steamboat "Clermont" steamed up the Hudson River. In 1800, while in France, Fulton built the submarine "Nautilus" under a grant from Napoleon Bonaparte. Completed in May 1801, this craft was made of copper sheets over iron ribs. A collapsing mast and sail were provided for surface propulsion, and a hand-turned propeller drove the boat when submerged. A precursor of a conning tower fitted with a glass-covered porthole permitted observation from within the craft. The "Nautilus" submerged by taking water into ballast tanks, and a horizontal "rudder"—a forerunner of the diving plane—helped keep the craft at the desired depth. The submarine contained enough air to keep four men alive and two candles burning for three hours underwater; later a tank of compressed air was added.

The "Nautilus" was intended to attach an explosive charge to the hull of an enemy ship in much the same manner as the "Turtle." Fulton experimentally sank an old schooner moored at Brest but, setting out to destroy British warships, was unable to overtake those he sighted. France's interest in Fulton's submarine waned, and he left for England, offering his invention to his former enemy. In 1805 the "Nautilus" sank the brig "Dorothy" in a test, but the Royal Navy would not back his efforts. Fulton then came to the United States and succeeded in obtaining congressional backing for a more ambitious undersea craft. This new submarine was to carry 100 men and be powered by a steam engine. Fulton died before the craft was actually finished, however, and the submarine, named "Mute," was left to rot, eventually sinking at its moorings.

During the War of 1812 between the United States and England, David Bushnell built another submarine, which attacked HMS "Ramillies" at anchor off New London, Conn. This time the craft's operator succeeded in boring a hole in the ship's copper sheathing, but the screw broke loose as the explosive was being attached to the ship's hull. The attempt failed, but again the submarine escaped.

The next U.S. attempt at submarine warfare came during the Civil War (1861–1865) when the Confederate States resorted to "unconventional" methods to overcome the Union Navy's superior strength, exerted in a blockade of Southern ports. In 1862 Horace L. Hunley of Mobile, Alabama, financed the building of a Confederate sub-

marine named "Pioneer," a craft that was 34 ft (10 m) long and was driven by a hand-cranked propeller operated by three men. It probably was scuttled to prevent its capture when Union forces occupied New Orleans (although some records say the "Pioneer" was lost with all hands during a dive while en route to attack Union ships).

The second submarine developed by the same builders was a remarkably advanced concept: a 25-ft (7.5 m) iron boat intended to be propelled by a battery and electric motors. Not surprisingly, no suitable motors could be found, so a four-man-cranked propeller was again adopted. The submarine sank without loss of life in heavy seas off Mobile Bay while seeking to attack the enemy.

Confederate submarines

The third submarine of the Confederacy was the "H.L. Hunley," a modified iron boiler lengthened to between 36 and 40 ft (11 and 12 m). Ballast tanks and a system of weights submerged the craft; it could travel at a speed of four miles an hour, powered by eight men cranking its propeller. Its armament consisted of a "torpedo," filled with 90 pounds (41 kilograms) of gunpowder, towed behind the submarine at the end of a 200-foot line. The "Hunley" was to dive under an enemy warship and drag the torpedo against its hull. After a successful test against a barge, the "Hunley" was moved by railroad to Charleston, South Carolina. There the vessel suffered a number of disasters, sinking three times and drowning a number of crewmen including Hunley himself. Manned for a fourth time, the "Hunley" was fitted with a "torpedo" on the end of a long spar and made several successful dives. On the night of February 17, 1864, the submarine attacked the Union warship "Housatonic" in Charleston harbour. The torpedo's detonation exploded the warship's magazines: the "Housatonic" sank in shallow water with the loss of five men, but the "Hunley" also was destroyed and her crew killed.

One of the more intrepid submarine inventors of the same period was Wilhelm Bauer, a noncommissioned officer of Bavarian artillery, who built two boats, "Le Plongeur-Marin" (1851) and "Le Diable-Marin" (1855). The first boat sank in Kiel harbour on February 1, 1851, but Bauer and his two assistants escaped successfully from a depth of 60 ft (18 m) after the craft had been on the bottom for five hours. His second craft, built for the Russian government, was successful and reportedly made 134 dives before being lost at sea. In September 1856, during the coronation of Tsar Alexander II, Bauer submerged his submarine in Kronstadt harbour with several musicians on board. An underwater rendition of the Russian national anthem was clearly heard by persons inside ships in the harbour.

Advent of the machine-powered submarine. A major limitation of these early submarines was their lack of a suitable means of propulsion. In 1880 an English clergyman, the Rev. George W. Garrett, successfully operated a submarine with steam from a coal-fired boiler that featured a retractable smokestack. The fire had to be extinguished before the craft would submerge (or it would exhaust the air in the submarine), but enough steam remained in the boilers for travelling several miles underwater.

Similarly, Swedish gun designer Torsten Nordenfelt constructed a steam-powered submarine driven by twin propellers. His craft could be submerged by vertical propellers to a depth of 50 ft (15 m) and was fitted with one of the first practical torpedo tubes. Several nations built submarines to Nordenfelt's design.

In an effort to overcome the problems of propulsion, two French naval officers built the 146-ft (45-m) submarine "Le Plongeur" in 1864, powered by an 80-horsepower (hp) compressed-air engine, but the craft quickly exhausted its air tanks whenever it got underway. Development of the electric motor finally made electric propulsion practicable. The submarine "Nautilus," built in 1886 by two Englishmen, was an all-electric craft. This "Nautilus" was propelled on the surface at a speed of six knots by two 50-hp electric motors operated from a 100-cell storage battery. But the battery had to be recharged and overhauled at short intervals, and the craft was never able to travel more than 80 miles (130 km) without a battery recharge. In France, Gustave Zédé launched the "Gymnote" in 1888; it, too, was propelled by an electric motor and was extremely manoeuvrable but tended to go out of control when it dived.

The end of the 19th century was a period of intensive submarine development, and Zédé collaborated in a number of designs sponsored by the French Navy. A most successful French undersea craft of the period was the "Narval," designed by Maxime Laubeuf, a marine engineer in the Navy (Figure 2). Launched in 1899, the

The French "Narval"

By courtesy of Henri Le Masson

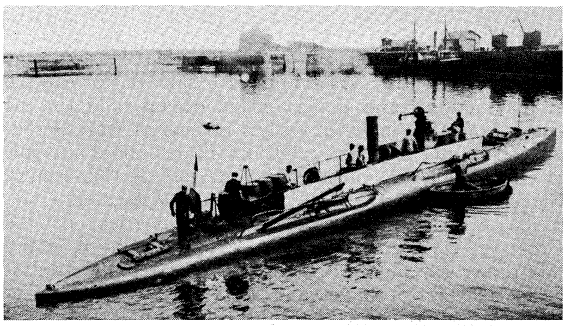


Figure 2: French "Narval" submarine, *c.* 1905.

"Narval" was a double-hulled craft, 111½ ft (34 m) long, propelled on the surface by a steam engine and by electric motors when submerged. The ballast tanks were located between the double hulls, a concept still in use today. The "Narval" made a large number of successful dives. To demonstrate her capabilities, the "Narval" once cruised for 48 hours in the English Channel during a simulated combat mission, then forced the narrow and carefully watched entrance of the roadstead at Brest and fired her four externally carried torpedoes at targets. French progress in submarines was marked by the four Sirène class steam-driven undersea craft completed in 1900–1901 and the "Aigrette," completed in 1905, the first diesel-driven submarine of any navy.

Similarly, there were submarine successes in the United States by rival inventors John P. Holland (an Irish immigrant) and Simon Lake. Holland launched his first undersea craft in 1875. This one and its successors were significant in combining water ballast with horizontal rudders for diving. In 1895, in competition with Nordenfelt, Holland received an order from the U.S. Navy for a submarine. This was to be the "Plunger," propelled by steam on the surface and by electricity when submerged. The craft underwent many design changes and finally was abandoned before completion. Holland returned the funds advanced by the navy and built his next submarine (his sixth) at his own expense. This was the "Holland," a 53¼-ft (16-m) craft launched in 1897 and accepted by the navy in 1900. For underwater propulsion the "Holland" had an electric motor and was propelled on the surface by a gasoline engine. The submarine's armament consisted of a bow torpedo tube, for which three torpedoes were carried, and two dynamite guns. The guns, one pointing forward and one aft, were built into the hull and aimed by pointing the submarine directly at or away from the target. With its nine-man crew the "Holland" was a successful boat and was modified many times to test different arrangements of propellers, diving planes, rudders, and other equipment.

Holland's chief competitor, Simon Lake, built his first submarine, the "Argonaut I," in 1894, powered by a gasoline engine and electric motor. This and Lake's other early boats were fitted with wheels for the submarine to roll over the ocean floor. He envisioned submarines sending out divers to cut cables, destroy enemy mines, and telephone enemy ship movements back to the submarine; in peacetime, such submarine-diver operations could be used for sea-floor mineral mining. In 1898 the "Argonaut I" sailed from Norfolk, Virginia, to New York under its

own power, predating the cruises of the French "Narval" and marking the first time an undersea craft operated extensively in the open sea. Lake's second submarine was the "Protector," launched in 1901. After the U.S. Navy rejected this craft, Lake offered the "Protector" to Russia and Japan. Tsarist Russia bought the vessel and five more like it. Simultaneously, Japan bought several Holland boats, but neither side used them during the Russo-Japanese conflict (1904–1905).

Of the major naval powers at the turn of the century, only Britain remained indifferent toward submarines. Finally, in 1901, the Royal Navy ordered five of the Holland-design undersea craft. Germany completed its first submarine, the *Unterseeboot* No. 1, in 1905. This craft was 139 ft (42 m) long, powered on the surface by a heavy oil engine and by an electric motor submerged, and was armed with one torpedo tube. Thus, the stage was set for the 20th-century submarine, a craft propelled on the surface by diesel engines and underwater by electric motors, submerging by diving planes and taking on water ballast, and armed with torpedoes for sinking enemy ships. The quarters inside these early craft were cramped, generally wet, and stank from diesel oil.

World War I. By the eve of World War I all of the major navies included submarines in their fleets, but these crafts were relatively small, were considered of questionable military value, and generally were intended for coastal operations. The most significant exception to the concept of coastal activity was the German Deutschland class of merchant U-boats, each 315 ft (95 m) long with two large cargo compartments. These submarines could carry 700 tons of cargo at 12- to 13-knot speeds on the surface and at seven knots submerged. Manned by a crew of 8 officers and 21 enlisted men under the command of Capt. Paul Koenig, a veteran merchant-ship skipper, the "Deutschland" herself made two successful cargo trips between Germany and the United States in 1916 despite a British naval blockade. The "Deutschland" became the "U-155" when fitted with torpedo tubes and deck guns and with seven similar submarines served in a combat role during the latter stages of the war. In comparison, the "standard" submarine of World War I measured slightly over 200 ft (60 m) in length and displaced less than 1,000 tons on the surface.

The prewar submarines generally had been armed with self-propelled torpedoes for attacking enemy ships. During the war submarines also were fitted with deck guns. This permitted them to approach enemy merchant ships on the surface and signal them to stop for searching (an early war policy) and later to sink small or unarmed ships that did not warrant expenditure of torpedoes. Most war-built submarines had one and sometimes two guns of about three- or four-inch calibre (diameter of the projectile); however, several later German submarines carried 5.9-inch guns (including the Deutschland class in military configuration).

Late in the war the British completed the first of three M class submarines, each armed with a 12-inch gun (in addition to four torpedo tubes and an anti-aircraft gun). These submarine monitors were intended to steam off an enemy coast with their decks submerged and only the muzzle of the 12-inch gun and a sighting periscope protruding through the water. Capable of firing far inland, these giant submarines were 296 ft (90 m) long, displaced 1,650 tons on the surface, and had diesel-electric propulsion that could drive them at 14 knots on the surface and 8 knots submerged.

Another armament variation was the submarine modified to lay mines during covert missions off an enemy's harbours. The Germans constructed several specialized submarines with vertical mine tubes through their hulls; some U-boats carried 48 mines in addition to their torpedoes.

Anti-submarine submarines

Also noteworthy was the development, during the war, of the concept of an antisubmarine submarine. British submarines sank 17 German U-boats during the conflict; the early submarine-versus-submarine successes led to British development of the R class submarine intended specifically for this role. These were relatively small craft, 163 ft (50 m) long and displacing 410 tons on the surface, with only one propeller (most contemporary submarines had two). Diesel engines could drive them at nine knots on the surface, but once submerged, large batteries permitted their electric motors to drive them at the high speed of 15 knots for two hours. Thus, they were both manoeuvrable and fast. Advanced underwater listening equipment (asdic/sonar) was installed, and six forward torpedo tubes made them potent weapons. Although these submarines appeared too late to have any actual effect on the war, they pioneered a new concept in the development of the submarine.

All World War I-era submarines were propelled by diesels on the surface and by electric motors submerged, except for the British Swordfish and K class. These submarines, intended to operate as scouts for surface warships, required the high speeds then available only from steam turbines. The Swordfish was of conventional size, but the K boats were 340 ft (100 m) long and displaced 1,780 tons on the surface; the steam plant of the K boats propelled them at 23½ knots on the surface, while electric motors gave them a ten-knot submerged speed, a common underwater speed for submarines until after World War II.

Interest in submarines continued high within the world's navies during the period between World Wars I and II. A particularly attractive notion was that of developing large, heavily gunned submarines for commerce raiding. The British "X-1," completed in 1925, was a long-range submarine with four 5.2-inch guns in twin, armoured turrets, in addition to six torpedo tubes. The French "Surcouf," completed in 1934, mounted two eight-inch guns in a single turret, in addition to ten torpedo tubes, and also had a small hangar to accommodate a seaplane. The aircraft was to be used to search for merchant ship victims. The seaplane–submarine team also was investigated by the Royal Navy, which removed the 12-inch gun from the submarine monitor "M-2"—the 12-inch gun being unsuitable for use against surface ships—and fitted her to carry a seaplane. The U.S. Navy similarly fitted the submarine "S-1" with a hangar and small seaplane, while the Japanese Navy modified a number of submarines to operate small seaplanes.

Commerce raiding

During this period the U.S. Navy built its first large, long-range submarine, the "Argonaut." Completed in 1928, she was 381 ft (116 m) long, displaced 2,710 tons on the surface, was armed with two six-inch guns and four forward torpedo tubes, and could carry 60 mines. The "Argonaut," the largest non-nuclear submarine ever built by the U.S. Navy, led to the highly successful Gato and Balao classes used in World War II.

During the 1930s the rejuvenated Soviet shipyards began producing large numbers of submarines, primarily coastal craft, in an attempt to make the U.S.S.R. a sea power without major expenditures for surface warships. But though the Soviet program achieved quantity, their ships were unsuitable for operations against the German Navy, their crews were poorly trained, and Soviet bases were blocked by ice much of the time.

World War II. World War II (1939–1945) saw extensive submarine campaigns on all of the world's oceans. In the Atlantic the principal German U-boat was the VII type, a relatively small but effective craft when properly employed. The Type VIIC variant was 220¼ ft (67 m) long, displaced 769 tons on the surface, and was powered by diesel-electric machinery at a speed of 17 knots on the surface and 7½ knots submerged. Armament consisted of one 3.5-inch deck gun, various anti-aircraft guns, and five torpedo tubes, four forward and one aft. Either 14 torpedoes or 14 tube-launched mines were carried. Manned by a crew of 44, these submarines had a surface endurance of 6,500 miles at 12 knots, but, when they were submerged, their batteries would remain active a little less than a day at four knots.

Three German designs warrant special mention: The VIIF design was an elongated boat 254¾ ft (78 m) long intended to replenish other U-boats at sea with torpedoes —25 spares were carried, above a normal complement of 14—and fuel oil—199 tons carried, compared with 114

tons in Type VII submarine. The ultimate diesel-electric submarine evolved in the war was the German Type XXI, a 250-ft (75-m), 1,600-ton craft that could attain 17½ knots submerged for more than an hour, could travel at six knots underwater for two days, or could "creep" at slower speeds for four days. These submarines were fitted with *schnorchel* devices (see below) that made it unnecessary for them to surface fully to recharge their batteries after operating submerged. The Type XXI had an operating depth of 850 ft (260 m), more than twice what was then normal, and was armed with four 33-millimetre guns and six forward torpedo tubes (23 torpedoes carried). Existing Allied antisubmarine forces would have had serious trouble coping with these craft had the war continued past the spring of 1945.

A final German war design of particular interest was the Walter turbine propulsion plant. The need for oxygen for combustion had previously prevented the use of steam turbines or diesels while the submarine was submerged, and air was at a premium. Hellmuth Walter, a German scientist, developed a turbine propulsion system using oxygen generated by hydrogen peroxide to operate the turbine while submerged. A simplified submarine, the "V-80," built in 1940 and propelled by a Walter turbine system, could attain speeds of more than 26 knots submerged for a short period of time. After many delays, the first Walter-propelled Type XVII combat submarines were completed and could reach 25 knots underwater for brief periods, and a submerged run of 20 knots for 5½ hours was achieved on trials. But these submarines, like the Type XXI, were not ready for full-scale operations when the war ended.

A notable German submarine development of World War II was the *schnorchel* device (anglicized by the U.S. Navy to "snorkel"). Its invention is credited to a Dutch officer, Lt. Jan J. Wichers, who in 1933 advanced the idea of a breathing tube to supply fresh air to a submarine while running submerged. (Robert Fulton had a similar device in his "Nautilus" of 1801.) The Netherlands Navy began using snorkels in 1936, and some fell into German hands in 1940. With the advent of radar to detect surfaced submarines, the Germans fitted hundreds of U-boats with snorkels to permit operation of diesels at periscope depth to recharge batteries for underwater propulsion with less of a possibility of detection by Allied radar-equipped ships and aircraft.

Japanese submarines

In the Pacific War the Japanese employed a large number of submarines of various sizes and types, including submarine-carried aircraft, midget submarines, and "human torpedoes" carried by larger submarines. Two Japanese designs are of special interest. The Japanese I-201 class was a high-speed submarine 259⅙ ft (79 m) long, displacing 1,291 tons, that had diesel propulsion for 15 knots on the surface; while underwater, large batteries and electric motors could drive the vessel at a speed of 19 knots for almost one hour. Each boat had two 25-millimetre guns and four forward torpedo tubes and carried ten torpedoes. The Japanese, who constructed one-man submarines, also built the largest conventionally powered undersea craft ever built, the I-400 class. These submarines were intended as underwater aircraft carriers and had a hangar that could accommodate three floatplanes and parts for a fourth. The I-400 submarines were 400¼ ft (122 m) long and displaced 5,700 tons, about the size of a light cruiser. The aircraft were housed in a hangar 12 ft (3.7 m) in diameter and 102 ft (31 m) long and launched by a catapult built into the deck. Each submarine was armed with a 140-millimetre deck gun plus seven 25-millimetre anti-aircraft guns, eight torpedo tubes, and 20 torpedoes. (They were never actually used for their intended purpose, which was to launch planes against the Panama Canal, but were employed as cargo submarines to supply bypassed island garrisons in the Pacific.)

The highly successful U.S. submarine campaign in the Pacific war was waged mainly with the Gato and Balao class submarines. These were approximately 311½ ft (94.9 m) long, displaced 1,525 tons, and had diesel-electric machinery for 20-knot surface and 9-knot underwater speeds. The principal difference between the two designs was the 300-ft (90-m) operating depth for the Gato class and 400-ft (120-m) depth for the Balao boats. Manned by 65 to 70, these submarines had one or two five-inch deck guns plus smaller anti-aircraft weapons, ten torpedo tubes (six forward, four aft), and carried 24 torpedoes.

After the war the Allies were quick to adopt advanced German submarine technology. The British built two peroxide turbine-propelled experimental submarines, the "Explorer" and "Excalibur," but this concept lost favour because of the unstable properties of hydrogen peroxide and because of American success with nuclear propulsion. The U.S.S.R. began building a modification of the Type XXI submarine. Some 235 of these W class submarines were completed between 1950 and 1958, more submarines than built by all of the world's other navies combined between 1945 and 1970. (In that period Soviet shipyards produced a total of 560 new submarines.)

In the postwar period the U.S. Navy studied German technology and converted 52 war-built submarines to the Guppy configuration (an acronym for Greater Underwater Propulsion Project with the "y" added for phonetics). These submarines had their deck guns removed and streamlined conning towers fitted; larger batteries and a snorkel were installed; four torpedoes and, in some craft, one of the four diesel engines were removed. The result was an underwater speed of 15 knots and increased underwater endurance. During the postwar years the U.S. Navy built 21 diesel-powered submarines, several of which were highly specialized craft for antisubmarine warfare, guided-missile launching, underwater research, training, and radar picket operations in support of surface ships. Several war-built submarines were converted to these roles and also were used for the transport of Marine reconnaissance troops and frogmen.

Nuclear propulsion. The first efforts to develop a submarine nuclear power plant were made at the U.S. Naval Research Laboratory in Washington, D.C., under the direction of Ross Gunn, beginning in the spring of 1939. A report submitted that summer by Gunn noted that such a power plant would not require oxygen to operate, "a tremendous military advantage [which] would enormously increase the range and military effectiveness of a submarine." All work on nuclear propulsion was halted, however, as efforts in this field were devoted toward the atomic bomb development. After the war Gunn and Philip Abelson of the Carnegie Institution resumed work on a nuclear power plant, resulting in a report in the spring of 1946 calling for a nuclear submarine. The U.S. Navy took no action in this regard until Capt. Hyman G. Rickover, appointed to head the navy's cooperation in a nuclear project at the Oak Ridge, Tenn., research centre became a crusader for this type of submarine propulsion. Subsequently, a land-based prototype submarine power plant was constructed, followed by installation of a duplicate plant in the USS "Nautilus."

The USS "Nautilus"

The "Nautilus" got underway on January 17, 1955. Resembling a conventional submarine in appearance, it is 320 ft (98 m) long, displaces 3,530 tons, and has six forward torpedo tubes. Her top speed of 20 knots, both surface and submerged, was good but not spectacular compared with her diesel-electric contemporaries. But her underwater range at high speed is virtually unlimited. Measured in terms of "years" rather than miles, this endurance made her a revolutionary ship, in fact the first true submarine. All of her predecessors had been merely surface craft capable of brief operation under water.

Seven additional nuclear submarines were built to conventional hull designs, including one for launching guided missiles and one for radar picket missions. The navy then combined the advanced "tear drop" hull design of the research submarine Albacore with a nuclear propulsion plant to produce the Skipjack class of very fast submarines, capable of underwater speeds estimated at over 30 knots. This design has been incorporated in the subsequent Thresher and Sturgeon classes. These later nuclear submarines, however, have their large bows devoted entirely to detection equipment (sonar) and have their four torpedo tubes amidships, angled out with two on each side (a concept tested by the Germans in World War II).

This places their sonar at the most advantageous position, as far forward as possible and away from the ship's machinery and propeller noises. The Sturgeon class submarines are 292 ft (89 m) long and displace 3,860 tons; however, they probably are slower than the earlier Skipjack class because they have the same propulsion machinery but larger hulls.

Almost all U.S. nuclear submarines are primarily intended to detect and attack enemy submarines, except for those armed with strategic Polaris or Poseidon missiles. Beginning with the "George Washington," the U.S. Navy completed 41 Polaris-armed submarines between 1960 and 1967. Of these, 31 are Lafayette class submarines that are to be converted to fire the improved Poseidon missile. The Lafayette class submarines, each carrying 16 missiles plus four torpedo tubes, are 425 ft (130 m) long and displace 6,650 tons. The only larger nuclear submarine is the "Triton" completed in 1959; she is a radar picket craft 447½ ft (136 m) long and displacing 5,940 tons. The "Triton" also was the first submarine propelled by two nuclear reactors; the Soviet Yankee class missile submarines also may have two reactors, according to some U.S. Navy estimates, but all other nuclear submarines have only one.

Soviet nuclear submarines

The U.S.S.R., maintaining the largest and newest submarine fleet in the post-World War II era, reached a maximum strength of almost 500 diesel-electric submarines in the early 1960s. The first Soviet nuclear submarine, designated November class, was completed in 1959 or 1960 and was 360 ft (110 m) long and displaced about 3,500 tons. The November class submarines generally are credited with a surface speed of about 25 knots and a submerged speed of 25 to 30 knots, with some of the later November submarines possibly exceeding 30 knots submerged. Armament consists of six bow torpedo tubes.

During the 1960s the U.S.S.R. produced large numbers of diesel-electric and nuclear submarines for three distinct missions: attack or torpedo-armed submarines to attack surface ships and other submarines; cruise missile-armed submarines to attack enemy surface warships, especially aircraft carriers; and ballistic-missile submarines armed with long-range strategic missiles. By 1971 the U.S. and Soviet navies each had built some 90 nuclear-powered undersea craft and had a rate of production of just over four atomic submarines per year for the United States and 12 to 14 per year for the Soviet Union.

The first British nuclear submarine was the "Dreadnought," completed in 1963, a torpedo-armed submarine of advanced design and fitted with an American propulsion system. Subsequent British nuclear submarines, attack and ballistic missile types, have British-designed and fabricated machinery. The first French nuclear submarine was to have been the hull designated "Q-244" begun in 1958; however, the project was cancelled, and the hull subsequently completed as the diesel-electric missile trials submarine "Gymnote." Subsequently the French built four nuclear-powered, ballistic-missile submarines, the first being "Le Redoutable," completed in 1969. During the 1960s the Soviet, British, and French navies continued constructing advanced diesel-electric submarines. In many roles the diesel-electric submarine can perform equally to a nuclear craft, especially when very quiet operation for limited periods is required. Also significant is the higher cost of nuclear propulsion; the European nations can sell conventional submarines to lesser navies that cannot afford nuclear craft.

The overall superiority of nuclear-propelled submarines cannot be questioned. Only this form makes possible the relatively invulnerable ballistic-missile submarines that can cruise the oceans for months without surfacing or extending an air-intake tube to the surface. Nuclear submarine performance can be measured by such exploits as: "Nautilus" (U.S.), making the first submerged polar transit from Point Barrow, Alaska, to the Greenland Sea, travelling 1,830 miles (2,945 km) under the polar ice pack (August 1–5, 1958); "Skate" (U.S.) (Figure 3), a similar polar-icepack voyage (August 9–20, 1958), travelling 2,405 mi (3,870 km) and surfacing 40 mi (64 km) from the North Pole; "Seawolf" (U.S.), remaining submerged

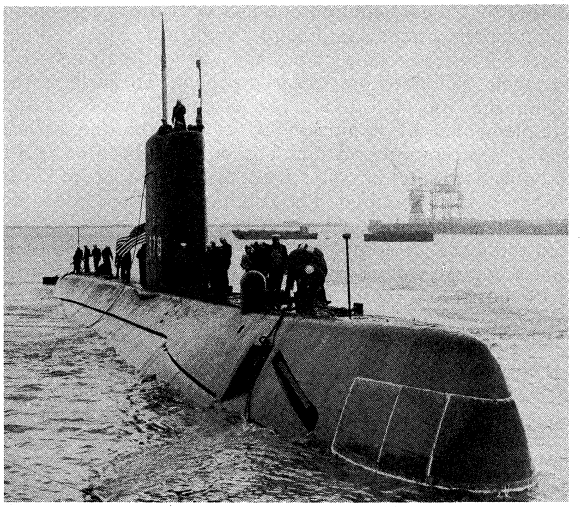


Figure 3: USS "Skate," a nuclear-powered submarine, arriving in Portland, Dorset.
Wide World Photos

for 60 consecutive days and travelling 13,761 mi (22,146 km; August 5–October 6, 1958); "Triton" (U.S.), circumnavigating the world submerged, following roughly the route taken by Magellan (1519–1592), and travelling 36,014 mi (57,946 km) February 24–May 10, 1960); "Leninsky Komsomol" (U.S.S.R.), reportedly operating under the polar ice pack and surfacing near the North Pole in 1962; a November class submarine (U.S.S.R.), remaining submerged for almost 50 consecutive days while steaming about 25,000 mi (40,000 km). None of these feats is considered possible with conventionally propelled submarines.

Like their conventional sisters, nuclear submarines have encountered difficulties in the hostile depths. To the list of submarines sunk accidentally in both peace and war, by 1971 three nuclear submarines had been added: the USS "Thresher," off New England, with 129 lives lost, April 10, 1963; USS "Scorpion," off the Azores, with 99 lives lost, May 1968; and a Soviet November class, in the eastern Atlantic, April 12, 1970 (no loss of life).

MODERN STRUCTURE AND USES

Construction features. To operate underwater a submarine has a strengthened steel pressure hull that provides protection from water pressure to the crew, machinery, and certain equipment. A double-hull technique is generally used, with the space between the inner (pressure) hull and outer hull providing room for water ballast and fuel tanks. This double-hull feature also provides additional protection for the pressure hull from nearby detonations of antisubmarine weapons. Until recently, above the hull and extending the length of the submarine there was a nonwatertight, free-flooding superstructure that formed the main deck for use when the submarine was surfaced. Modern high-speed submarines, both diesel-electric and nuclear, do not have the superstructure deck, as it interferes with optimum underwater shape. Also, nuclear-powered submarines do not have the normal double-hull configuration because they require only small amounts of diesel fuel for their auxiliary machinery; hence, these fuel tanks and water-ballast tanks are fitted within the pressure hull and a partial double hull.

Above the submarine's hull is a conning tower or sail structure. In older submarines this was a watertight conning tower that served as the command centre during combat. On post-World War II submarines there is a streamlined "sail" structure that does not contain a watertight compartment. The conning tower and sail structure both house the retracting periscopes, radar and radio antennas, and snorkel tubes and provide a platform or bridge for the navigating officers and lookouts when the submarine is on the surface.

Normally, the submarine submerges by flooding the water-ballast tanks to attain a neutral buoyancy and then

diving by forward motion coupled with the use of small diving planes, somewhat like rotating wings. Most submarines have stern diving planes and a second pair installed on their sail structure or at the bow. To surface, the submarine uses the diving planes and forward motion, supplemented by expulsion of water from the ballast tanks with compressed air. The compressed air can be produced within the submarine and stored in tanks or "bottles" until required. The submarine is kept in a state of neutral buoyancy with water-trim tanks that are partially flooded to compensate for changes in the submarine's volume caused by increasing pressure as the craft goes deeper and by the effects of changing water temperature and salinity or by weight changes from firing torpedoes or missiles.

Control room

The control room in a modern submarine is carefully designed to achieve the maximum efficiency during operations, with centralized command positions from which the submarine and its weapons can be directed. The modern submarine relies more on its sonar than on the periscope for detecting and attacking an enemy; in addition, the modern submarine has a complex inertial-navigation system that provides information on its position independent of any outside sources except gravity and earth rotation. These data are periodically checked by a special star-tracking periscope device or radio navigation antenna that is projected above the water for very brief periods. The actual control of the submarine is achieved by aircraft-line controls, with one or two men sitting in "pilot" chairs and manoeuvring the craft. Some undersea craft have analog "pictures" of a simulated roadway, presented to the "pilots" on a television-like screen, in addition to the conventional dials and pointers that indicate course, speed, and depth with numbers.

The modern submarine is a relatively comfortable craft, dry inside and fully air-conditioned, the latter feature for both crew habitability and for the great amount of sophisticated electronic equipment that requires carefully maintained humidity and temperature conditions. Modern, especially designed electric kitchens provide food that is traditionally considered of high quality. The crew sleeping or dining spaces also are used for motion pictures and as game rooms. The officers have their own sleeping area and wardroom.

Oceangoing submarines can remain at sea for two or three months, carrying all provisions required by the crew during that period. Diesel-electric submarines periodically raise a snorkel device to draw in air for the diesel engines and crew and to carry away exhaust gases. Nuclear-powered submarines need not draw in air for their engines and can manufacture oxygen for the crew from the surrounding seawater to supplement oxygen carried in tanks.

Two-way communication with submerged submarines is difficult because radio waves cannot be produced underwater. Thus, the modern submarine cannot send radio messages while fully submerged but must raise an antenna mast up into the air to communicate by radio. The submarine can receive radio transmissions while submerged, however, through built-in antennas or by trailing a wire antenna through the water.

Sensors. The prime sensors of a submarine are sonars, periscopes, and radars (see also SONAR; RADAR). Submarines employ both passive sonar (sound detection and ranging), which listens for the noises generated by another submarine or a surface ship, and active sonar, which sends out an acoustic pulse that is reflected upon hitting another submarine or ship. The submarine uses sonar for detection of an enemy, for sea-floor and under-ice navigation, and for directing torpedoes.

The periscope

The periscope is the "eye" of the submarine. Early submarines had small glass portholes for direct observation. This required the submarine to be partially surfaced. In 1854 E-H. Marié-Davy of France designed a submarine sight tube containing two mirrors, one above the other, held at a 45° angle and facing in opposite directions. The mirror-tube provided some degree of sight from a submerged submarine but was faulty at best, and in 1872 prisms were substituted for mirrors.

Subsequently, the prism periscope was improved and made to telescope, so that it could be housed within the submarine's hull when not in use and extended above the submarine for reconnaissance. Modern periscopes are about 40 ft (12 m) long and taper to about one inch in diameter at the top to reduce detection from the surface. Special features permit considerable magnification, night reconnaissance, photography through periscopes, sextant use for celestial navigation while submerged, and viewing of the air around the submarine as well as the surface.

Radar (radio detection and ranging) was fitted to most U.S., British, and German submarines during World War II, initially to detect aircraft while the submarine was operating at periscope depth or on the surface. Subsequently, surface-search radar was provided, and during the latter part of the war radar-approach attacks against surface ships were common. Submarine radars are now used primarily for navigation in coastal waters. The antenna is mounted on a retractable mast and is withdrawn into the sail structure when not in use.

Propulsion. Shortly after the beginning of the 20th century, most submarines were fitted with diesel engines for surface propulsion and to recharge the batteries needed for electric-motor submerged propulsion. (The steam-turbine and peroxide-turbine variants are discussed above.) Engines burning diesel oil had a significant advantage over gasoline motors for surface propulsion because of the highly volatile properties of gasoline, and advantages over steam propulsion because of the high temperatures generated by the boilers necessary for steam turbines before the advent of nuclear propulsion.

Once submerged, the diesel engines are shut down, and electric motors are started, turning the same propeller shafts. The electric motors, drawing their energy from storage batteries, require no combustion and can be used without regard for the submarine's air supply. Electric motors have the additional advantage of being quiet, an important factor in submarine operations.

The nuclear power plant

The submarine nuclear power plant consists of an atomic reactor that generates enormous amounts of heat through nuclear fission (see also NUCLEAR REACTOR). The reactor is controlled by inserting and withdrawing control rods that penetrate the reactor core. When the rods are completely inserted the reaction ceases, and when they are completely withdrawn the reaction is at its maximum. The heat is carried by the primary coolant (pressurized water) to a steam generator, where the heat is transferred to a secondary water system that creates steam to drive the submarine's turbines. The primary coolant water is radioactive from passing near the radioactive core and thus cannot be used directly with the turbine. On most nuclear submarines reduction gears are used between the turbines and the propeller shaft; however, a few incorporate turbo-electric drive (notably the USS "Tullibee" and USS "SSN-685"). Here the steam turbine drives an electric generator, which in turn drives the electric motors coupled to the propeller shaft.

Nuclear-powered submarines are "refuelled" by replacement of the uranium core in their nuclear reactor. The "Nautilus" steamed 62,562 mi (100,681 km) on her first nuclear core, which was replaced in 1957; 91,324 mi (146,968 km) on her second core, which was replaced in 1959; and approximately 150,000 mi (240,000 km) on her third core, replaced in 1964. The cores now being installed in U.S. nuclear submarines will have an endurance of approximately 400,000 mi (640,000 km).

All nuclear-powered submarines are fitted with auxiliary diesel-electric machinery (and snorkel) for use in the event of a main machinery casualty and to generate power when the main machinery is shut down while in port.

Weapons. The primary weapon of the submarine traditionally has been the self-propelled torpedo. Modern submarines are fitted with internal torpedo-launching tubes 21 inches (53 centimetres) in diameter and about 21 ft (6.4 m) long. The tubes are horizontal in the bow or amidships, with some submarines also having stern tubes. The tubes are loaded from within the submarines and normally are loaded with reloads kept within the submarine when the craft goes to sea. Each tube has doors at both ends so that the inner door can be opened when the outer

door is closed and vice versa, permitting firing and reloading without water entering the submarine. The tubes are fired electrically or mechanically by hand.

Most torpedoes today are intended for use against other submarines; they either home on their target acoustically with self-contained sonar or are guided by electronic commands passed to them through a threadlike wire paid out behind the torpedo as it travels away from the launching submarine. Another submarine-to-submarine weapon is the U.S. Navy's submarine rocket (Subroc). Fired from a torpedo tube while the craft is submerged, Subroc streaks to the surface, leaves the water, and, propelled by a rocket, travels through the air on a ballistic trajectory. The weapon re-enters the water an estimated 25 to 30 mi (40 to 48 km) from the launching submarine, and a nuclear warhead detonates.

During World War II, the German Navy experimented with small guided missiles and large ballistic missiles launched from submarines and submarine-towed containers, respectively. After the war, the U.S. and Soviet navies developed guided cruise missiles, jet- or rocket-propelled vehicles with stub wings that resembled pilotless aircraft. These were carried in watertight hangars on the submarine deck and fired after the submarine surfaced. The U.S. Navy operated the Regulus I missile with a 500-mi (800-km) range from several submarines between 1958 and 1964 and developed but did not use the Regulus II missile with a range of 1,000 mi (1,600 km). The Soviets developed similar missiles and now use the Shaddock missile with a maximum range of 450 mi (720 km) from a number of conventional and nuclear submarines. These missiles, which have a variable flight path, are considered the most potent threat to surface warships. (Some newer Soviet submarine-launched cruise missiles may be capable of underwater launching.)

Submarine-launched ballistic missiles are strategic weapons intended for use against cities and other fixed land targets. They cannot be effectively used against ships. The Soviet Navy began deploying conventional submarines with short-range ballistic missiles in the late 1950s, while the first U.S. Navy strategic missile submarine, the "George Washington," was first deployed in 1960; it carried 16 Polaris A-1 missiles, each with a nuclear warhead and a range of 1,380 mi (2,220 km). These and subsequent ballistic missiles can be fired when the submarine is submerged. Subsequent U.S. Polaris submarines were fitted with the A-2 missile with a range of 1,725 mi (2,775 km) and then the A-3 missile with a 2,875-mi (4,625-km) range. Some A-3 missiles also have three-vehicle warheads that can "shotgun" three nuclear bombs at the target. In 1970 the U.S. Navy began to rearm most Polaris submarines with the Poseidon missile, which has about the range of the A-3 missile but which carries ten small warheads that can be aimed at individual targets. Similarly, British strategic missile submarines are armed with the Polaris A-3 missile, French submarines with the MSBS missile with a range of 1,900 mi (3,000 km), and Soviet submarines with the 1,300–1,500-mi (2,000–2,400 km) SS-N-6 missile. (A longer range, 2,700–3,000-mi submarine missile, developed by the Soviets, was scheduled for deployment during the 1970s.) All of these strategic submarines carry 16 missiles in addition to torpedoes. The Soviet Navy also operates a number of smaller nuclear and conventional submarines that fire three strategic missiles with lesser range.

The Polaris missile

Submarine Strength
(active units)

	U.S.	Russia, U.S.S.R.	Germany	France	U.K.	Japan	China
1914	30	36	30	67	76	15	—
1919	60	few	0	43	100	20	—
1939	75	170	57	80	55	65	—
1945	250	200	154	10	90	50	—
1955	1* 128†	400†	0	13	53	1	13
1970	90* 63†	90* 265†	12‡	1* 21†	9* 26†	10	30
1975§	110* 40†	140* 200†	36‡	4* 21†	12* 21†	15	25

*Nuclear propelled. †Diesel-electric propelled. ‡Coastal. §Estimate.

BIBLIOGRAPHY. W.R. ANDERSON, *Nautilus 90 North* (1959), and J. CALVERT, *Surface at the Pole* (1960), popular accounts of nuclear submarine operations, by the commanding officers of the "Nautilus" and the "Skate"; E. ROSSLER, *U-Boottyp XXI* (1966), a description of one of the most successful submarine designs of World War II (in German); W. FRANK, *The Sea Wolves* (1955), a discussion of the political, material, and operational aspects of German submarines in World War II; M. HASHIMOTO, *Sunk!* (1954), an account of Japanese submarines and their activities during World War II; A.R. HEZLET, *The Submarine and Sea Power* (1967), a comprehensive look at submarines and their influence on warfare by a leading British naval strategist; W.J. HOLMES, *Undersea Victory* (1966), an excellent study of U.S. submarine operations in the Pacific during World War II; *Jane's Fighting Ships* (annual), dimensions and characteristics of most of the world's submarines; W.S. JAMESON, *The Most Formidable Thing* (1965), a British naval officer's analysis of submarine development and operations during World War I; P.K. KEMP, *H.M. Submarines* (1952), an official treatment of British submarine operations; H. LE MASSON, *Du Nautilus (1800) au Redoutable* (1969), the story of submarine development in France, written by a leading warship authority; H.T. LENTON, *German Submarines*, 2 vol. (1965), detailed dimensions and particulars of German U-boats of the World War II era; F.W. LIPSCOMB, *The British Submarine* (1954), British submarine development and operations; S. LAKE, *Submarine* (1938), an interesting autobiography of a pioneer designer providing considerable insight into early approaches to submarine development; R.K. MORRIS, *John P. Holland* (1966), a biography of a pioneer submarine designer, with considerable data on the politics of early submarine development; N. POLMAR, *Atomic Submarines* (1963), a detailed description of nuclear-powered submarine development from the late 1930s to the loss of the submarine "Thresher," and *Death of the Thresher* (1964); D.A. THOMAS, *Submarine Victory* (1961), a description of the contribution of British submarines to the Allied triumph in World War II.

(N.C.P.)

Subsidies

A subsidy is a direct or indirect governmental payment, economic concession, or special privilege granted to private firms, households, or other governmental units in order to promote a public objective. Identification of a subsidy is generally complicated because of the amorphous variety of subsidy instruments, the multiplicity of the public objectives that they are designed to serve, and the complexity of their effects.

Varieties and functions of subsidies

Subsidies to transportation, housing, agriculture, mining, and other industries have been instituted on the grounds that preservation or expansion of these industries, even at a cost to the general public, is in the public interest. Subsidies to the arts, sciences, humanities, and religion have also been instituted in many nations because of the inability of the private economy to support these functions at a level consistent with public policy. The term also comprehends grants of money or other aid made by a central government to a local one to promote objectives in which the central government has an interest (*e.g.*, grants-in-aid). More broadly defined, subsidies include welfare payments designed to ameliorate inequalities in the distribution of income and also other governmental programs designed to mitigate the effects of market forces. Regardless of the form that subsidies take, their purpose is to alter the results created by otherwise free markets and unimpeded competition in a direction considered more consistent with the objectives of public policy. The effect of subsidies is to encourage the growth of subsidized industries relative to industries that do not receive subsidies and thus to alter the uses to which an economy puts its resources.

Subsidies have a long history in all nations and were extensively employed by governments during the mercantilist period preceding the Industrial Revolution when it was thought that the accumulation of gold through a favourable balance of trade required the protection of domestic manufacturers. Such protectionist doctrines have often been viewed with skepticism since the British economist Adam Smith published *Wealth of Nations* in

1776, refuting mercantilist doctrine and establishing a modern theoretical basis for the abandonment of government protection in favour of unimpeded trade and competition. Nevertheless, protectionism continues as a part of national economic policy in most nations of the world, which try to justify subsidy programs on the grounds of a need for national self-sufficiency, for the development of emerging industries, for the aid of small businesses, for the encouragement of home ownership, for the elimination of slums, for the preservation of local government, for the alleviation of depression and unemployment, for the promotion of religion, marriage, and other cherished institutions, and so on. Subsidy as a distinct tool of public policy has somewhat more significance in nations, such as the United States, that rely primarily on a competitive market to perform basic economic functions and that seek only to induce or nudge the action of such markets. In nations in which a strong central government exerts influence over the price and production policies of domestic industries, the subsidy device is replaced by comprehensive economic planning.

Subsidies are implemented through a variety of financial techniques, such as (1) direct payments in cash or kind, (2) governmental provision of goods or services at prices below the normal market price, (3) governmental purchase of goods or services at prices in excess of the market price, and (4) tax concessions and similar inducements. In addition, there are numerous governmental policies that have subsidy effects, such as regulatory statutes that soften the full force of competition, policies that require the purchase of goods from favoured producers or nations, and protective wage and price legislation. A distinction is sometimes made between direct or visible subsidies (such as direct payments for ship construction and airline operation), which are easy to identify and measure, and indirect or concealed subsidies (such as price ceilings or floors, tariffs, and tax concessions), which are difficult to identify and always difficult to measure. Concealed subsidies are more susceptible to abuse by special-interest groups and politicians and, from an economic point of view, achieve their objectives at a higher real cost to the community. Since the term subsidy often carries a connotation of opprobrium, it is seldom employed in official descriptions; most subsidies are implemented indirectly through government manipulation of the prices received or paid by producers. Narrowly defined and conservatively measured, the subsidy programs of the United States government, for example, amounted to about $10,000,000,000 a year during the 1960s; estimates vary upward from this figure to as high as $30,000,000,000 or $40,000,000,000 depending on the comprehensiveness of the definitions employed.

Direct subsidies. Historically, direct subsidies have been employed most extensively to promote the development of the transportation industries. Ocean shipping and aviation have been viewed as important instruments of defense and foreign policy by all nations, and the development of land transportation has been considered a prerequisite for domestic economic development. Direct maritime subsidies in France, Italy, Japan, the United States, and other nations involve (a) construction-differential subsidies, whereby the government absorbs the excess cost of construction in foreign shipyards, or (b) operating-differential subsidies, which compensate certain operators of ships for higher operating costs than those borne by their foreign competitors.

Transportation subsidies

In no instances are railroads a purely private endeavour. The initial costs of construction are high, the risks are great, and the buildup of traffic and revenue is delayed. When a railroad network is finally completed, it is frequently discovered that many miles of track yield little revenue but that the maintenance of a basic minimum service is socially desirable. As a result, many national railroads are either government owned or considered instruments of public policy and, if necessary, receive direct subsidy payments to cover deficits. The principal types of subsidies include cash contributions, tax concessions, government purchase of railroad securities at favourable prices, and land grants. In recent years, the survival of railroads has been threatened by competition from other carriers. As a result, such countries as Great Britain have inaugurated a system of grants for the operation of passenger trains.

Direct subsidies to the aviation industry include cash payments to recover operating deficits and government financing of the costs of developing new aircraft. Competition between nations, notably Great Britain, France, the United States, and the Soviet Union, has led to huge public expenditures for the development of commercial aircraft.

Direct cash subsidies are also awarded, especially in Europe and Asia, to finance the construction and operation of various public utilities, including power plants, water plants, piers, communications and broadcasting systems, and the like. In such instances, management is usually vested in quasi-public bodies that receive direct governmental payments. Cash payments have been employed to promote the cultivation of idle lands as well as to encourage the removal of other lands from production. Cash payments and readjustment allowances for disabled veterans and for the victims of floods, fires, and other disasters are used throughout the world. Bounties have been given for a wide range of acts, including the apprehension of criminals and tax evaders, for the destruction of rats, and for the birth of children. The promotion of the arts, science, humanities, and research is frequently implemented through cash grants to individuals, foundations, and private firms.

Grants-in-aid

An important form of direct cash payment made by most central governments is the intergovernmental grant-in-aid received by local governments. The British Industrial Development Act of 1966, for example, instituted an elaborate system of cash grants to local governments and private firms to encourage the economic growth of less developed regions. Similarly, the United States, France, and other countries have instituted grants to local governments in order to encourage urban redevelopment programs and slum removal. In international affairs, nations have subsidized the wars of smaller or less affluent allies through direct cash payments, the provision of war materials, and favourable loans (see FOREIGN AID PROGRAMS).

Indirect subsidies. Indirect subsidies arise when governments purchase directly from private producers at higher-than-market prices, maintain higher prices through manipulation of markets, provide services to private enterprises at prices below the cost of providing the service, or grant special tax concessions.

Perhaps the best examples of direct purchases at high prices have been governmental purchases of metals for monetary reserves and strategic materials for defense stockpiles. For many years the U.S. government purchased silver from domestic producers, sometimes at prices more than double the world-market price. It also paid premium prices to producers of rare metals during and following World War II as a part of the stockpiling programs. The amount of subsidy involved in such transactions is the difference between the free-market or world price and the price actually paid. Rules and regulations requiring the domestic government to buy all or a specified minimum of the purchases from domestic producers are common and typically result in subsidies to high-cost domestic producers. Properly employed, such purchase programs can encourage infant industries, mitigate short-term balance-of-payments problems, and contribute to national self-sufficiency. At the same time, they increase the costs of operating the government and may, unless carefully monitored, result in the preservation of inefficient producers, and the storage of essentially useless supplies.

Price-support programs

Indirect subsidies are frequently implemented through price-support and tariff programs. All developed nations have used price supports in one form or another to protect the income of producers in the presence of surplus production; similarly, tariffs have been used to exclude foreign competition or alleviate balance-of-pay-

ments problems. The farm price-support program in the United States, for example, is designed to maintain the prices of farm products—principally wheat, corn, and cotton—at levels designed to maintain a parity relationship between farm prices and other prices in the economy. If the market price is below the government support price, the government lends the farmer an amount equal to the difference. If the market price rises above the support price, the grower retains the crop and repays the loan. When the market price remains below the support price, the government becomes the owner, and storage and disposal of the surplus production becomes a costly undertaking. Since price-support programs encourage surplus production, they are usually supplemented by production controls and quotas, surplus-removal programs, soil-conservation programs, and other devices aimed at offsetting the undesirable side effects.

Great Britain and a number of other nations allow the price of farm products to seek the market level and then subsidize the difference between the market price and one considered fair to the farmer. This type of subsidy program has the advantage of lowering prices to the final consumer while reducing or eliminating storage and surplus removal costs. South Africa requires millers to pay the support price to farmers and then subsidizes millers who sell their flour at a loss to bakers; the consumer receives the benefits of a low price for bread. A major problem with farm price-support programs is that large, efficient, and affluent farmers may receive the bulk of the subsidy. To avoid this, Great Britain places a limit on the amount of output that will be subsidized; and the United States is experimenting with a dollar limit on the total amount of subsidy a grower may receive. Farm price-support programs may also prevent the development of effective trade relations between nations. One of the major stumbling blocks in the development of the Common Market in Europe has been the inability to reach agreement on the prices of farm products. To enter the Common Market, for example, Great Britain would have to buy agricultural products from France and Germany at prices well above the prices available in other Commonwealth countries.

Tariffs

Import tariffs are fundamentally a form of selective price support for particular domestic industries. Such tariffs frequently seem motivated by a desire to protect the employment, incomes, and profits of producers threatened by foreign competition. Though tariffs, especially if they are temporary, may be justified in certain instances, they deprive the domestic consumer of inexpensive foreign products, discourage advantageous international trade, and deny the domestic economy the most advantageous use of its resources. Though international efforts to remove tariff barriers have met with some success, they have also been frustrated by the imperatives of domestic policy and politics.

Governments may also encourage price-stabilization programs by permitting or actually administering concerted trade association activity designed to raise or maintain prices when competition threatens the existing structure of prices, employment, or balance-of-payments. Resale-price maintenance, fair mark-up, and other types of price regulation are common in almost every nation.

Free or low-priced government services

Subsidies are also created when governments provide services to private firms at prices that do not recover costs. Statistical, research, and informational services and technical, managerial, and marketing aids are often provided to farmers, small businesses, and various other groups at no charge. Postal rates on some classes of mail may not be remunerative. Transportation industries may receive indirect subsidy in the use of government-financed rights-of-way—highways, airports, airways, waterways, and canals.

Government promotion of housing and home ownership takes a variety of forms, including financing of mortgages at reduced interest rates, provision of publicly owned housing for low-income families at rents below cost, and broadening the availability of credit by insuring or guaranteeing home mortgages. Governments provide agricultural credit to meet emergencies created by crop failure, to finance the purchase and enlargement of farms and farm buildings, and to purchase feed, fertilizer, livestock, and machinery. Favourable loan terms also have been extended to rural electric cooperatives, railroads, and other enterprises characterized by public service.

Tax concessions

Tax exemptions, tax credits, differential tax rates, and other forms of special tax treatment represent another major source of indirect subsidy. Central governments employ their income tax systems to encourage consumer expenditures on housing, health services, charity programs, and education. Systems of business taxation may promote particular industries, particular forms of business and social organization, or specific business practices. The oil-depletion allowance in the United States, for example, has promoted the exploration of oil by permitting firms to deduct more than the actual cost of production in calculating taxable income. Provisions for accelerated depreciation and investment tax credits have stimulated business investment and encouraged improved levels of economic activity. Partial or complete tax exemptions have been used by governments to attract industry, to reward veterans, and to aid the aged and disabled.

Economic effects. Although subsidies are initiated and justified in terms of benefits that inure to the general public, they result in either a higher level of general taxation or higher prices for consumer goods. They may also encourage the preservation of inefficient producers. The test of the desirability of subsidy hinges on a comparison of the public benefits (which are usually diffuse and difficult to measure) with their costs in terms of higher prices, taxes, and inefficiency. In some cases, the public benefits are more illusory than real; and, in cases in which the subsidy is implemented through interference with the price system, it may create inefficiencies that more than offset the gain. When subsidies are not required in order to maintain production or other economic activities at a socially desirable level, they amount to a pure income transfer from the public to the favoured industry. In such instances, subsidies still may be justified on the ground that they may raise the income of subsidized individuals and industries to levels that the community considers desirable. Even in these cases, however, resources that the economy could employ in more productive uses are encouraged to remain in the subsidized industry, where the value of output to the community may be low.

As a general rule, subsidy programs should be so contrived as to eliminate the necessity for their continuation. Failure to adapt subsidy policy to changing circumstances may perpetuate an unnecessary payment and may prevent the most efficient use of resources in the subsidized activities as well as in other types of economic endeavour. Many economists have a preference for direct subsidies. This improves the level of social bookkeeping, creates less confusion in the political process, encourages a more direct comparison between public and private benefits and costs, and minimizes interference in the operation of the price system.

BIBLIOGRAPHY. W. ADAMS and H.M. GRAY, *Monopoly in America: The Government as Promoter* (1955), a classic work describing those governmental policies that encourage and preserve monopoly, with a description and evaluation of subsidies, in both their concealed and direct forms; K. BRANDT, *Farm Price Supports: Rigid or Flexible*? (1954), an examination of the advantages and disadvantages of alternative devices for developing farm price supports and implementing subsidy in the case of agriculture; H.G. JOHNSON, *Economic Policies Toward Less Developed Countries* (1967), a carefully reasoned economic analysis of various policies toward less developed countries, including tariff, aid, and other subsidy-like policies; C.L. DEARING and W. OWEN, *National Transportation Policy* (1949), a classic treatment of transportation policy, including the role of direct and indirect subsidy in the transportation industries; C. WILCOX, *Public Policies Toward Business*, rev. ed. (1960), a leading textbook in the area of the relations between government and business,

with a chapter on subsidy policy; C. KAYSEN, "On Defining a Subsidy," in *Public Policy*, Yearbook of the Harvard Graduate School of Public Administration, pp. 3–10 (1953), an effort to define subsidy in both its economic and political dimensions.

(M.Z.K.)

Sudan, The

The Democratic Republic of the Sudan (Jumhūrīyat as-Sūdān ad-Dīmuqraṭīyah) is the largest African nation. It is bounded on the north by Egypt; on the east by the Red Sea and Ethiopia; on the south by Kenya, Uganda, and Zaire; on the west by the Central African Republic and Chad; and on the northwest by Libya. Its total area of 967,491 square miles (2,505,805 square kilometres) represents more than 8 percent of the African continent and almost 2 percent of the world's total land area. The population is small in relation to the country's size, with an estimated 16,489,000 persons. Khartoum, the national capital, is located in the northern half of the country at the junction of the Blue and White Nile rivers. The name Sudan comes from the ancient Arabic expression *bilād as-sūdān* ("land of the blacks") for the Upper Nile Valley.

An ancient corridor between Egypt and Central Africa, The Sudan is the home of Arabs in the north and black Africans in the south. Its economy is based on pastoralism and subsistence agriculture, and the limited development in commercial agriculture and industry is concentrated in the north. Plagued with political instability since its independence in 1956, the country has no formal constitution. Among its problems, perhaps the most important is that of the unification of its peoples. Successive regimes were unable to placate the fears and suspicions of the southern tribes, and guerrilla warfare continued throughout the south from the 1950s to the early 1970s. Costly in both men and money, the war delayed development in both sections of the country and prevented The Sudan from fulfilling its promise as representative of both the Arab and black African worlds. The settlement ending the war, reached in March 1972 at Addis Ababa, granted the Sudan's southern provinces a certain measure of autonomy.

(For related physical features see NILE RIVER; RED SEA; and SAHARA DESERT. For history see NILOTIC SUDAN, HISTORY OF THE.)

THE LANDSCAPE

Relief features. The Sudan is comprised of an immense plain that is bounded on the west by the Nile–Congo watershed and the highlands of Darfur and on the east by the Ethiopian Highlands and the Red Sea Hills (ʿAtbāy). The plain can be divided into the northern area of rock desert of the Sahara; the western Qawz, an area of undulating sand dunes that merges northward into the rock desert; and the central and southern wedge-shaped clay plain, the centre of which is occupied by an enormous swamp known as as-Sudd.

Three physical regions

The centre of the clay plain is diversified by inselbergs (isolated mountains rising abruptly from the plains), the largest group of which forms the Nuba Mountains (Jibāl an-Nūbah). The volcanic pile of the Jabal Marrah (*jabal*, "mountains" in Arabic) rises out of the Darfur Plateau in the west to altitudes between 3,000 and 10,000 feet above sea level. The mountains form the Nile–Congo watershed and the western boundary of the central plain. Local small plateaus, such as Jabal Maydūb and Boma in Upper Nile province (*mudīrīyah*) in the southeast, are formed by outlines of lava; volcanic craters retain their form in many areas, such as Jabal Marrah, and Maydūb and Bayyūḍah, in Northern province.

The western plain is mainly composed of Nubian sandstones, which form a dissected plateau region with flat-topped mesas and buttes. Along the western frontier, the Erdi Plateau stands at 4,760 feet (1,450 metres), the Wādī Howar Pleateau at 4,170 feet (1,270 metres), and Jabal al-ʿUwaynāt at 2,790 feet (850 metres).

The Red Sea Hills region is an uplifted escarpment. The scarp slope facing the Red Sea forms rugged hills that are deeply incised by streams. The western slope dips more gradually into the plain, with inselbergs, partly drowned in sediments. On the east, the escarpment overlooks a narrow coastal plain that is ten to 25 miles wide and festooned with dunes and coral reefs. Further south the eastern uplands constitute the foothills of the Ethiopian highland massif.

In the south there are two contrasting upland areas. The Ironstone Plateau lies between the Nile–Congo watershed and the southern clay plain; the level country is diversified with inselbergs. On the Uganda border there are massive ranges with peaks rising to more than 10,000 feet. The Imatong Mountains contain Mt. Kinyeti (10,456 feet [3,187 metres]), the highest in The Sudan.

Drainage and soils. The drainage pattern of The Sudan may be divided into the four hydrographic regions: the southeastern perennial streams; the central seasonal watercourses; the northern zone of no surface drainage; and the southern clay plain, where—despite heavy and frequent rainfall—the level surface impedes drainage.

The Nile River

All streams and rivers drain either into or toward the Nile. The White Nile enters the country through rapids at Nimule on the Uganda border. It flows northward as the Baḥr al-Jabal until its confluence with the left-bank tributary of the Baḥr al-Ghazāl, after which it is known as the Baḥr al-Abyad. The White Nile receives most of its water from the right-bank Sobat River, which flows from the Ethiopian Highlands to join the Nile near Malakāl. The main stream loses most of its water in as-Sudd as it flows northward to Khartoum.

Most of the Nile system's waters are brought from Ethiopia by the Blue Nile (al-Baḥr al-Azraq), which rises in Lake Tana and flows south and west to enter The Sudan near Juba, Ethiopia. The river continues northwestward to meet with the White Nile at Khartoum and then flows in a great, curving northward course as the Nile (Nahr an-Nīl).

The land varies from an absence of soils in the extreme north, through a range of arid and semi-arid soils characterized by gypsum, salt, and calcium carbonate, to leached humid soils exhibiting lateritic tendencies (*e.g.*, containing iron oxides and aluminum hydroxide).

The surface of the soilless desert is either bare rock, a mantle of bare waste, or sandy expanses of mobile dunes known as ergs. In the semi-arid zone, the layer of rock waste is slightly modified to form immature soils; in the Qawz region, soils are brownish red and of low fertility. Alluvial soils occur at the desert deltas of the Gash and Barakah rivers, along the two Niles, and in the alluvial plains of rivers radiating from Jabal Marrah.

The alkaline soils of the central and southern plains are heavy cracking clays. The soil of Gezira (al-Jazīrah) plain south of Khartoum is deep-cracking, uniform clay that has been deposited during the annual inundations of the Blue Nile, while the clays of as-Sudd were deposited in the area of impeded drainage. The more steeply sloping plain between Gezira and as-Sudd is an extensive area of cracking clay.

Alkaline soils are also found in the extensive crystalline outcrops of the central plain. The soils of the Ironstone Plateau vary from deep-reddish loams to light soils with a high content of ironstone (a hard rock rich in iron) pebbles. Immature montane soils are found in the Jabal Marrah, the Boma Plateau, and the Imatong Mountains.

Climate. The fundamental climatic factor is the seasonal migration and pulsation of the northern tropical continental air mass and the southern maritime continental air mass, which are divided by the Intertropical Convergence Zone. In winter, the north winds of the tropical air mass blow across The Sudan toward the front, which may be as far south as the Tropic of Capricorn. These winds are relatively cool and dry and bring no rain except when they blow across the Red Sea. By April the front begins to move northward across the country, and moist unstable air of the maritime air mass is drawn in from the South Atlantic Ocean. Central and southern Sudan have rainy seasons the total lengths of which vary according to their latitude.

The Sudan is a hot country, for temperatures are rarely

mitigated by altitude. The central region experiences the highest mean annual temperatures; at Khartoum temperatures of more than 100° F (38° C) can be recorded during any month of the year. The highest temperatures normally occur just before the rainy season.

Rainfall patterns

Rainfall varies from almost nothing in the north to more than 57 inches annually in the extreme south. Along the Red Sea the climate is alleviated by sea breezes, and most of the rain falls during winter. Precipitation usually occurs during the summer in showers or thunderstorms in the afternoon or evening. Dust storms are common in the north, while the rainy season lasts for eight to nine months a year in the south.

Vegetation and animal life. The Sudan includes five main vegetational belts that run from east to west, more or less in coincidence with rainfall patterns. The desert region in the north forms about one-fourth of the total area. It is followed southeastward by semidesert, low- and high-rainfall savanna (grassland) with inland flood plains, and mountain vegetation regions.

Natural vegetation is largely absent because of the effects of centuries of cultivation and grass fires that annually sweep across more than half the country. Further dangers to plant life are the effects of overstocking, soil erosion, the lowering of the water table, and the advance of the desert into the central region.

Wild life includes the lion, leopard, and cheetah as well as the elephant, giraffe, zebra, wild ass, rhinoceros, buffalo, hippopotamus, ibex (a type of wild goat), wild sheep, and numerous antelopes. The chimpanzee, baboon, grivet (a white and green monkey), and colobus monkey are found in the forests. Resident birds include bustards, guinea fowl, several kinds of partridge, geese, cranes, Egyptian vultures, storks, pelicans, plover, weaverbirds, shrikes, and starlings. Thousands of seabirds, waterfowl, and other birds winter in The Sudan during their migration north. Reptiles include the crocodile and various lizards; snakes are not numerous, but eight poisonous varieties are found. Insect life is abundant; the vicinity of rivers and swamps is infested with mosquitoes, seroot flies (large blood-sucking houseflies) are a scourge during the wet months, and scorpions attain eight inches in length. The tsetse fly is found south of latitude 12° N whenever suitable conditions are found.

Traditional regions. The northern Sudan largely corresponds to Northern province (*mudīrīyah*) and includes the narrow ribbon of cultivable land beside the Nile to the Sixth Cataract at Wad Hāmid. The western Sudan includes the entire area of mixed nomadism and peasant agriculture that lies west of the White Nile in Kordofan and Darfur provinces (*mudīrīyāt*). This region is subdivided into the four subregions of western Darfur, the Qawz sands, the semidesert, and the Nuba Mountains. The central clay plains area roughly corresponds to Blue Nile province and the southern extent of Kassala province.

Most of Kassala province falls in the eastern Sudan, which is generally devoted to nomadism, but includes the irrigated deltas of the Gash and Baraka rivers. The southern clay plains area includes all of the heavy clay soil subject to flooding in the Upper Nile basin. Throughout the region the sources of livelihood are cattle raising, agriculture, and fishing. The Ironstone Plateau and the southern hills region is comprised of the area between the Upper Nile basin and the Nile–Congo watershed, the large hill masses east of the Nile in Equatoria province, and the arid hills and plains of the southeast.

The landscape under human settlement. Rural settlement is invariably clustered because of problems of water supply, especially during the dry months. Along the northern Nile villages are often linear and run parallel to the river. In the extreme south, however, dispersed settlement is favoured.

House types vary from north to south. In the north, flat-topped roofs and sun-dried bricks daubed with clay prevail, while, in the central and southern regions, round huts are built with thatched conical roofs made of grass, millet stalks, and wood poles. In the central region, compounds of buildings are often surrounded by walls made of millet stalks.

Towns are few and widely scattered; only 10 percent of the population can be considered urban. The prime function of the town is undoubtedly administrative rather than commercial. Each of the three southern provinces of Bahr el-Ghazal, Upper Nile, and Equatoria has one major town of between 10,000 and 15,000 inhabitants and no other settlement of comparable size. Urbanization is slightly more advanced in Kordofan and Darfur provinces, where trade is more highly developed. The urban proportion of 11 percent of the total population of Northern province is attributed to ʿAṭbarah, the administrative centre that contains the main workshops of Sudan Railways. The high proportion of 16 percent urban population in Kassala province is caused by Port Sudan, The Sudan's only outlet to the sea and the numerous towns in the cotton-growing deltas of the Gash and Barakah rivers. In Blue Nile province the proportion of urban population is low—7 percent. With few exceptions all towns lie along one of the Niles, because a permanent water supply is necessary in a region of few wells and little surface drainage.

The Three Towns

Khartoum, the smallest of the provinces, contains the Three Towns of Khartoum (population 300,000), Omdurman (289,000), and Khartoum North (149,000). Khartoum was founded on a defensive site by the Egyptians in 1824. As the seat of the government and higher education and the centre of commercial and industrial undertakings, it is firmly established as the centre of power in the country. Omdurman contrasts with Khartoum in being more traditionally Sudanese than Khartoum, which is Arabized. Khartoum North has a more functional atmosphere because of its new light-industrial area.

The Sudan, Area and Population

	area		population	
	sq mi	sq km	1956 census	1972 estimate
Provinces (*mudīrīyāt*)				
Bahr el-Ghazal (Baḥr al-Ghazāl)	82,529	213,751	991,000	1,537,000
Blue Nile (an-Nīl al-Azraq)	54,879	142,138	2,070,000	3,398,000
Darfur (Dārfūr)	191,648	496,369	1,329,000	1,825,000
Equatoria (al-Istiwāʾīyah)	76,495	198,121	904,000	1,403,000
Kassala (Kassalā)	131,527	340,655	941,000	1,754,000
Khartoum (al-Khurṭūm)	8,097	20,971	505,000	945,000
Kordofan (Kurdufān)	146,929	380,546	1,762,000	3,027,000
Northern (ash-Shamālīyah)	184,198	477,074	873,000	1,220,000
Upper Nile (Aʿālī an-Nīl)	91,189	236,180	889,000	1,380,000
Total Sudan	967,491	2,505,805	10,263,000*	16,489,000

*Figures do not add to total given because of rounding.
Source: Official government figures.

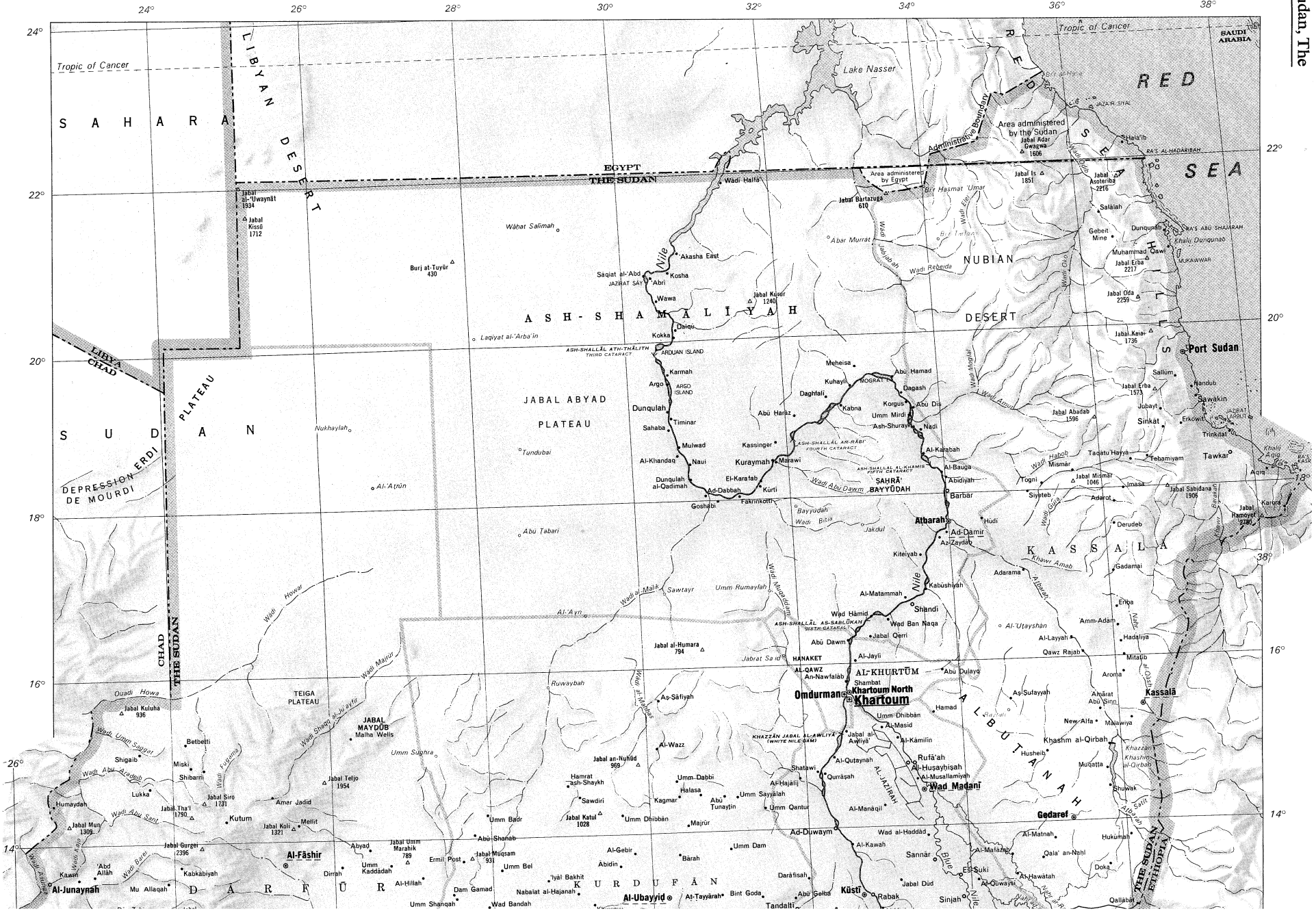
PEOPLE AND POPULATION

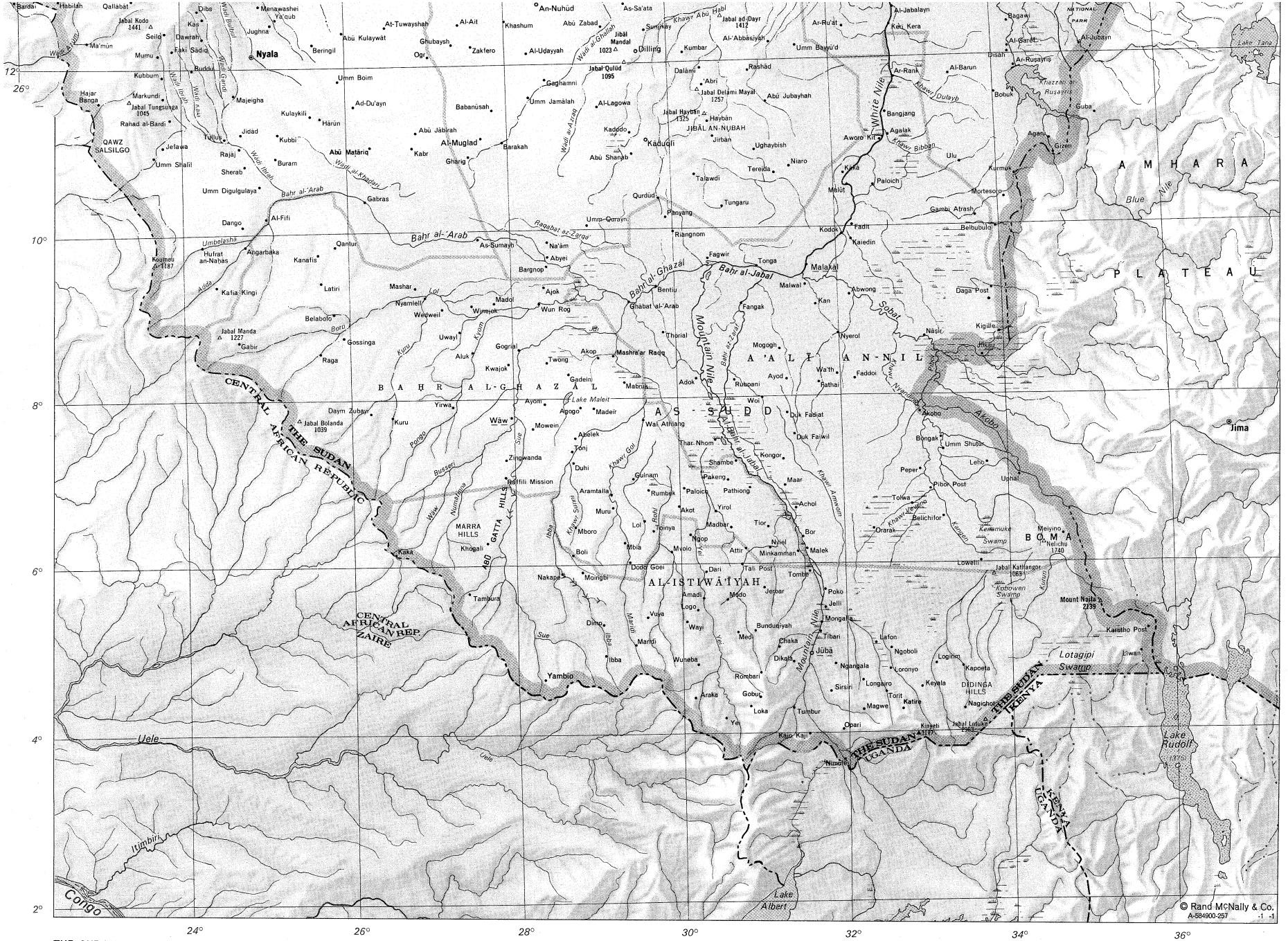
Population groups. The northern Sudan is predominantly Muslim, and its people speak Arabic, the official language. Although commonly called Arabs, the people are of mixed Arab, Hamitic, and Negro ancestry. The actual infusion of Arab blood may be slight, but almost all northern Sudanese are agreed on the acceptance of the Arabic cultural heritage. Discernible ethnic groups include the Arabs, the Nubians, the Beja, and the black Funj (Fung), Nuba, and tribes of the Darfur regions.

The Nubians are the dominant element of the population of the northern Nile. The inhabitants of the Halfa district and Dunqulah as far south as ad-Dabbah speak the Nubian language, although most Nubians speak Arabic. Situated astride the Nile corridor linking Egypt with The Sudan, the ancient Nubian kingdom was the country's historic avenue of contact with the Mediterranean world.

The ancestors of the Beja, a Hamitic group of cattle owners, arrived in the eastern Sudan between 4000 and 2500 BC. They are remarkably similar to predynastic Egyptians but have an admixture of Arab and Negro blood. The Beja are unique among the northern peoples because their language has survived Arabization.

The principal Arab migrations into The Sudan from the north and east began in the 14th century and continued spasmodically until the 19th century. The Arabs occupy most of the fertile lands along the Nile and constitute the


24°
26°
28°
30°
32°
34°
36°
38°
Tropic of Cancer
SAUDI ARABIA
RED SEA
Lake Nasser
LIBYAN DESERT
SAHARA
EGYPT
THE SUDAN
Area administered by the Sudan
Area administered by Egypt
Administrative Boundary
Jabal Adar Gwagwa 1606
Jabal Is 1851
Jabal Asoteriba 2216
Jabal Bartaugga 610
Jabal al-'Uwaynāt 1934
Jabal Kissū 1712
Wāhat Salimah
Burj at-Tuyūr 430
Akasha East
Saqiat al-'Abd
Kosha
Abri
Wawa
Jabal Kusur 1240
Daiqu
Kokka
ASH-SHAMALIYAH
NUBIAN DESERT
Laqiyat al-'Arba'in
ASH-SHALLĀL ATH-THĀLITH THIRD CATARACT
ARDUAN ISLAND
Karmah
Argo
ARGO ISLAND
Dunqulah
Timinar
Sahaba
Mulwad
Al-Khandaq
Naui
Kuraymah
Marawi
Dunqulah al-Qadimah
El-Karafab
Ad-Dabbah
Kurti
Goshabi
Fakirinkotti
Kassinger
ASH-SHALLĀL AR-RĀBI FOURTH CATARACT
ASH-SHALLĀL AL-KHAMIS FIFTH CATARACT
SAHRĀ' BAYYŪDAH
Bayyudah
Wadi Abu Dawm
Meheisa
Kuhayli
MOGRAT IS.
Daghlali
Abu Hamad
Dagash
Kabna
Korgus
Abu Dis
Umm Mirdi
Ash-Shuraya
Nadi
Abu Haraz
Al-Karabah
Al-Bauga
Abidiyah
Barbar
Atbarah
Ad-Damir
Az-Zaydab
Hudi
Kiteiyah
Kabūshiyah
Al-Matammah
Shandi
Wad Hamid
Wad Ban Naqa
ASH-SHALLĀL AS-SABŪKAH SIXTH CATARACT
Abu Dawm
Jabal Qerri
Al-Jayli
HANAKET
AL-QAWZ
An-Nawfalab
Shambat
AL-KHURTŪM
Khartoum North
Khartoum
Omdurman
Abu Dulayq
Umm Dhibban
Al-Masid
Al-Kāmilīn
Jabal al-Awliyā'
KHAZZĀN JABAL AL-AWLIYĀ (WHITE NILE DAM)
Al-Qutaynah
Shatawi
Qurrāṣah
Al-Hajalij
Rufā'ah
Al-Husayhisah
Al-Musallamiyah
Wad Madanī
Al-Jazīrah
ALBUTANAH
Gedaref
Al-Manāqil
Ad-Duwaym
Wad al-Haddād
Al-Kawah
Al-Matnah
Hukumah
Al-Matāzah
Qala' an-Nahl
Doka
Al-Hawātah
Al-Quwaysī
Es-Suki
Sannār
Blue Nile
Sinjah
Jabal Dūd
Rabak
Kūstī
Abū Gelba
Bint Goda
At-Tayārah
Al-Ubayyid
Tandalti
Nabalat al-Hajanah
'Iyāl Bakhit
Dārāfisah
Abidin
Bārah
Umm Dam
Al-Gebir
Umm Bel
Umm Shanqah
Wad Bandah
Dam Gamad
Al-Hillah
Umm Kaddādah
Dirrah
Ermil Post
Jabal Muqsam 931
Abyad
Jabal Umm Marahik 789
Jabal Gurgei 2396
Kabkābiyah
Mu Allaqah
Al-Junaynah
Kawm
'Abd Allāh
Al-Fāshir
DĀRFŪR
KURDUFĀN
Umm Badr
Abū Shanab
Jabal Kuli 1321
Mellit
Kutum
Amar Jadid
Jabal Sirū 1731
Jabal Tha'l 1790
Jabal Mun 1309
Humaydah
Lukka
Shibarni
Miski
Shigaib
Betbetti
Jabal Kuluba 936
Ouadi Howa
TEIGA PLATEAU
CHAD
THE SUDAN
LIBYA
CHAD
PLATEAU
SUDAN
DEPRESSION ERDI DE MOURDI
JABAL ABYAD PLATEAU
Nukhaylah
Tundubai
Al-'Atrūn
Abū Ṭabari
Al-'Ayn
Umm Rumaylah
Sawtayr
Wadi al-Malik
Wadi Howar
Wadi Mayur
Wadi Shaqq al-Jayyil
Wadi Umm Saggel
Wadi Abu Arageib
Wadi Fugeima
Wadi Abu Sant
Wadi Kajo
Wadi Barei
Wadi Asheim
JABAL MAYDŪB
Malha Wells
Umm Sughra
Jabal Teljo 1954
Amar Jadid
Abd Allah
Jabal Gurgei 2396
Jabal Umm Marahik 789
Jabal Muqsam 931
Jabal Katul 1028
Jabal an-Nuhūd 969
Jabal al-Humara 794
Jabrat Sa'd
Ruwaybah
Wadi al-Mahas
Aṣ-Ṣāfiyah
Al-Wazz
Hamrat ash-Shaykh
Sawdiri
Umm Dabbi
Halasa
Kagmar
Abū Tunaytin
Umm Qantur
Umm Dubbān
Majrūr
Al-Gebir
Abidin
Bārah
Umm Dam
Darāfisah
Abū Gelba
Jabal Dūd
Sinjah
Nile
Blue Nile
Atbara
Setit
Gash
Baraka
Khawr Amabi
Al-Buzaqi
Khawr Amab
Wadi Mukattab
Wadi Habob
Wadi Mismar
Wadi Gita
Wadi Ambur
Wadi Rebeida
Wadi Dabbah
Wadi Bitia
Wadi Jakdul
Wadi Nuqaidam
Wadi Qitr
Wadi Qitr
Khaliq Dunqunab
Khaliq Aqiq
Khaliq Kasr
Jabal Kasr 1736
Jabal Erba 1573
Jabal Abadab 1596
Jabal Mismar 1046
Jabal Sabidana 1906
Jabal Ramoyet 1700
Jabal Erba 2217
Jabal Oda 2259
Jabal Asoteriba 2216
Jabal Is 1851
Jabal Adar Gwagwa 1606
Jabal Bartaugga 610
Jabal Kissū 1712
Jabal al-'Uwaynāt 1934
Jabal Kuluba 936
Jabal Tha'l 1790
Jabal Sirū 1731
Jabal Teljo 1954
Jabal Mun 1309
Jabal Kuli 1321
Jabal Katul 1028
Jabal an-Nuhūd 969
Jabal al-Humara 794
Jabal Qerri
Jabal al-Awliyā'
Jabal Kusur 1240
Jabal Erba 2217
Jabal Oda 2259
Jabal Kasr 1736
Jabal Erba 1573
Jabal Abadab 1596
Jabal Mismar 1046
Jabal Sabidana 1906
Jabal Ramoyet 1700
Port Sudan
Sawākin
Sinkāt
Tawkar
Kassalā
Gadamai
Erba
Hadaliya
Mitaib
Aroma
Amarat
Abū Sinn
New Alfa
Malawiya
Khashm al-Qirbah
Khazzan Khashm al-Qirbah
Muqatta
Shuwak
Aṣ-Ṣufayyah
Qawz Rajab
Al-Layyah
Adarama
Derudeb
Siyateb
Aderot
Togni
Imasa
Barbar
Abidiyah
Ad-Damir
Hudi
Atbarah
Jakdul
Wadi Bitia
Wadi Abu Dawm
Bayyudah
Fakirinkotti
Goshabi
Ad-Dabbah
Kurti
Dunqulah al-Qadimah
El-Karafab
Marawi
Kuraymah
Naui
Al-Khandaq
Mulwad
Timinar
Sahaba
Dunqulah
Argo
Karmah
ARDUAN ISLAND
ASH-SHALLĀL ATH-THĀLITH THIRD CATARACT
Laqiyat al-'Arba'in
Wadi Halfa
Abar Murrat
Bir Hasmat 'Umar
Bir Tarfawi
Wadi Elei
Wadi Jabbabah
Wadi Rebeida
Wadi Qitr
Khaliq Dunqunab
Khaliq Aqiq
Khaliq Kasr
Wadi Mukattab
Wadi Habob
Wadi Mismar
Wadi Gita
Wadi Ambur
Wadi Abu Dawm
Wadi Bitia
Wadi Jakdul
Wadi Nuqaidam
Wadi Howar
Wadi Mayur
Wadi Shaqq al-Jayyil
Wadi Umm Saggel
Wadi Abu Arageib
Wadi Fugeima
Wadi Abu Sant
Wadi Kajo
Wadi Barei
Wadi Asheim
Wadi al-Malik
Wadi al-Mahas
Ruwaybah
Jabrat Sa'd
Umm Rumaylah
Sawtayr
Al-'Ayn
Abū Ṭabari
Tundubai
Nukhaylah
Al-'Atrūn
Burj at-Tuyūr 430
Wāhat Salimah
Lake Nasser
Bir al-Hiya
JAZA'IR SIYAL
RA'S AL-HADARIBAH
RA'S ABŪ SHAJARAH
RA'S KASR
JAZIRAT ARKUIT
Tropic of Cancer
LIBYAN DESERT
SAHARA
EGYPT
THE SUDAN
LIBYA
CHAD
THE SUDAN
CHAD
PLATEAU
SUDAN
DEPRESSION ERDI DE MOURDI
JABAL ABYAD PLATEAU
TEIGA PLATEAU
JABAL MAYDŪB
Malha Wells
ASH-SHAMALIYAH
NUBIAN DESERT
KURDUFĀN
DĀRFŪR
RED SEA
SAUDI ARABIA
THE SUDAN
ETHIOPIA
KASSALĀ
ALBUTANAH
AL-JAZĪRAH
KHAZZĀN JABAL AL-AWLIYĀ (WHITE NILE DAM)
Area administered by the Sudan
Area administered by Egypt
Administrative Boundary
Al-Junaynah
Al-Fāshir
Al-Ubayyid
Kūstī
Wad Madanī
Gedaref
Khartoum
Khartoum North
Omdurman
Port Sudan
Kassalā
Sawākin
Tawkar
Sinkāt
Atbarah
Barbar
Shandi
Abū Dulayq
Al-Qawz
Al-Khurtūm
Hanaket
Abū Dawm
Al-Matammah
Kabūshiyah
Nile
Blue Nile
Atbara
Setit
Gash
Baraka
Qala' an-Nahl
Doka
Al-Hawātah
Al-Quwaysī
Es-Suki
Sannār
Sinjah
Jabal Dūd
Rabak
Abū Gelba
Bint Goda
At-Tayārah
Tandalti
Nabalat al-Hajanah
'Iyāl Bakhit
Dārāfisah
Abidin
Bārah
Umm Dam
Al-Gebir
Umm Bel
Umm Shanqah
Wad Bandah
Dam Gamad
Al-Hillah
Umm Kaddādah
Dirrah
Ermil Post
Jabal Muqsam 931
Abyad
Jabal Umm Marahik 789
Jabal Gurgei 2396
Kabkābiyah
Mu Allaqah
Kawm
'Abd Allāh
Abū Shanab
Umm Badr
Jabal Kuli 1321
Mellit
Kutum
Amar Jadid
Jabal Sirū 1731
Jabal Tha'l 1790
Jabal Mun 1309
Humaydah
Lukka
Shibarni
Miski
Shigaib
Betbetti
Jabal Kuluba 936
Ouadi Howa
Jabal Katul 1028
Jabal an-Nuhūd 969
Jabal al-Humara 794
Jabal Qerri
Jabal al-Awliyā'
Jabal Kusur 1240
Jabal Erba 2217
Jabal Oda 2259
Jabal Kasr 1736
Jabal Erba 1573
Jabal Abadab 1596
Jabal Mismar 1046
Jabal Sabidana 1906
Jabal Ramoyet 1700
Jabal Asoteriba 2216
Jabal Is 1851
Jabal Adar Gwagwa 1606
Jabal Bartaugga 610
Jabal Kissū 1712
Jabal al-'Uwaynāt 1934
Burj at-Tuyūr 430
Wāhat Salimah
Lake Nasser
Bir al-Hiya
JAZA'IR SIYAL
RA'S AL-HADARIBAH
RA'S ABŪ SHAJARAH
RA'S KASR
JAZIRAT ARKUIT
Tropic of Cancer
Khaliq Dunqunab
Khaliq Aqiq
Khaliq Kasr
Wadi Mukattab
Wadi Habob
Wadi Mismar
Wadi Gita
Wadi Ambur
Wadi Abu Dawm
Wadi Bitia
Wadi Jakdul
Wadi Nuqaidam
Wadi Howar
Wadi Mayur
Wadi Shaqq al-Jayyil
Wadi Umm Saggel
Wadi Abu Arageib
Wadi Fugeima
Wadi Abu Sant
Wadi Kajo
Wadi Barei
Wadi Asheim
Wadi al-Malik
Wadi al-Mahas
Wadi Qitr
Wadi Qitr
Wadi Rebeida
Wadi Dabbah
Wadi Elei
Abar Murrat
Bir Hasmat 'Umar
Bir Tarfawi
Ruwaybah
Jabrat Sa'd
Umm Rumaylah
Sawtayr
Al-'Ayn
Abū Ṭabari
Tundubai
Nukhaylah
Al-'Atrūn
Burj at-Tuyūr 430
Wāhat Salimah
Laqiyat al-'Arba'in
ASH-SHALLĀL ATH-THĀLITH THIRD CATARACT
ARDUAN ISLAND
ARGO ISLAND
ASH-SHALLĀL AR-RĀBI FOURTH CATARACT
ASH-SHALLĀL AL-KHAMIS FIFTH CATARACT
ASH-SHALLĀL AS-SABŪKAH SIXTH CATARACT
KHAZZĀN JABAL AL-AWLIYĀ (WHITE NILE DAM)
Area administered by the Sudan
Area administered by Egypt
Administrative Boundary
Nile
Blue Nile
Atbara
Setit
Gash
Baraka
Khawr Amabi
Al-Buzaqi
Khawr Amab
Wadi Mukattab
Wadi Habob
Wadi Mismar
Wadi Gita
Wadi Ambur
Wadi Abu Dawm
Wadi Bitia
Wadi Jakdul
Wadi Nuqaidam
Wadi Howar
Wadi Mayur
Wadi Shaqq al-Jayyil
Wadi Umm Saggel
Wadi Abu Arageib
Wadi Fugeima
Wadi Abu Sant
Wadi Kajo
Wadi Barei
Wadi Asheim
Wadi al-Malik
Wadi al-Mahas
Wadi Qitr
Wadi Qitr
Wadi Rebeida
Wadi Dabbah
Wadi Elei
Ruwaybah
Jabrat Sa'd
Umm Rumaylah
Sawtayr
Al-'Ayn
Abū Ṭabari
Tundubai
Nukhaylah
Al-'Atrūn
Burj at-Tuyūr 430
Wāhat Salimah
Lake Nasser
Bir al-Hiya
JAZA'IR SIYAL
RA'S AL-HADARIBAH
RA'S ABŪ SHAJARAH
RA'S KASR
JAZIRAT ARKUIT
Tropic of Cancer
Khaliq Dunqunab
Khaliq Aqiq
Khaliq Kasr
Jabal Kasr 1736
Jabal Erba 1573
Jabal Abadab 1596
Jabal Mismar 1046
Jabal Sabidana 1906
Jabal Ramoyet 1700
Jabal Erba 2217
Jabal Oda 2259
Jabal Asoteriba 2216
Jabal Is 1851
Jabal Adar Gwagwa 1606
Jabal Bartaugga 610
Jabal Kissū 1712
Jabal al-'Uwaynāt 1934
Jabal Kuluba 936
Jabal Tha'l 1790
Jabal Sirū 1731
Jabal Teljo 1954
Jabal Mun 1309
Jabal Kuli 1321
Jabal Katul 1028
Jabal an-Nuhūd 969
Jabal al-Humara 794
Jabal Qerri
Jabal al-Awliyā'
Jabal Kusur 1240
Port Sudan
Sawākin
Sinkāt
Tawkar
Kassalā
Gadamai
Erba
Hadaliya
Mitaib
Aroma
Amarat
Abū Sinn
New Alfa
Malawiya
Khashm al-Qirbah
Khazzan Khashm al-Qirbah
Muqatta
Shuwak
Aṣ-Ṣufayyah
Qawz Rajab
Al-Layyah
Adarama
Derudeb
Siyateb
Aderot
Togni
Imasa
Barbar
Abidiyah
Ad-Damir
Hudi
Atbarah
Kiteiyah
Kabūshiyah
Al-Matammah
Shandi
Wad Hamid
Wad Ban Naqa
Abū Dawm
Jabal Qerri
Al-Jayli
Hanaket
Abū Dulayq
Al-Qawz
Al-Khurtūm
Shambat
Khartoum North
Khartoum
Omdurman
Umm Dhibban
Al-Masid
Al-Kāmilīn
Jabal al-Awliyā'
Al-Qutaynah
Shatawi
Qurrāṣah
Al-Hajalij
Rufā'ah
Al-Husayhisah
Al-Musallamiyah
Wad Madanī
Gedaref
Al-Manāqil
Ad-Duwaym
Wad al-Haddād
Al-Kawah
Al-Matnah
Hukumah
Al-Matāzah
Qala' an-Nahl
Doka
Al-Hawātah
Al-Quwaysī
Es-Suki
Sannār
Sinjah
Jabal Dūd
Rabak
Kūstī
Abū Gelba
Bint Goda
At-Tayārah
Al-Ubayyid
Tandalti
Nabalat al-Hajanah
'Iyāl Bakhit
Dārāfisah
Abidin
Bārah
Umm Dam
Al-Gebir
Umm Bel
Umm Shanqah
Wad Bandah
Dam Gamad
Al-Hillah
Umm Kaddādah
Dirrah
Ermil Post
Jabal Muqsam 931
Abyad
Jabal Umm Marahik 789
Jabal Gurgei 2396
Kabkābiyah
Mu Allaqah
Al-Junaynah
Kawm
'Abd Allāh
Al-Fāshir
Abū Shanab
Umm Badr
Jabal Kuli 1321
Mellit
Kutum
Amar Jadid
Jabal Sirū 1731
Jabal Tha'l 1790
Jabal Mun 1309
Humaydah
Lukka
Shibarni
Miski
Shigaib
Betbetti
Jabal Kuluba 936
Ouadi Howa
Jabal Katul 1028
Jabal an-Nuhūd 969
Jabal al-Humara 794
Jabal Qerri
Jabal al-Awliyā'
Jabal Kusur 1240
Jabal Erba 2217
Jabal Oda 2259
Jabal Kasr 1736
Jabal Erba 1573
Jabal Abadab 1596
Jabal Mismar 1046
Jabal Sabidana 1906
Jabal Ramoyet 1700
Jabal Asoteriba 2216
Jabal Is 1851
Jabal Adar Gwagwa 1606
Jabal Bartaugga 610
Jabal Kissū 1712
Jabal al-'Uwaynāt 1934
Burj at-Tuyūr 430
Wāhat Salimah
Lake Nasser
Bir al-Hiya
JAZA'IR SIYAL
RA'S AL-HADARIBAH
RA'S ABŪ SHAJARAH
RA'S KASR
JAZIRAT ARKUIT
Tropic of Cancer
Khaliq Dunqunab
Khaliq Aqiq
Khaliq Kasr
Wadi Mukattab
Wadi Habob
Wadi Mismar
Wadi Gita
Wadi Ambur
Wadi Abu Dawm
Wadi Bitia
Wadi Jakdul
Wadi Nuqaidam
Wadi Howar
Wadi Mayur
Wadi Shaqq al-Jayyil
Wadi Umm Saggel
Wadi Abu Arageib
Wadi Fugeima
Wadi Abu Sant
Wadi Kajo
Wadi Barei
Wadi Asheim
Wadi al-Malik
Wadi al-Mahas
Wadi Qitr
Wadi Qitr
Wadi Rebeida
Wadi Dabbah
Wadi Elei



THE SUDAN

Size of symbol indicates relative size of town • ○ ◉ ▣

Elevations in metres

MAP INDEX

Political subdivisions

A'alī an-Nīl (Upper Nile) 9·00n 32·00e
al-Istiwā'īyah (Equatoria) 5·00n 31·00e
al-Khurṭūm 15·45n 32·30e
an-Nīl al-Azraq (Blue Nile) 13·00n 33·00e
ash-Shamālīyah (Northern) 19·00n 30·00e
Baḥr al-Ghazāl 8·00n 28·00e
Dārfūr 13·00n 25·00e
Kassalā 18·00n 36·00e
Kurdufān 13·00n 30·00e

Cities and towns

'Abd Allāh 13·30n 23·02e
Abekr 12·43n 28·55e
Abelek 7·23n 28·46e
'Ābidīn 13·33n 29·38e
'Abidīyah 18·14n 33·57e
'Abrī 11·40n 30·28e
'Abrī 20·48n 30·20e
Abū Dawm 16·16n 32·36e
Abū Dīs 19·08n 33·34e
Abu Dulayq 15·54n 33·49e
Abū Gelba 13·11n 31·52e
Abū Ḥamad 19·32n 33·19e
Abū Ḥarāz 12·58n 29·52e
Abū Ḥarāz 19·04n 32·07e
Abū Jābirah 11·04n 26·51e
Abū Jubayhah 11·27n 31·14e
Abū Kulaywāt 12·20n 26·00e
Abū Maṭāriq 10·58n 26·17e
Abū Na'āmah 12·44n 34·08e
Abū Shanab 10·47n 29·32e
Abū Shanab 13·57n 27·47e
Abū Ṭunayṭin 14·24n 31·01e
Abū Zabad 12·21n 29·15e
Abwong 9·07n 32·12e
Abyaḍ 13·46n 26·28e
Abyei 9·36n 28·26e
Achol 6·34n 31·31e
Adarama 17·05n 34·54e
Adarot 17·50n 36·07e
ad-Dabbah 18·03n 30·57e
ad-Dāmir 17·35n 33·58e
ad-Du'ayn 11·26n 26·09e
ad-Duwaym 14·00n 32·19e
Adok 8·11n 30·19e
Agalak 11·01n 32·42e
Agaru 10·59n 34·44e
Agogo 7·49n 28·52e
Ajok 9·15n 28·27e
'Akasha East 21·05n 30·43e
Akobo 7·47n 33·01e
Akop 8·21n 29·05e
Akot 6·33n 30·03e
al-'Abbāsīyah 12·10n 31·18e
al-Ait 12·22n 27·27e
al-'Atrūn 18·11n 26·36e
al-Barun 11·44n 33·30e
al-Bauga 18·16n 33·55e
al-Fāshir (el-Fasher) 13·38n 25·21e
al-Fifī 10·03n 25·01e
al-Garef 12·03n 34·19e
al-Gebir 13·43n 29·49e
al-Ḥajālīj 14·36n 31·54e
al-Ḥawātah 13·25n 34·38e
al-Ḥillah 13·27n 27·08e
al-Ḥuḍayb 13·00n 32·50e
al-Ḥuṣayḥiṣah 14·44n 33·18e
al-Jabalayn 12·36n 32·48e
al-Jaylī 16·01n 32·36e
al-Jubayn 12·07n 35·10e
al-Junaynah (Geneina) 13·27n 22·27e
al-Kāmilīn 15·05n 33·11e
al-Karabah 18·33n 33·42e
al-Kawah 13·44n 32·30e
al-Khandaq 18·36n 30·34e
al-Khurṭūm, see Khartoum
al-Khurṭūm Baḥrī, see Khartoum North
al-Lagowa 11·24n 29·08e
al-Layyah 16·16n 35·25e
al-Mafāzah 13·36n 34·33e
al-Manāqil 14·15n 32·59e
al-Masīd 15·15n 32·57e
al-Matammah 16·43n 33·22e
al-Matnah 13·47n 35·03e
al-Muglad 11·02n 27·44e
al-Qaḍārif, see Gedaref
al-Quṭaynah 14·52n 32·21e
al-Quwaysī 13·20n 34·05e
al-Ubayyiḍ (el-Obeid) 13·11n 30·13e
al-Uḍayyah 12·03n 28·17e
Aluk 8·26n 27·27e
al-Wazz 15·01n 30·10e
Amadi 5·31n 30·20e
'Amārat Abū Sinn 15·21n 35·45e
'Amm-Adam 16·22n 36·06e
Amar Jadīd 14·28n 25·14e
Angarbaka 9·44n 24·44e
an-Nawfalāb 15·52n 32·32e
an-Nuhūd 12·42n 28·26e
'Aqīq 18·14n 38·12e
Araka 4·20n 30·23e
Aramtalla 6·47n 29·13e
Argo 19·31n 30·25e
Aroma 15·49n 36·08e
ar-Rahad 12·43n 30·39e
ar-Rank 11·45n 32·48e
ar-Ru'at 12·21n 32·17e
ar-Ruṣayriṣ 11·51n 34·23e
ash-Shurayk 18·48n 33·34e
as-Sa'ata 13·37n 29·59e
aṣ-Ṣāfīyah 15·31n 30·07e
aṣ-Ṣufayyah 15·30n 34·42e
as-Sumayḥ 9·49n 27·39e
'Aṭbarah (Aṭbara) 17·42n 33·59e
Aṭ-Ṭayyārah 13·12n 30·47e
Attir 6·04n 30·50e
aṭ-Ṭuwayshah 12·21n 26·32e
Aworo Kit 10·59n 32·38e
Ayod 8·07n 31·26e
Ayom 7·52n 28·23e
az-Zaydāb 17·26n 33·53e
Babanūsah 11·20n 27·48e
Bagawi 12·19n 34·21e
Bangjang 11·23n 32·42e
Bārah 13·42n 30·22e
Barakah 10·58n 27·59e
Barbar (Berber) 18·01n 33·59e
Bardai 12·43n 22·00e
Bargnop 9·30n 28·28e
Belabolo 8·57n 25·51e
Belbubulo 9·57n 34·04e
Belichifor 6·33n 33·16e
Bentiu 9·14n 29·50e
Berber, see Barbar
Beringil 12·10n 25·41e
Betbetti 15·06n 24·12e
Bint Goda 13·17n 31·33e
Bobuk 11·30n 34·05e
Boli 6·01n 28·43e
Bongak 7·27n 33·14e
Bor 6·12n 31·33e
Buddu 11·54n 24·08e
Bunduqiyah 5·06n 30·53e
Buram 10·49n 25·10e
Būr Sūdān, see Port Sudan
Chaka 4·49n 31·14e
Daga Post 9·12n 33·58e
Dagash 19·22n 33·24e
Daghfalī 19·17n 32·30e
Dalāmī 11·52n 30·28e
Dalqū 20·07n 30·37e
Dam Gamad 13·17n 27·28e
Dango 10·00n 24·45e
Daqqāq 12·56n 26·58e
Darāfīsah 13·23n 31·59e
Dari 5·48n 30·21e
Dār Zubi 13·07n 23·40e
Dawrah 12·22n 24·19e
Daym Zubayr 7·43n 26·13e
Derudeb 17·32n 36·06e
Dibs 12·34n 24·14e
Dikala 4·41n 31·23e
Dilling 12·03n 29·39e
Dimo 5·19n 29·10e
Dirrah 13·37n 26·06e
Disah 12·02n 34·19e
Dodo Goei 5·57n 27·26e
Doka 13·31n 35·46e
Dongola, see Dunqulah
Duhi 7·07n 28·45e
Duk Fadiat 7·45n 31·25e
Duk Faiwil 7·30n 31·29e
Dunkuj 12·50n 32·49e
Dunqulah (Dongola) 19·10n 30·29e
Dunqulah-al-Qadīmah 18·13n 30·45e
Dunqunāb 21·06n 37·05e
el-Fasher, see al-Fāshir
el-Karafab 18·10n 31·36e
el-Obeid, see al-Ubayyiḍ
Eriba 16·37n 36·04e
Erkowit 18·46n 37·07e
Ermil Post 13·37n 27·36e
es-Suki 13·20n 33·54e
Faddoi 8·07n 32·07e
Fadit 9·58n 32·13e
Fagwir 9·33n 30·25e
Faki Ṣādiq 12·08n 23·55e
Fakrinkotti 18·01n 31·20e
Fangak 9·04n 30·53e
Fathai 8·05n 31·48e
Fongfong 12·56n 23·14e
Gabir 8·35n 24·40e
Gabras 10·16n 26·14e
Gadamai 17·09n 36·06e
Gadein 8·11n 28·44e
Gaghamni 11·41n 28·19e
Galegu 12·36n 35·02e
Gambi Aṭrash 10·03n 33·47e
Gebeit Mine 21·03n 36·19e
Gedaref (al-Qaḍārif) 14·02n 35·24e
Geneina, see al-Junaynah
Ghābat al-'Arab 9·02n 29·29e
Gharig 10·47n 27·33e
Ghubaysh 12·09n 27·21e
Gizen 10·49n 34·48e
Gobur 4·20n 31·04e
Gogrial 8·32n 28·07e
Goshabi 17·58n 31·06e
Gossinga 8·39n 25·59e
Gulnam 6·55n 29·30e
Habīlah 12·41n 22·33e
Hadalīya 16·10n 36·06e
Ḥajar Banga 11·30n 23·00e
Halasa 14·26n 30·39e
Ḥamad 15·19n 33·43e
Ḥamrat ash-Shaykh 14·35n 27·58e
Handub 19·14n 37·16e
Hārūn 11·20n 25·43e
Haybān 11·13n 30·31e
Hūdī 17·42n 34·17e
Ḥufrat an-Naḥās 9·45n 24·19e
Ḥukūmah 13·52n 36·07e
Humaydah 14·22n 22·31e
Husheib 14·54n 35·07e
Ibba 4·48n 29·06e
Imasa 18·01n 36·12e
'Iyāl Bakhīt 13·25n 28·41e
Jabal al-Awliyā' 15·14n 32·30e
Jabal Dūd 13·25n 33·09e
Jabal Qerri 16·15n 32·48e
Jefawa 10·57n 23·48e
Jelli 5·22n 31·48e
Jerbar 5·39n 31·05e
Jidād 11·05n 24·44e
Jirbān 11·03n 30·36e
Jokau 8·24n 33·49e
Jūbā 4·51n 31·37e
Jubayt 18·57n 36·50e
Jughna 12·24n 25·06e
Kabkābīyah 13·39n 24·05e
Kabna 19·10n 32·41e
Kabr 10·54n 26·50e
Kabūshīyah 16·53n 33·42e
Kadodo 11·04n 29·31e
Kāduqlī 11·01n 29·43e
Kafia Kingi 9·16n 24·25e
Kagmar 14·24n 30·25e
Kaiedin 9·45n 32·11e
Kajo Kaji 3·53n 31·40e
Kākā 10·36n 32·11e
Kan 9·01n 31·47e
Kanafīs 9·48n 25·40e
Kapoeta 4·47n 33·35e
Karmah 19·38n 30·25e
Karora 17·42n 38·22e
Karotho Post 5·11n 35·50e
Kas 12·30n 24·17e
Kassalā 15·28n 36·24e
Kassinger 18·45n 31·54e
Katire 4·02n 32·47e
Kawm 13·31n 22·50e
Keri Kera 12·21n 32·46e
Keyala 4·27n 32·52e
Khartoum (al-Khurṭūm) 15·36n 32·32e
Khartoum North (al-Khurṭūm Baḥrī) 15·38n 32·33e
Khashm al-Qirbah 14·58n 35·55e
Khashum 12·27n 28·02e
Khogali 6·08n 27·47e
Khuwayy 13·05n 29·14e
Kigille 8·40n 34·02e
Kodok 9·53n 32·07e
Kokka 20·00n 30·35e
Kongor 7·10n 31·21e
Korgus 19·13n 33·29e
Kosha 20·49n 30·32e
Kosti, see Kūstī
Kubbī 11·08n 25·14e
Kubbum 11·47n 23·47e
Kuḥaylī 19·25n 32·50e
Kulaykilī 11·21n 25·36e
Kumbar 12·03n 30·16e
Kuraymah 18·33n 31·51e
Kurmuk 10·33n 34·17e
Kūrtī 18·07n 31·33e
Kuru 7·43n 26·31e
Kūstī (Kosti) 13·10n 32·40e
Kutum 14·12n 24·40e
Kwajok 8·19n 28·00e
Lafon 5·02n 32·27e
Latiri 9·10n 25·43e
Leho 7·07n 33·52e
Liwan 4·54n 35·40e
Logirim 4·43n 33·14e
Logo 5·20n 30·18e
Loka 4·16n 31·01e
Lol 6·26n 29·37e
Longairo 4·30n 32·17e
Loronyo 4·39n 32·38e
Lowelli 5·59n 33·45e
Lukka 14·33n 23·42e
Maar 6·54n 31·33e
Mabrūk 8·07n 29·25e
Madbar 6·19n 30·40e
Madeir 7·50n 29·12e
Madol 9·02n 27·46e
Magwe 4·08n 32·17e
Majeigha 11·33n 24·40e
Majrūr 14·01n 30·27e
Malakāl 9·31n 31·39e
Malawiya 15·16n 36·12e
Malek 6·04n 31·36e
Malha Wells 15·08n 26·12e
Malūṭ 10·26n 32·12e
Malwal 9·19n 31·35e
Ma'mūn 12·15n 22·41e
Mapoi 5·28n 27·40e
Marawi (Merowe) 18·29n 31·49e
Marīdī 4·55n 29·28e
Markundi 11·33n 23·49e
Mashar 9·14n 26·52e
Mashra'ar-Raqq 8·25n 29·16e
Mbia 6·15n 29·19e
Mboro 6·18n 28·45e
Medi 5·04n 30·44e
Meheisa 19·37n 32·57e
Meiyino 6·12n 34·40e
Mellit 14·08n 25·33e
Menawashei 12·40n 24·59e
Merowe, see Marawī
Minkamman 6·03n 31·32e
Miskī 14·51n 24·13e
Mismār 18·13n 35·38e
Misterei 13·07n 22·09e
Mitatib 16·03n 36·11e
Modo 5·29n 30·38e
Mogogh 8·26n 31·19e
Moingbi 5·46n 28·49e
Mongalla 5·12n 31·46e
Mortesoro 10·12n 34·09e
Mowein 7·36n 28·11e
Mu'Allaqah 13·28n 23·57e
Muḥammad Qawl 20·54n 37·05e
Mulwad 18·39n 30·35e
Mumu 12·06n 23·42e
Muqaṭṭa' 14·40n 35·51e
Murnei 12·57n 22·52e
Muru 6·36n 29·15e
Mvolo 6·03n 29·56e
Na'ām 9·42n 28·27e
Nabalat al-Ḥajanah 13·13n 29·02e
Nadi 18·40n 33·42e
Nagichot 4·16n 33·34e
Nakape 5·47n 28·37e
Nāṣir 8·36n 33·04e
Naui 18·28n 30·43e
Ngangala 4·42n 31·55e
Ngoboli 4·57n 32·37e
Ngop 6·16n 30·12e
Niaro 10·38n 31·31e
Nimule 3·36n 32·03e
Nyala 12·03n 24·53e
Nyamlell 9·07n 26·58e
Nyerol 8·41n 32·02e
Nyiel 6·06n 31·13e
Omdurman (Umm Durmān) 15·38n 32·30e
Opari 3·56n 32·03e
Orarak 6·15n 32·23e
Pakeng 6·55n 30·40e
Paloich 6·45n 30·08e
Paloich 10·28n 32·32e
Panyang 10·04n 29·58e
Pathiong 6·46n 30·54e
Peper 7·04n 33·00e
Pibor Post 6·48n 33·08e
Poko 5·38n 31·50e
Port Sudan (Būr Sūdān) 19·37n 37·14e
Qala 'an-Naḥl 13·38n 34·57e
Qallābāt 12·43n 23·26e
Qallābāt 12·58n 36·09e
Qantur 9·45n 25·52e
Qawz Rajab 16·04n 35·34e
Qurdūd 10·17n 29·56e
Qurrāṣah 14·38n 32·12e
Rabak 13·09n 32·44e
Raffili Mission 6·53n 27·58e
Raga 8·28n 25·41e
Rahad al-Barāī 11·18n 23·53e
Rashād 11·51n 31·04e
Rajāj 10·55n 24·43e
Riangnom 9·55n 30·01e
Rombari 4·33n 31·02e
Ruboani 8·06n 30·45e
Rufā'ah 14·46n 33·22e
Rumbek 6·48n 29·41e
Sahaba 18·55n 30·28e
Salālah 21·19n 36·13e
Salim 12·52n 28·40e
Sallūm 19·23n 37·06e
Sannār (Sennar) 13·33n 33·38e

MAP INDEX (continued)

Sāqiat al-'Abd...20·48n 30·19e
Sawākin (Suakin).....19·07n 37·20e
Sawdirī.....14·25n 29·05e
Seilo.....12·20n 23·50e
Sennar, see Sannār
Shambat.....15·40n 32·32e
Shambe..... 7·07n 30·46e
Shandī.....16·42n 33·26e
Shatawī.....14·39n 32·06e
Sherab.....10·43n 24·47e
Shibarnī.....14·50n 24·25e
Shigaib.....15·01n 23·36e
Shuwak.....14·23n 35·52e
Sinjah.....13·09n 33·56e
Sinkāt.....18·50n 36·50e
Sirsiri..... 4·24n 31·53e
Siyeteb.....18·00n 35·01e
Suakin, see Sawākin
Sumayḥ.....12·43n 30·50e
Sunjikāy.....12·20n 29·46e
Suq'at al-Jamal..12·48n 27·42e
Talawdī.....10·38n 30·23e
Tali Post..... 5·54n 30·47e
Tambura..... 5·36n 27·28e
Tandaltī.....13·01n 31·52e
Taqāṭu 'Hayyā...18·20n 36·22e
Ṭawkar.....18·26n 37·44e
Tehamiyam.....18·20n 36·32e
Tereida.....10·35n 31·17e
Thar Nhom..... 7·26n 30·29e
Thorial..... 8·40n 29·56e
Tibari..... 5·01n 31·43e
Timinar.....19·02n 30·29e
Tior..... 6·23n 31·11e
Togni.....18·04n 35·13e
Toinya..... 6·17n 29·44e
Tolwa..... 6·38n 32·37e
Tombe..... 5·49n 31·41e
Tonga..... 9·28n 31·03e
Tonj..... 7·17n 28·45e
Torit..... 4·24n 32·34e
Trinkitat.....18·41n 37·43e
Tullus.....11·03n 24·33e
Tumbur..... 4·20n 31·34e
Tungaru.....10·14n 30·42e
Twong..... 8·18n 28·20e
Ughaybish.....10·52n 31·05e
Ulu.....10·43n 33·29e
Umm Badr.....14·14n 27·57e
Umm Bayyū'd....12·05n 31·40e
Umm Bel.....13·32n 28·04e
Umm Boim.....11·43n 25·57e
Umm Dabbī.....14·37n 30·23e
Umm Dam.....13·45n 30·59e
Umm Dhibbān...15·26n 32·51e
Umm Digulgulaya.....10·29n 24·57e
Umm Durmān, see Omdurman
Umm Jamālah...11·27n 28·12e
Umm Kaddādah.13·36n 26·42e
Umm Kuwaykah.13·00n 32·17e
Umm Mirdi.....18·59n 33·32e
Umm Qantur....14·17n 31·22e
Umm Qurayn.... 9·58n 28·55e
Umm Ruwābah..12·54n 31·13e
Umm Sayyālah...14·25n 31·10e
Umm Shalīl.....10·51n 23·42e
Umm Shanqah...13·14n 27·14e
Umm Shuṭūr..... 7·17n 33·14e
Uphal..... 6·58n 34·16e
Uwayl..... 8·46n 27·24e
Vuya..... 5·21n 29·40e
Wad al-Ḥaddād..13·49n 33·32e
Wad Bandah.....13·06n 27·57e
Wad Ban Naqa...16·30n 33·08e
Wad Ḥāmid.....16·30n 32·48e
Wādī Ḥalfā'.....21·56n 31·20e
Wad Madanī.....14·25n 33·28e
Wal Athiang..... 7·42n 29·40e
Wa'th..... 8·10n 32·07e
Wāw..... 7·42n 28·00e
Wawa.....20·26n 30·21e
Wayi..... 5·11n 30·10e
Wedweil..... 9·00n 27·12e
Winejok..... 9·01n 27·34e
Woi..... 7·53n 31·10e
Wuneba..... 4·50n 30·20e
Wun Rog..... 9·00n 28·21e
Yambio..... 4·34n 28·23e
Ya'qūb.....12·29n 25·11e
Yei..... 6·15n 30·13e
Yirol..... 6·33n 30·30e
Yirwa..... 7·47n 27·15e
Zakfero.....12·10n 27·35e
Zalingei.....12·54n 23·29e
Zingwanda..... 7·10n 27·56e

Physical features and points of interest

Abadab, Jabal, *mountain*.....18·53n 35·59e
Abū 'Aradeib, Wādī, *ravine*....14·12n 23·06e
Abū Dawm, Wādī, *watercourse*.....18·28n 31·49e
Abū Gatta Hills.. 6·06n 27·44e
Abū Ḥabl, Khawr, *watercourse*.....12·49n 31·15e
Abū Sanṭ, Wādī, *ravine*.....14·11n 23·06e
Abū Shajarah, Ra's, *point*.....21·04n 37·14e
Abū Ṭabarī, *well*.17·35n 28·31e
Abyaḍ, al-Baḥr al-, see White Nile
Adda, *river*..... 9·51n 24·50e
Akobo, *river*..... 7·48n 33·03e
al-'Ayn, *well*.....16·36n 29·19e
al-Gharbiya, Sahara, see Libyan Desert
al-Jazīrah (The Gezira), *physical region*..14·25n 33·00e
Amab, Khawr, *ravine*.....17·08n 34·51e
'Amūr, Wādī, *ravine*.....18·56n 33·34e
Amwom, Khawr, *watercourse*..... 7·50n 31·13e
an-Nuhūd, Jabal, *mountain*.....14·50n 29·53e
'Aqīq, Khalij, *bay*.....18·20n 38·10e
'Arab, Baḥr al, *river*..... 9·02n 29·28e
Arduan Island...19·55n 30·22e
Argo Island.....19·25n 30·27e
ar-Rīḥ, Jazīrat, *peninsula*.....18·10n 38·27e
Asoteriba, Jabal, *mountain*.....21·51n 36·30e
as-Sudd, *physical region*.. 8·00n 31·00e
'Atbarah, *river*...17·40n 33·56e
Azraq, al-Baḥr al-, see Blue Nile
Azraq, Wādī al-, *watercourse*.....10·33n 28·40e
'Azūm, Wādī, *river*.....11·33n 20·09e
Barakah, Khawr, *river*.....18·13n 37·35e
Barei, Wādī, *ravine*.....13·27n 23·57e
Bartazuga, Jabal, *mountain*.....21·44n 33·33e
Bayyūḍah, *well*...17·32n 32·07e
Bayyūḍah, Saḥrā', *desert*...18·00n 33·00e
Bibban, Khawr, *watercourse*.....11·00n 32·41e
Bitia, Wādī, *watercourse*.....17·34n 32·15e
Blue Nile (al-Baḥr al-Azraq), *river*.....15·38n 32·31e
Bolanda, Jabal, *mountain*..... 7·44n 25·28e
Boma, *physical region*..... 6·10n 34·28e
Boro, *river*..... 8·52n 26·11e
Bulbul, Wādī, *ravine*.....10·59n 24·33e
Busseri, *river*.... 7·41n 28·03e
Dayr, Jabal ad-, *mountain*.....12·27n 30·42e
Delami Mayal, Jabal, *peak*.....11·38n 30·23e
Didinga Hills, *mountains*..... 4·20n 33·35e
Dindar, Nahr ad-, *river*.....14·06n 33·40e
Dindar National Park.....12·40n 35·20e
Dulayb, Khawr, *watercourse*.....11·45n 32·47e
Dunqunāb, Khalij, *bay*.....21·05n 37·08e
Elei, Wādī, *ravine*.....22·04n 34·27e
Erba, Jabal, *mountain*.....20·45n 36·50e
Erba, Jabal, *mountain*.....19·04n 36·46e
Erdi Plateau.....19·10n 24·00e
Fifth Cataract, see Khamis, ash-Shallal al-
Fourth Cataract, see Rabi', ash-Shallal ar-
Fugama, Wādī, *ravine*.....14·43n 24·36e
Gandi, Wādī, *ravine*.....11·23n
Gash, see Qāsh, Nahr al-,
Gezira, The, see al-Jazīrah
Ghallah, Wādī al-, *watercourse*.....10·25n 27·32e
Ghazāl, Baḥr al-, *river*..... 9·31n 30·25e
Gol, Khawr, *watercourse*..... 6·55n 30·16e
Gura, Wādī, *ravine*.....17·28n 35·10e
Gurgei, Jabal, *mountain*.....13·50n 24·19e
Habob, Wādī, *ravine*.....18·07n 35·01e
Ḥadāribah, Ra's al-, *cape*.....22·04n 36·54e
Hamoyet, Jabal, *mountain*.....17·33n 38·02e
Ḥasmat 'Umar Bi'r, *well*.....21·46n 34·00e
Haybān, Jabal, *mountain*.....11·15n 30·31e
Howar, Wādī, *river*.....17·30n 27·08w
Humara, Jabal al-, *mountain*....16·16n 30·59e
Ibba, *river*..... 7·09n 28·41e
Ibrah, Wādī, *ravine*.....10·36n 24·58e
Is, Jabal, *mountain*.....21·49n 35·39e
Jabal Abyad Plateau.....19·00n 29·00e
Jabal, Baḥr al-, *river*..... 9·30n 30·30e
Jabjabah, Wādī, *watercourse*.....22·37n 33·17e
Jabrat Sa'īd, *well*.16·06n 31·50e
Jakdūl, *pond*.....17·39n 32·59e
Jur, *river*..... 8·39n 29·18e
Kaia, Wādī, *ravine*.....11·31n 24·15e
Kaiai, Jabal, *peak*.....19·54n 36·48e
Kaja, Wādī, *river*.12·02n 22·28e
Kangen, *river*.... 6·47n 33·09e
Kasr, Ra's, *point*.18·02n 38·35e
Kathangor, Jabal, *mountain*..... 5·45n 33·59e
Katul, Jabal, *mountain*.....14·16n 29·23e
Kenamuke Swamp..... 5·55n 33·48e
Khadari, Wādī al-, *ravine*.....10·29n 26·15e
Khāmis, ash-Shallal al- (Fifth Cataract), *rapids*.....18·23n 33·47e
Khashm al-Qirbah, Khazzan, *reservoir*.....14·40n 35·55e
Khazzān Jabal al-Awliya, (White Nile Dam), *dam*.....15·10n 32·27e
Kinyeti, *peak*..... 3·57n 32·54e
Kissū, Jabal, *mountain*.....21·35n 25·09e
Kobowen Swamp..... 5·38n 33·54e
Kodo, Jabal, *mountain*.....12·26n 23·38e
Koli, Jabal, *mountain*.....14·05n 25·31e
Koumou, *mountain*..... 9·34n 23·38e
Kuluha, Jabal, *mountain*.....15·31n 23·25e
Kurun, *river*..... 5·30n 34·17e
Kuru, *river*..... 9·08n 26·57e
Kurur, Jabal, *mountain*.....20·31n 31·32e
Kyom, *river*..... 8·58n 28·13e
Laqīyat al-Arba'īn, *oasis*...16·25n 34·30e
Larbut, Jazīrat, *island*.....18·47n 37·33e
Libyan Desert (Sahara al-Gharbiya)....22·00n 26·00e
Lol, *river*..... 9·13n 28·59e
Lotagipi Swamp, *plain*..... 4·40n 34·50e
Lotuke, Jabal, *mountain*..... 4·07n 33·48e
Mahbas, Wādī al-, *watercourse*.....15·50n 29·45e
Majrūr, Wādī, *watercourse*.....15·44n 26·26e
Maleit, Lake..... 7·55n 28·35e
Malik, Wādī al-, *watercourse*.....18·02n 30·58e
Manda, Jabal, *mountain*..... 8·39n 24·27e
Mandal, Jibal, *hill*.....12·09n 29·31e
Maridi, *river*..... 4·55n 29·28e
Marrah, Jabal, *mountain*.....14·04n 24·21e
Marra Hills..... 6·05n 27·33e
Maydūb, Jabal, *hills*.....15·14n 26·30e
Mismār, Jabal, *mountain*.....18·06n 35·42e
Moglal, Wādī, *ravine*.....19·18n 34·29e
Mountain Nile, *river*..... 9·30n 30·30e
Mukawwar, *island*.....20·48n 37·13e
Mun, Jabal, *mountain*.....14·08n 22·42e
Muqaddam, Wādī, *watercourse*.....18·04n 31·30e
Muqsam, Jabal, *mountain*.....13·38n 27·42e
Murrāt, Ābār, *well*.....21·03n 32·55e
Naita, Mount, *mountain*..... 5·32n 35·19e
Nasser, Lake.....21·30n 31·00e
Nelichu, *mountain*..... 6·08n 34·25e
Nile (Nahr an-Nīl), *river*....21·10n 30·40e
Nūbah, Jibāl an-, *mountains*.....12·00n 30·45e
Nubian Desert...20·30n 33·00e
Nukhaylah, *oasis*.19·03n 26·19e
Numatinna, *river*. 7·14n 27·37e
Nyanding, Khawr, *watercourse*..... 8·40n 32·41e
Oda, Jabal, *mountain*.....20·21n 36·39e
Odib, Wādī, *ravine*.....22·38n 36·06e
Oko, Wādī, *ravine*.....21·15n 35·56e
Pibor, *river*..... 8·26n 33·13e
Pongo, *river*..... 8·42n 27·40e
Qāsh, Nahr al- (Gash), *river*....16·48n 35·51e
Qawz, Hanakat al-, *dunes*.....15·59n 32·01e
Qulūd, Jabal, *hill*.....11·41n 29·31e
Rabi', ash-Shallal ar-(Fourth Cataract), *rapids*.....18·47n 32·03e
Rahad, Nahr ar-, *river*.....14·28n 33·31e
Rayrah, *well*.....15·21n 34·41e
Rebeida, Wādī, *ravine*.....20·45n 34·06e
Red Sea.....23·00n 37·00e
Red Sea Hills, *mountains*.....23·00n 35·00e
Rohl, *river*..... 6·22n 29·46e
Rusayris, Khazzan al-, *reservoir*.....11·40n 34·20e
Ruwaybah, *well*..15·39n 28·45e
Sabidana, Jabal, *mountain*.....18·04n 36·50e
Sablūkah, Shallāl as- (Sixth Cataract), *rapids*.....16·20n 32·42e
Salīmah, Wāḥat, *oasis*.....21·22n 29·19e
Salsilgo, Qawz, *dunes*.....10·49n 22·54e
Satīt, *river*.....14·20n 35·50e
Sawtayr, *well*.....17·03n 30·24e
Sāy, Jazirat, *island*.....20·42n 30·20e
Shaqq al-Ju'ayfir, Wādī, *ravine*....15·16n 26·00e
Siro, Jabal, *mountain*.....14·23n 24·23e
Sixth Cataract, see Sablūkah, Shallāl as-
Sobat, *river*..... 9·22n 31·33e
Suakin Archipelago, *islands*.....18·42n 38·30e
Sunnī, Khawr, *watercourse*..... 7·09n 28·41e
Sue, *river*..... 7·41n 28·03e
Tarfawi, Bi'r, *well*.....21·04n 34·08e
Teiga Plateau....15·38n 25·40e
Teljo, Jabal, *mountain*.....14·42n 25·56e
Tha'l, Jabal, *mountain*.....14·13n 24·14e
Thālith, ash-Shallālath- (Third Cataract), *rapids*.....19·49n 30·19e
Third Cataract, see Thālith, ash-Shallālath-

MAP INDEX (continued)

Ṭiwāl, Wādī, *watercourses*....10·25n 23·15e
Tundubai, *well*...18·31n 28·33e
Tungsunga, Jabal, *mountain*.....11·29n 23·21e
Ṭuyūr, Burj aṭ-, *mountain*.....20·55n 27·55e
Umbelasha, *river*. 9·51n 24·50e
Umm Marahik, Jabal, *hill*.....13·40n 26·53e
Umm Rumaylah, *well*.....16·55n 31·40e
Umm Saggāt, Wādī, *ravine*....15·15n 23·12e
Umm Sughra, *well*.....15·03n 27·12e
'Uwaynāt, Jabal al-, *mountain*...21·54n 24·58e
Veveno, Khawr, *watercourse*..... 6·40n 32·58e
Wāw, *river*..... 7·03n 27·13e
White Nile (al-Baḥr al-Abyaḍ), *river*.....15·38n 32·31e
White Nile Dam, see Khazzān Jabal al-Awliya
Yei, *river*..... 6·15n 30·13e
Zarāf, Baḥr az-, *river*..... 9·25n 31·10e
Zarqā', Raqabat az-, *watercourse*. 9·14n 29·44e

greater part of the nomads of northern and central Sudan.

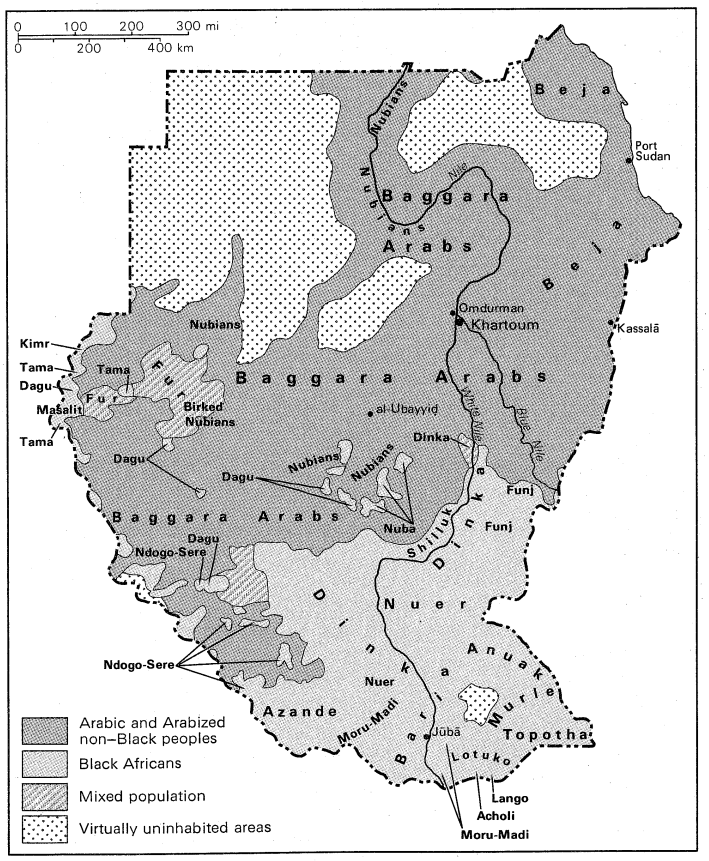
The black African peoples

Of the black northern peoples, the Funj inhabit the land on either side of the Blue Nile upstream of Sannār; the Nuba, who are probably related to groups that stretch toward West Africa, are found in northern Kordofan province; and the Fur, Daju, Beigo, Birked, Masalit, Gimira, and Tama occupy central Darfur province.

The Nilotes of southern Sudan include the Dinka, Nuer, Shilluk, Anuak, Acholi, and Lango. All are conspicuous for their absorbing interest in cattle, which enters widely into their customs and traditions. The Nilo-Hamitic people, including the Bari, Lotuko, Taposa, and Murle, are by tradition cattle owners and, like the Nilotes, are to be found both within and beyond the borders of The Sudan. The tribes of the southwest belong to the eastern Sudanic linguistic group. They include the Azande, the Ndogo-Sere, the Moru-Madi, and the Bango-Baka.

The southern peoples speak their individual tribal languages. Although attempts have been made to establish Arabic as a lingua franca, it is used only in the towns.

Islām is the religion of the Arabs and Arabized peoples of the north; it provides a common cultural basis and unifies social and political ideas. In the south the majority hold to a multiplicity of animist religions, while a small percentage have converted to Roman Catholicism and Protestantism.



General ethnic composition of The Sudan.

Demography. About 90 percent of the population of 16,489,000 lives in rural areas, with an average population density of 17 persons per square mile. The population is concentrated along the Nile and its tributaries, with the highest densities occurring in Khartoum (117 per square mile) and Blue Nile (62 per square mile) provinces. It is a young nation, with almost half the population under 15 years of age and less than 4 percent over 60 years. A joint study of the 1955 census by the Sudanese government and the United Nations revealed that there were about 52 births per 1,000 population and 24 deaths per 1,000, with a natural increase of 28 per 1,000 in that year. It is believed that improved medical facilities since then have resulted in a lower death rate and a higher rate of natural increase. The United Nations estimate for the period of 1965–70 had indicated a birth rate of 49 per thousand, a death rate of 18 per thousand, and a natural increase rate of 31 per thousand. Life expectancy at birth is 48 years of age.

Since the mid-1950s there has been significant emigration from the south in response to the civil war. The United Nations estimated about 180,000 Sudanese refugees in the Central African Republic, Zaire, Ethiopia, and Uganda. It is possible that there are actually about 500,000 refugees, most of whom have been absorbed into their host countries and have not been officially counted. Further dislocation was caused by the government's peace villages; southern Sudanese were moved from contested areas or areas of suspected guerrilla activity to villages guarded by the army. Eight months after the end of the war it was estimated that 250,000 refugees had returned from surrounding countries.

In 1955 immigration amounted to only 249,000 persons, or less than 3 percent of the population. Aliens are concentrated in Kassala, Blue Nile, Kordofan, and Khartoum provinces. They are comprised mainly of West Africans in the south and Egyptians, Asians, and Ethiopians in the north. (M.e.D.S.)

THE NATIONAL ECONOMY

Mineral potential

The extent and distribution of resources. The Sudan is believed to be rich in mineral resources, although it has not been adequately surveyed. There are small quantities of iron ore near the Red Sea, deposits of gold, manganese, chromite, quartz, and marble and reserves of copper in Darfur province, and asbestos near al-Qaḍārif. The search for oil and natural gas has so far been unfruitful, but there is mild optimism concerning offshore oil deposits in the Red Sea. Mining is conducted on a small scale and contributes little to the overall economy.

There is a great potential for the generation of hydroelectric power from the Nile River system. Thermal resources are limited because of the inaccessibility of the southern forests and the lack of coal or oil.

Sources of national income. About one third of the gross domestic product is created in agriculture. The main crops of cotton, durra (common millet), *dukhn* (bulrush millet), sorghum, peanuts (groundnuts), sesame, maize (corn), wheat, and castor are raised on land that is either rain fed (three-fourths of the cultivated area), irrigated, or subjected to flooding. The amount of rain-fed and flooded land varies greatly annually, but irrigation is gradually being extended and irrigated land doubled between 1960 and 1970. The irrigated areas in the north and east produce the bulk of the commercial crops; they are centred on the Gezira Scheme (al-Jazīrah)—with its Mangil extension—between the Blue and White Niles south of Khartoum and the Gash and Baraka deltas. Further irrigation is provided by the Khashm al-Qirbah Dam on the 'Aṭbarah River, which in 1972 irrigated 470,000 acres (190,000 hectares); ar-Ruṣayriṣ Dam, completed in the 1960s was expected in the early 1970s to provide irrigation water for the Rahad Scheme, consisting of some 310,000 acres to the east of the Blue Nile. Smaller irrigation schemes also obtain water from

ar-Ruṣayriṣ. Parts of the rain-fed areas are mechanized, and some irrigation is carried out by use of pumps. Crop production varies annually because of changes in weather conditions and the incidence of pest attacks, of which grasshoppers and desert locusts are the most virulent.

The Gezira Scheme

Cotton is the major cash crop; it provides the country with about 60 percent of its total earnings of foreign exchange. The bulk of the cotton crop is produced on the irrigation schemes of Gezira and its extensions, and future expansions in the irrigated area promise considerable increases in production. In the early 1970s production accounted for about 2 percent of the world's cotton. The Sudan's importance in the world cotton market, however, derives from the fact that the country produces between 40 and 45 percent of the total world production of extra-long-staple cotton (*i.e.*, one and three-eighths of an inch or longer). Less than 1 percent of the crop is used domestically. Cotton production is substantially nationalized through The Sudan Gezira Board, which is responsible for the operation of the Gezira Scheme. Profits are shared among the government, which owns and irrigates the land; the tenant farmers; the Board; local governments; and the Social Development Fund, which provides services to the tenants.

Although extensive areas are potentially suited for growing vast quantities of fruit and vegetables, production is limited mainly to the north and east. Important fruits include dates, mangoes, guavas, citrus fruits, and bananas; the most important vegetable crop is onions. Tobacco is raised in Darfur provinces. Commercial agriculture is undeveloped in the three southern provinces, where most farms are self-sufficient and some produce is sold in local markets. The civil war greatly disrupted farm production.

Livestock is ubiquitous. Camels, sheep, and goats are kept in the north and west, and cattle are raised in the central region and the south. There is some export of animals, hides, and skins, but most livestock is raised for local use.

The Sudan is the world's largest producer of gum arabic, a water-soluble gum obtained from acacia trees and used in the production of adhesives, candy, and pharmaceuticals. The southern forests yield hardwood timber such as mahogany and *sant* (a type of acacia) and softwoods, and there are sawmills near Wāw, Loka, and Katire. The northern woodlands have been deforested by the extraction of wood for fuel and charcoal.

The Nile rivers are an important source of fish, especially Nile perch. Most of the catch is consumed locally. Significant quantities of fish and shellfish are believed to be produced from the Red Sea.

Manufacturing

The manufacturing industry, which is concentrated in the three towns, contributed about 9 percent of the gross domestic product in 1969. Manufacturers include cement, petroleum products, ginned cotton, textiles, leather and footwear, and processed foods. Most factories operate well below full capacity, and low incomes have depressed demand. A considerable part of large-scale manufacturing production takes place in the public sector. The Industrial Development Corporation controls and administers government-owned factories that produce sugar, cardboard, and processed foods. The government tannery processes skins from government-owned and administered slaughtering houses.

In 1970 The Sudan had an installed electrical capacity of about 117,000 kilowatts. The Sannār Dam supplies electricity to Gezira (al-Jazīrah) and Khartoum, and hydroelectric dams have also been built at Khashm al-Qirbah on the ʿAṭbarah River and ar-Ruṣayriṣ on the Blue Nile.

All banks operating in The Sudan were nationalized in 1970. The Bank of Sudan issues all currency and acts as banker to the government. It also holds bilateral trade accounts on behalf of the government and supervises the operations of the seven commercial banks. The Agricultural Bank and the Industrial Bank provide development funds and credit. There are also more than 40 insurance companies. The banking system is geared primarily to the finance of foreign trade and especially the cotton trade.

The Sudan suffers from a chronic unfavourable balance of trade and a growing national debt. Most foreign exchange is earned from cotton, which accounts for more than half of total exports. Other exports include gum arabic, oil seeds, and livestock products. Imports consist largely of manufactured goods, such as textiles, machinery, vehicles, petroleum products, and iron and steel. The most important trading partners are members of the European Economic Community, India, Japan, and Egypt. China and the Soviet Union became substantial trading partners after 1967.

The government's economic role

Management of the economy. All major economic units, whether in agriculture, mining, or manufacturing, are owned or operated by the government. The private sector consists largely of subsistence activities, mainly in farming and pastoralism, but also includes small-scale industry and services. The middle income group is limited, and entrepreneurship is scarce. The central government plays an important economic role through direct participation, such as the Gezira Scheme, and fiscal policy. Traditionally, the government aimed at budgeting for a small surplus that was used, together with foreign assistance, to finance economic development. Since the mid-1960s, however, the government has engaged in deficit finance and has borrowed large sums from the banking system. This has resulted in a decline of private initiative in internal trade, exports, and medium-scale industry.

Little government revenue is raised by direct taxation, while indirect taxes, such as those on imports and excise and consumption duties, provide half the revenue. Other revenue is derived from government monopoly profits on sugar, tea, coffee, and salt and proprietary receipts from agricultural and other schemes.

The smallness of the industrial sector and the predominance of rural life have tended to constrain the development of workers' and employers' associations. In 1958 all trade unions were dissolved but were permitted to function again in 1961 within limits prescribed by legislation. There are trade unions representing the workers in government ministries, the Sudan Railways, the Central Electricity and Water Administration, and other bodies. Unions normally are affiliated to the three central federations of the Sudanese Workers' Union, the Federation of Workers' Trade Unions in the Private Sector, and the Federation of Workers' Trade Unions of the Public Sector. Political activity by trade unions is discouraged.

Leading businessmen are members of the Sudan Chamber of Commerce, and an important pressure group is the Gezira Tenants' Association, which attempts to promote the tenants' interests as opposed to those of the Sudan Gezira Board.

A five-year plan was declared in 1970. The plan did not make special provisions for the development of the south, but emphasized the development of agriculture and livestock as well as transport and communications. Instead of the average annual growth rate of the gross domestic product of 5 percent realized in the period between 1965 and 1970, the planned growth rate for 1970 to 1975 was set at 8 percent per annum. Development was to be funded both by government revenues and foreign loans.

Since 1969 the government has leaned heavily on Socialist principles, but its intentions to federate with Egypt, Syria, and Libya have been impeded by troubles in the southern Sudan. Government intervention in the economy is likely to grow, and the activities of the private sector and the trade unions is likely to be drastically curtailed. Given political stability, the immense resources with which the country is endowed are capable of being translated into considerable economic growth over the long term. (E.I.U.)

Transport. The transport system is underdeveloped and represents a serious constraint on economic growth. The country's vast area and the availability of only one outlet to the sea throw a heavy burden on the limited facilities and especially on the government-owned Sudan Railways, which handles almost all foreign trade and employs about one-fifth of all wage earners.

There are almost 3,400 track route miles of railways

The railway network

throughout the country. The main line runs north from al-Ubbayyiḍ (el-Obeid) via Khartoum to Lake Nasser and the submerged terminal of Wādī Ḥalfā. There are branch lines from Sannār and ʿAṭbarah to Port Sudan and from Sannār to ar-Ruṣayriṣ. There is also a westward extension from al-Ubbayyiḍ to Nyala with a branch line south to Wāw. In the early 1970s, extensions were planned to Jūbā, Aswān, Egypt, and al-Junaynah (in the extreme west of Darfur province); another line will eventually reach from al-Junaynah across Chad to northern Nigeria.

There are more than 6,500 miles of roads. In the north, roads are generally motorable, unimproved tracks except in the towns, and they are impassable after rain. There are several gravelled all-weather roads in the south, where other roads, including the through route from Jūbā to Khartoum, are impassable during the rainy season. In the early 1970s all-weather paved roads were planned to link The Sudan with the Red Sea (at Port Sudan), Ethiopia, Egypt, and Libya.

There are more than 2,300 miles of navigable waterways. The White Nile and the Baḥr al-Ghazal are navigable throughout the year. The Jur and Sobat rivers are seasonal. The Blue Nile is not navigable, and the Nile below Khartoum is only navigable in short stretches because it is broken by cataracts. The government operates steamer services on the White Nile between Kūstī and Jūbā and on the main Nile between Wādī Ḥalfā and Aswān and between Marawī and Kuraymah.

Port Sudan, 850 miles south of Suez, Egypt, is the country's main port on the Red Sea. Minor ports include Sawākin, Trinkitat, and ʿAqīq.

The government-owned Sudan Airways operates domestic and international services from the main airport at Khartoum.

ADMINISTRATION AND SOCIAL CONDITIONS

The government. By December 1972 The Sudan still had no formal constitution. After the military coup d'etat of 1969 the Constituent Assembly and Supreme Council were abolished and the provisional constitution suspended. In 1971 legislative, judicial, and executive powers were vested in the president, Major General Gaafar al-Nimeiry, who was also commander of the armed forces. In January 1972 the Sudanese Socialist Union approved the establishment of a Sudanese People's Assembly and approved a draft permanent constitution that was to be amended and endorsed by the People's Assembly upon its creation. The proposed Sudanese People's Assembly was to consist of 207 members, of whom 32 members were to be appointed by President al-Nimeiry and the rest elected by popular vote. The election of members of the Sudanese People's Assembly took place during September–October of that year. The new government, headed by Prime Minister (former President) al-Nimeiry, was announced in October 1972, and the Assembly opened later the same month. Their first task was the drafting of a new constitution embodying the agreements reached in the peace settlement of March 1972.

Local government

Local government was reorganized in 1971 to increase the power and functions of local councils. The country is divided into nine provinces, each headed by a Province Popular Executive Council. The provincial councils are chaired by the governor of the province, who is appointed by the Council of Ministers on recommendation of the Minister of Local Government. The membership is comprised of a two-thirds majority of members elected by the rural and town councils and of representatives of the central government. Beneath the provincial councils are subregional councils and rural and town councils. Their membership is also comprised of central government officials and members elected by functional organizations such as village development committees and trade unions. The smallest political units—village and block councils—are comprised only of directly elected members.

Coordination of local government activities is the responsibility of the provincial councils. The central government appropriates funds to the provincial councils, which then distribute the funds and powers to the local administrative organs.

In 1972 the Southern Region, comprised of Bahr el-Ghazal, Equatoria, and Upper Nile provinces, was established with its capital at Jūbā. It was granted limited self-government within the national structure. The People's Regional Assembly was created as the legislative body, its members elected by direct, secret ballot. The High Executive Council, appointed by the president, was formed as the administrative body and was charged with the establishment of a civil service in the south. Although the agreement establishing the southern government was ratified by representatives of both the north and south, some of its provisions were disputed by some southern guerrilla leaders.

All political parties were banned in 1969. Two years later, however, the Sudan Socialist Union was established as the country's only party. Its structure is comprised of a National Congress, Executive Committee, and Supreme Polit Bureau on the national level (all three chaired by the president); provincial assemblies; and basic units in the villages and towns. Ancillary organizations include the May Youth Corps and the Socialist Sudanese Women's Association. Membership to the union is open to all those over 18 years of age who are loyal to the revolutionary government.

Civil justice is administered through the High Court of Justice in Khartoum, province courts, and district judges. There is a body of civil law; where there is no special enactment, cases are decided according to "justice, equity, and good conscience." The criminal code is based upon the Indian Penal Code and is dispensed in major, minor, and magistrate's courts. There are also people's courts that try a substantial proportion of civil and criminal cases in accordance with popular custom. Muslims are subject to the Sharīʿah, the religious law of Islām in questions of inheritance, marriage, divorce, family relationship, and charitable trusts.

The armed forces have been greatly expanded since 1969, and the military's political role has been defined through the Sudan Socialist Union. In the early 1970s, the forces consisted of a 35,000-man army, a 600-man navy, and a 1,500-man air force. There were also 3,000 men in the paramilitary national guard and border police. As part of the agreement ending the civil war in 1972, a southern command of the army was to be formed; it would consist of 12,000 men, one-half of whom would be from the southern provinces.

Educational services

Administrative services. In the Muslim areas of the north, boys have long been instructed in religious subjects according to traditional methods. Primary education was begun in the northern Sudan after 1898, and secondary education commenced in 1913. Postsecondary schools in arts and science, agriculture, veterinary science, engineering, and law were established between 1938 and 1941. In 1947 these schools were brought together within Gordon Memorial College, which was linked in special relationship with the University of London until 1959. Education in medicine began in 1924 with the foundation of the Kitchener School of Medicine, which in 1951 was united with Gordon Memorial College in the University College of Khartoum. In 1956 University College became the University of Khartoum.

In the south, Christian missionaries were responsible for education until 1957, when the government assumed direct control. As a result, Arabic was given greater prominence in the curriculum, but education was to suffer severely from the civil war. There have been no formal schools since they were closed in 1962. During the war, education was offered in English and vernacular languages in bush schools that operated in areas controlled by the Anya-nya, the rebel army.

In 1970 the government instituted a new four-tiered educational system. It consists of a six-year primary level, a three-year general secondary level, a three-year level of either higher secondary or vocational, technical, and teacher-training programs, and three or more years of higher education. In addition, the government has instituted a special program to combat the country's high illiteracy rate of 81 percent in the early 1970s. The *maʿhad ʿilmī* in the mosque at Omdurman, attended by pupils from

Muslim areas, provides theological education and trains religious leaders (*'ulamā'*) and lawyers (*qāḍīs*). The principal centre for the training of intermediate school teachers is the Institute of Education at Bakht ar-Ruda near ad-Duwaym for men and at Omdurman for women. Secondary school teachers are trained at a Higher Institute for Teacher Training at Omdurman.

The University of Khartoum has faculties of agriculture, arts, science, economics and social studies, engineering and architecture, law, medicine, veterinary science, pharmacy, and extramural studies. English is the medium of instruction in the university and the secondary schools, but it is being replaced by Arabic in the latter. A branch of Cairo University in Khartoum provides evening classes in arts, law, commerce, and economy that are conducted in Arabic. The Khartoum Polytechnic Institute was founded in 1950.

Early in the 20th century The Sudan was reputed to be unhealthy, and epidemics of sleeping sickness, cerebrospinal meningitis, and yellow fever were frequent. The progress of medical and sanitary services under the Sudan Medical Service (later the Ministry of Health), supplemented in many places by missionary institutions, has reduced mortality. By the early 1970s there were almost 100 government hospitals of different grades and about 100 health centres located in all important areas. Rural areas were served by more than 1,600 dispensaries and dressing stations under medical assistants. There are doctors in private practice in the larger towns and a rural midwife service.

Towns are planned so that there are separate first-, second-, and third-class residential areas, as well as commercial and industrial zones. Buildings must be erected in conformity with regulations such as those governing the materials to be used and the number of stories. Outside the limits of towns are *deims* (Arabic: *admās*), where unregulated houses may be erected. In rural areas where the population is sedentary, housing conforms to tribal customs, while nomads occupy tents.

The Sudan Police Force, under the authority of the Ministry of the Interior, is responsible for the policing of the entire country. Its duties include internal security, criminal investigation, and the administration of passport migration, nationality, and alien registration. There are also several police reserve companies that are specially trained on quasi-military lines.

Social conditions. In 1964 more than half of the total labour force was engaged in agriculture, forestry, and fishing. Income is higher in urban and semi-urban areas than in the larger rural sector. In industry, commerce, and public administration there is legislation to cover employment exchanges, conditions of employment in factories and in domestic service, and workmen's compensation. Under legislation applying to cooperative societies, many such bodies have been registered, of which about 150 are agricultural and more than 100 are for marketing and credit.

Social and economic divisions

The greatest social and economic divisions are those between the north and the three southern provinces, which can be traced to British administrative policy under the Anglo-Egyptian Condominium of 1899 to 1955. Until 1946 the British government administered the southern provinces separately from the north because of the ethnic and linguistic differences between the peoples of the two areas. Because it was of easier access, the Islāmic north underwent economic and educational advances, while the south continued its traditional existence. Christian missionaries were allowed to operate in the south, where Arab traders and educators were banned. After 1946 this policy was amended and the two areas were treated as one unit; independence was granted in 1956 to the whole region as one nation.

The creation of one political state has not erased the historic differences between the regions. The southern peoples have long feared the Muslim north, and there was guerrilla activity against the national government from 1955 until the signing of a ceasefire in 1972. The war wreaked havoc in the south, and between 500,000 and 1,000,000 southerners were killed in the fighting. The rebel movement was divided between those who desired regional autonomy, federation, or independence, and it was feared that the agreements of 1972 would not be wholly acceptable to the rebel leaders.

CULTURAL LIFE

The Sudan is part of both Arab and black Africa and has a dual cultural heritage. The way of life is similar in both areas—most of the population in the north and the south are pastoralists—but history has exacerbated differences of race, religion, and language. The Islāmic Arabs were long slave traders who preyed upon the animist southern tribes. Fear of the northerners as slavers has been replaced by the fear of cultural assimilation. Although denied by the government, it is believed by many southerners that Islām and Arabic will be forced upon the southern tribes in a drive to create a wholly Arab state. If these suspicions can be ameliorated by the formation of the semi-independent Southern Region, it will be possible to create a nation in which both cultural traditions thrive.

Cultural institutions. Most cultural institutions are centred in Khartoum, where learned societies such as the Antiquities Service, the Educational Documentation Centre, and the Industrial Research Institute—as well as the British Council and the Soviet Cultural Centre—are located. Libraries in Khartoum are connected to educational institutes, and there is a public library in Omdurman and a research library of the Ministry of Agriculture in Wad Madanī. Museums include the Ethnographical Museum and the Sudan Museum in Khartoum, the Marawī Museum, with exhibits of antiquities in Marawī (Merowe), the archaeological Shaykhan Museum in al-Ubbayyiḍ, and the Sudan Natural History Museum, affiliated with the University of Khartoum.

Press and broadcasting. The government-controlled Sudan Broadcasting Service broadcasts daily radio programs in Arabic and English. Programs intended for the southern provinces are aired in the vernacular languages. A weekly half-hour program is also broadcast in Somali, and it is planned to greatly extend external broadcasting by 1975. The Sudan Television Service offers daily service to the Three Towns area from its station at Omdurman. In the early 1970s service was being extended north to 'Aṭbarah and south to the entire Gezira region.

The press was nationalized in 1970 and its ownership and supervision transferred to the Sudan Socialist Union. There are two Arabic daily newspapers, two Arabic weeklies, and one English weekly. The Ministry of Information publishes a weekly broadcasting magazine, a monthly cultural magazine, and daily bulletins for the press and embassies. There is also a weekly magazine aimed at youth movement.

PROSPECTS FOR THE FUTURE

The Sudan's future economic development is dependent upon cotton. The great extent of the country has also need for a resolution of internal political problems, which have diverted energies and resources needed for development, and the necessity for evolving a more balanced economy by reducing the degree of dependence upon cotton. The great extent of the country has also inhibited the rapid development of an effective transportation network. It is hoped, however, that in the years ahead the transportation problem will progressively be overcome. For example, if The Sudan's railway system should be linked with that of Nigeria through Chad, a trans-African railroad running from Lagos in Nigeria to Port Sudan on the Red Sea would come into existence.

BIBLIOGRAPHY. K.M. BARBOUR, *The Republic of the Sudan: A Regional Geography* (1961), a survey of the political and historical background of The Sudan, with chapters on relief, drainage, soils, vegetation, Nile hydrology, and human and economic geography; H.S. BLUNT, *Gum Arabic, with Special Reference to Its Production in the Sudan* (1926), a study of the botanical aspects of the gum tree, its distribution, exploitation, properties, and uses; AUDREY BUTT, *The Nilotes of the Sudan and Uganda* (1964), an account of tribal groupings, their distribution, physical environment, social conditions, political and economic structure, religious beliefs, cults, and art; O.G.S. CRAWFORD, *Castles and Churches*

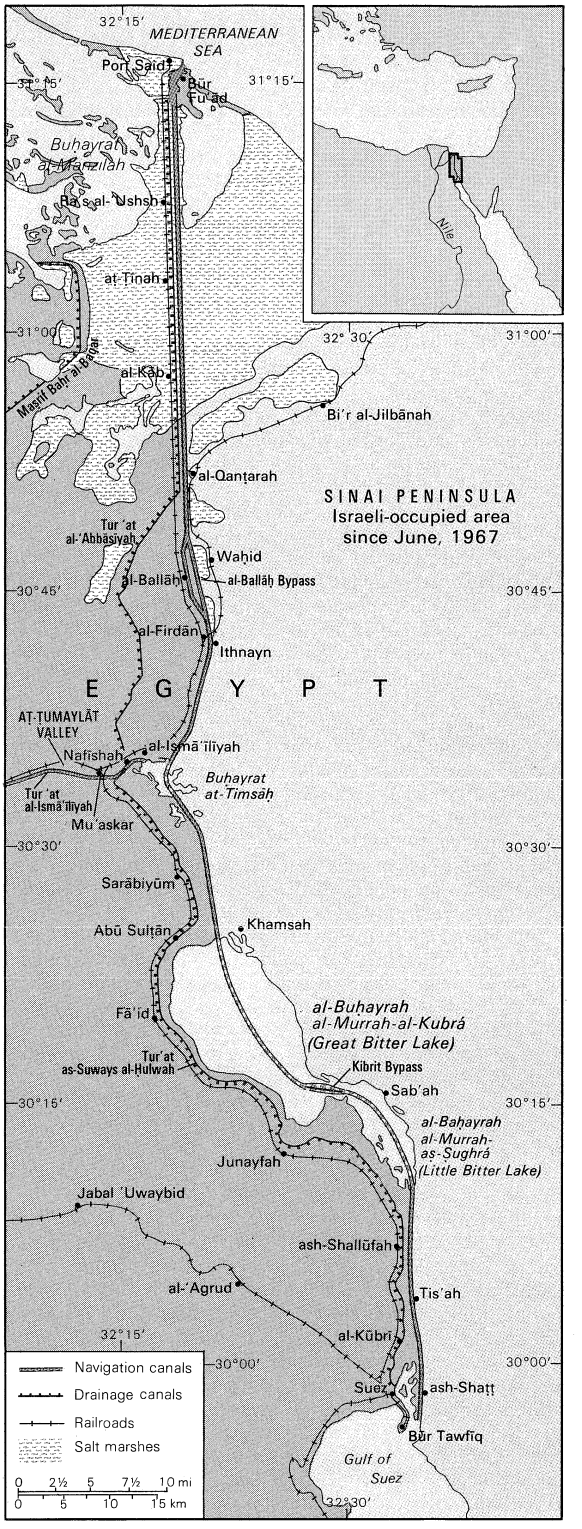
in the Middle Nile Region (1953), a survey of the ancient sites of 37 castles and churches between Atbara and Abu Hamed, with an added note on the inscriptions by M.F. LAMING MACADAM; J.K. JACKSON and M.N. HARRISON, "Ecological Classification of the Vegetation of the Sudan," *Khartoum, Forests Department Bulletin*, n.s., no. 2 (1958), a survey of rainfall and soils, followed by a detailed description and classification of the vegetation; J.H.G. LEBON, *Land Use in Sudan* (1965), a regional monograph that contains a series of maps; C.G. and B.Z. SELIGMAN, *Pagan Tribes of the Nilotic Sudan* (1932, reprinted 1965), a comprehensive ethnographic survey of the tribes of the Southern Sudan; MINISTRY OF INFORMATION, KHARTOUM, *Sudan Almanac: An Official Handbook* (annual), a valuable source of information supported by reliable statistics; *Sudan Notes and Records* (semi-annual since 1918), the leading journal exclusively devoted to Sudanese studies—particularly significant in the humanities and social sciences; J.D. TOTHILL (ed.), *Agriculture in the Sudan, Being a Handbook of Agriculture As Practised in the Anglo-Egyptian Sudan* (1948), an authoritative but nontechnical reference work containing background chapters on geology, soils, climate, and irrigation; J.S. TRIMINGHAM, *Islam in the Sudan* (1949, reprinted 1965), a description of the religion of the muslim Sudanese, covering beliefs and practices, religious orders, and other aspects; A.J. WHITEMAN, *The Geology of the Sudan Republic* (1971), a comprehensive and systematic account of The Sudan's geology.

M.e.D.S.)

Suez Canal

The Suez Canal (Qanāt as-Suways) a sea-level waterway across the Isthmus of Suez in the Arab Republic of Egypt connecting the Mediterranean and the Red seas, for almost 100 years until its closing by the Arab-Israeli War of June 1967, provided the shortest maritime route between Europe and the lands lying around the Indian and western Pacific oceans and was one of the world's most heavily used shipping lanes. The canal extends 100 miles (160 kilometres) between Port Said (Būr Saʿīd) in the north and Suez in the south, with dredged approach channels north of Port Said into the Mediterranean, and south of Suez. The canal does not take the shortest route across the Isthmus, which is only 75 miles, but utilizes three intervening lakes, from north to south, Buḥayrat al-Manzilah (Lake Menzala), Buḥayrat at-Timsāh (Lake Timseh), and al-Buḥayrāt al-Murrah (the Bitter Lakes). It is an open cut, without locks, and, though extensive straight lengths occur, there are eight major bends. To the west of the canal lies the low-lying delta of the Nile; to the east is the higher, rugged, and arid Sinai Peninsula. Prior to construction of the canal, completed in 1869, the only important pre-existing settlement was Suez, which in 1859 had about 3,000 inhabitants, the rest of the towns along its banks having grown up since, with the possible exception of al-Qanṭarah, which was a tiny village. The name Suez is held to derive from Qolzuom, the pharaonic name of the settlement. This became Arabicized into Suways, originally translated by the French as Souyès. Classical writers, Herodotus especially, rendered the original name as Clysma.

Physical description. *Geology.* The Isthmus of Suez, the sole land bridge between Africa and Asia, is of recent geological origin. Both continents, it would appear, once formed a single large continental mass, but during Tertiary times (about 2,500,000 to 65,000,000 years ago) the great fault structures of the Red Sea and Gulf of Aqaba developed, with the subsequent drowning of the Red Sea trough and its prolongation as far as the Gulf of Suez. In the succeeding Quaternary Era, there was considerable oscillation of sea level leading to marine erosion in the southeast Mediterranean and, finally, transgression by the Mediterranean far to the south, thus linking the two seas. Later uplift led to the emergence of a low-lying isthmus bounded to the east and southwest by higher plateau formations and narrow in the south but broadening northwards to a very low-lying open coastal plain. There, the Nile Delta once extended further east (as a result of periods of abundant rainfall coincident with the Pleistocene Epoch, or Ice Age), and two former river arms, or distributaries, crossed the northern isthmus, one branch reaching the Mediterranean at the narrowest



The Suez Canal.

point of the Isthmus and the other entering the sea some nine miles east of present Port Said. Topographically, the Isthmus is not uniform; there are three shallow, water-filled depressions—al-Manzilah, at-Timsāh, and the Bitter Lakes, these last distinguished as Great and Little though they form one continuous sheet of water. Between these lake depressions are east–west aligned ridges: Jisr al-Jīr north of at-Timsāh, Serapium between at-Timsāh and the Bitter Lakes, and Jabal ash-Shallūfah between Little Bitter Lake and the Gulf of Suez. A number of more resistant bands of limestone and gypsum obtrude in the south of the Isthmus, and another significant feature

is a narrow valley (aṭ-Ṭumaylāt) leading from Buḥayrat at-Timsāh southwestward toward the middle Nile Delta and Cairo. The Isthmus is composed of marine sediments, coarser sands, and gravels deposited in the early periods of abundant rainfall, Nile alluvium (especially to the north), and wind-blown sands.

Original size of the canal

The canal. When first opened in 1869, the canal consisted of a channel with a depth of barely 20 feet (six metres), 72 feet (22 metres) wide at the bottom, and 190 feet (58 metres) wide at the surface. Passing bays, 1,000 to 1,500 feet long and 16 feet wide were built, as far as possible, every 16 miles. Construction involved the excavation and dredging of 97,000,000 cubic yards of sediments. Between 1870 and 1884 some 3,000 groundings of ships had occurred as results of the narrowness and tortuousness of the channel. Major improvements began in 1876, and, after successive widenings and deepenings at the time the canal was closed by the Arab–Israeli war in June 1967, it had a minimum width of 179 feet and an uninterrupted depth of almost 40 feet at low tide; passing bays had been greatly enlarged and new bays constructed; bypasses were made in the Bitter Lakes and at al-Ballāḥ; stone or cement cladding and steel piling for bank protection had been almost entirely completed in areas particularly liable to erosion; tanker anchorages were deepened in Buḥayrat at-Timsāh; the new berths were dug at Port Said to facilitate grouping of ships in convoy. During 1964 plans had been made to deepen the canal to 60 feet and double its width by 1975, but these, never realized, have since been supplanted by further plans to accommodate supertankers (see below *Present situation and future prospects*).

Operation. In 1870, the canal's first full year of operation, there were 486 transits, or fewer than two per day. In 1966–67 there were 20,326, an average of 56 per day, with net tonnage increasing from 437,000 (1870) to 267,000,000. Originally, passing involved one ship entering a siding and stopping, but after 1947 a system of convoys was adopted. Transit time at first averaged 40 hours; by 1939 it was 13, but as traffic increased after 1942 it went up to 15 hours in 1967 despite convoying, suggesting that the canal was operating at near capacity. Convoys left twice daily, passing at Kabrit and al-Ballāḥ. On entering at Port Said or Suez a ship was assessed for tonnage and cargo (passengers were not charged for after 1950) and handled by three pilots, one for the approach roadstead and two for canal transit with changeover at al-Ismāʿīlīyah. Over 300 pilots of 15 to 20 nationalities were employed.

The nature of traffic greatly altered, the chief feature being the enormous growth of oil movement from the Persian Gulf after 1950. In 1913 oil and petroleum by-products in northbound traffic amounted to 291,000 tons; in 1966 to 166,000,000 tons. Passenger traffic, from an all time peak of 984,000 in 1945, declined considerably due to competition from aircraft. Northbound cargoes consisted of petroleum, ores and crude metals, bulk foodstuffs (chiefly grain), and textile fibres (cotton and wool). Southbound cargoes were chiefly manufactured goods, machinery, vehicles, and fertilizers. Of cargo shipped through the canal after 1960, 82 percent was northbound, 17 percent southbound (the rest going to Egypt), with oil accounting for over 70 percent of the total. Much southbound traffic was of empty tankers, and because less than half rates were charged for ships in ballast there were two problems—the obvious one of revenue, and the other of differing speeds as between empty and full vessels, which controlled convoy procedure. In 1913 exactly 60 percent of canal shipping was registered in Great Britain, followed by The Netherlands (6 percent) and France (5 percent). In 1966–67 the chief flags were United Kingdom (16 percent), Liberia (13 percent) Norway (11 percent), Greece (8 percent), and the Soviet Union (7 percent).

Traffic and cargoes

Communications and towns. Construction of the canal led to the growth of settlements in what had been, except for Suez, almost uninhabited arid territory. Over 70,000 acres have been brought under cultivation, and about 8 percent of the total population was engaged in agriculture, with approximately 10,000 commercial and industrial activities of various sizes. Settlements of importance are Port Said (population [1966 census] 283,000), with its east-bank counterpart, Būr Fu'ād; al-Ismāʿīlīyah (Ismailia) (population [1966 census] 144,000), on the north shore of Buḥayrat at-Timsāh with a westward extension, Muʿaskar (Moascar); and Suez (population [1966 census] 264,000), with its west-bank outport, Būr Tawfīq. Most of the population has been evacuated from the canal area since 1967.

There are two roads on the west bank: one, originally a company service road, follows closely the canal bank; the other, the Treaty Road, part of the state system, runs between Suez and al-Ismāʿīlīyah, then northward toward Port Said with ferries at al-Kūbrī and at al-Firdān to al-Qanṭarah. From this last, a road continues along the east bank to Būr Fu'ād, and another eastward through the Sinai to Israel. A major road runs westward from al-Ismāʿīlīyah through the aṭ-Ṭumaylāt valley, and there are direct routes from Suez to Cairo and from Port Said to Cairo via the delta. Military exigencies since 1967 have led to changes in the road pattern throughout the entire area. Rail connections, part of the Egyptian state system, existed on both banks of the canal. A swing bridge just south of al-Firdān carried the main line into Sinai.

History. *Construction.* The first canal in the region seems to have been dug *c.* 1800 BC, when an irrigation channel navigable at flood period was constructed into the Wadi Tumelat (Wādī aṭ-Ṭumaylāt). This was extended by the Ptolemys via the Bitter Lakes as far as the Red Sea. From the region of Buḥayrat at-Timsāh a northward arm appears to have reached a former branch of the Nile. Extended under the Romans, when it was known as Trajan's Canal, neglected by the Byzantines, and reopened by the early Arabs, this canal was deliberately filled in for military reasons during AD 775. Throughout, the aim appears to have been to facilitate trade from the delta lands to the Red Sea rather than to provide a passage to the Mediterranean.

Ideas for a direct Suez Canal were put forward by the Venetians in the 15th century and at subsequent intervals, but all encountered Ottoman hostility. In 1798 Napoleon personally investigated the remains of the ancient canal and ordered a survey for a direct canal, which was carried out in 1799. It was concluded that the level of the Red Sea was 33 feet above that of the Mediterranean, and, therefore, locks would be needed. Studies of a possible canal were made in 1834 and again in 1846. In 1854 Ferdinand de Lesseps received an Act of Concession from the viceroy (khedive) of Egypt, Muḥammad Saʿīd, to construct a canal; and in 1856 a second act conferred on the Compagnie Universelle du Canal Maritime de Suez (the Suez Canal Company) the right to operate a maritime canal for 99 years after completion of the work. Construction began in 1859 and took ten years instead of the six envisaged; climatic difficulties, a cholera epidemic in 1865, and early labour troubles all slowed down operations. An initial project was the cutting of a small canal from the delta along the Wadi Tumelat with a southern branch (the Turʿat as-Suways al-Ḥulwah [Sweet Water Canal]) to Suez and a northern one (al-ʿAbbāsīyah canal) to Port Said. This supplied drinking water in an otherwise arid area and was completed in 1863.

The Suez Canal Company

At first, hand digging with picks and baskets was used, fellahin being drafted as forced labour. Later, dredgers and steam shovels with European labour took over, and, as dredging proved cheaper than dry excavation, the terrain was artificially flooded and dredged wherever possible. Other than in the few areas where rock strata were met, the entire canal was driven through sand or alluvium. In August 1869 the waterway was complete and was officially opened with an elaborate ceremony in November 1869.

Finance. The Suez Canal Company had been incorporated as an Egyptian joint stock company with its head office in Paris. Despite much early official coolness, even hostility, on the part of Great Britain, de Lesseps was anxious for international participation and offered shares widely. Only the French responded, however, and the

remainder was largely taken up by the khedive, who came to own 44 percent as compared with 52 percent by French interests. The first board of directors included representatives of 14 countries.

In 1875 financial troubles compelled the khedive, Ismāʿīl, to sell his holding, which (at the instigation of the prime minister, Benjamin Disraeli) was at once bought by the British government. Until 1875 the shares remained below their issue price, but with the British purchase (at 568 francs for a Fr. 500 share) steady appreciation took place, from a low point of Fr. 215 in 1870 to over Fr. 3,600 in 1900.

By the 1950s the board of directors numbered 32, representing France, Great Britain, The Netherlands, the United States, and eventually Egypt. Originally allocated 15 percent of profits, Egypt later relinquished these and, after the sale of Ismāʿīl's 176,602 shares, remained unrepresented until 1949 when it was, in effect, reinstated as a board member and allotted 7 percent of gross profits. From 1936 onward Egyptian participation in the organization and working of the canal increased. Besides more directors it was agreed in 1949 that 90 percent of new clerical jobs and 80 percent of technical appointments would be offered to Egyptians. The Canal Company also providing hospitals, schools, and other amenities.

In 1956, 13 years before the concession was due to expire, the canal was nationalized by Pres. Gamal Abdel Nasser. Since that time the Egyptian government has exercised complete control, though the original company continues in Paris as a conglomerate.

International status. Although the canal was built to serve, and profit from, international trade, its international status remained undefined for many years. A convention was finally signed in 1888 making the canal open to all ships of all nations both in peace and war, acts of hostility in its waters being forbidden. Britain did not join the convention until 1904.

The history of international use of the canal during wartime included denial of passage by Spanish warships during the Spanish–American War of 1898; passage by a squadron of the Russian Navy during the Russo-Japanese War in 1905; and passage of Italian vessels during the Italo-Turkish War of 1911–12. Theoretically, the canal was open to all belligerents during World Wars I and II, but the naval and military superiority of the Allied forces denied effective use of the canal to enemy shipping.

Following the armistice between Israel and its Arab opponents in 1949, Egypt denied use of the canal to Israel and to all ships trading with Israel. The Arab–Israeli War in June 1967 closed the canal to all shipping.

Present situation and future prospects. Since June 1967 the canal has remained closed, and significant new trading patterns have emerged. Most precious cargoes of small bulk are now increasingly routed by air, as are a large majority of passengers, with the rest routed around Africa via the Cape of Good Hope. Containerization has partly offset the resultant higher costs. New sources of crude petroleum have been developed closer to the markets of Europe (Algeria, Libya, Nigeria), while Kuwait has increasingly close links with Japan, its principal trade partner. Much of the rerouted oil traffic is now by giant tanker; with the restraints of canal size removed, vessels of 200,000–300,000 tons are in operation, with supertankers of 500,000 tons planned. Enlargement of the canal will be a necessity, since 73 percent of revenues came from oil.

The challenge of containers, supertankers, and pipelines

Reopening would involve clearance, repairs, reinstallation of equipment, re-establishment of the operation teams (pilots and other personnel), and, inexorably, enlargement. Besides 15 ships trapped in the canal and now written off by their owners (but easily movable), ten more were sunk in the southern portion, and there is debris (shattered equipment) in the north. Repairs to portions of the banks damaged by military action will be necessary, although little silting is believed to have occurred. Plans in 1971 called for, in effect, the construction of a new canal by setting back the eastern bank by 328 feet and dredging to a depth of 66 feet. This could possibly be completed within four to five years, allowing vessels of 100,000 tons (the break-even point in competition with the Cape route) to use the canal within one year. Some regard these estimates as optimistic.

Another important factor is an oil pipeline between Suez and the Mediterranean at Alexandria, planned for completion in the 1970s with a capacity of 80,000,000 tons per annum. This, with the Israeli pipeline from Elat (now used by Iran to reach the markets of Europe), could provide formidable competition to an enlarged canal.

Nevertheless, the immense significance of the canal remains. In the 1960s it handled 15 percent of world sea traffic (up to 40 percent for India and Pakistan, 20 percent for Australia and Great Britain). Some 50 percent of total proved oil reserves occur in the three Persian Gulf states of Iran, Kuwait, and Saudi Arabia; and there would be a considerable political significance in a reopened canal allowing easier access by the Soviet Union to the Indian Ocean. The utility of the canal is also shown by the accompanying table.

Distances Between Selected Principal Ports

	by cape	by canal	saving (percentage)
London or Liverpool to			
Bombay	10,800	6,300	42
Kuwait	11,300	6,500	42
Melbourne	12,200	11,000	10
Calcutta	11,700	7,900	32
Singapore	11,800	8,300	30
Marseille to			
Bombay	10,400	4,600	56
Melbourne	11,900	9,400	21
New York to			
Bombay	11,800	8,200	31
Singapore	12,500	10,200	18
Ras Tanura (Saudi Arabia)	11,900	8,300	30

BIBLIOGRAPHY. A. ANDERSON, *Observations on the Practicability and Utility of Opening Communication Between the Red Sea and the Mediterranean by a Ship-Canal* (1843), a primary source by a canal pioneer on the technical feasibility of a sea-level canal linking Suez with the Mediterranean; B.B. BROMS, *The Legal Status of the Suez Canal* (1961), a recent examination of the various positions on the canal's legal status; JULES CHARLES-ROUX, *L'Isthme et le canal de Suez*, 2 vol. (1901), reprints of selected official documents with an extensive bibliography and maps relating to the canal; F.M. DE LESSEPS, *Percement de l'isthme de Suez: exposé et documents officiels*, 6 vol. (1855–66), contemporary history, memoir, and documents by the French engineer responsible for the canal concession and formation of the canal company; D.A. FARNIE, *East and West of Suez: The Suez Canal in History, 1854–1956* (1969), near definitive volume tracing events from the concession to the nationalization of the canal company, with extensive statistics and bibliography; E. LAUTERPACHT (ed.), *The Suez Canal Settlement* (1960), a selection of French, British, Egyptian, United Nations, and canal company documents covering nationalization, closure, re-opening, and settlement between 1956–59; J.M. LEPERE, "Mémoire sur la communication de la mer des Indes à la Méditerranée par la mer Rouge et l'isthme de Soueys," *Description de l'Egypte: État moderne*, vol. 1 (1809), initial survey by French engineer on the possibility of a waterway across the Isthmus of Suez; KENNETH LOVE, *Suez: The Twice-Fought War* (1969), highly readable, well-researched volume on the historical events leading to the nationalization of the canal company and the succeeding events up to 1967; JOHN MARLOWE, *World Ditch: The Making of the Suez Canal* (1964), soundly documented presentation from the canal's inception up to the 1882 British occupation of Egypt; J.A. OBIETA, *The International Status of the Suez Canal* (1960), gives legal aspects of, and claims to, the Suez Canal; H.J. SCHONFIELD, *The Suez Canal in Peace and War, 1869–1969* (1969), a brief general survey of the canal, together with some documents; F.P. VOISIN, *Le Canal de Suez*, 6 vol. (1902–06), primary source and contemporary history by the canal company's chief administrator during the waterway's construction; A.T. WILSON, *The Suez Canal: Its Past, Present, and Future*, 2nd ed. (1939), classic evaluation of the British interest in the canal based upon primary sources with a useful collection of documents.

(W.B.Fi.)

Sugar Production

Sugar is a sweet, crystalline substance, white or colourless when pure, derived from the juices of various plants, chiefly the sugarcane and the sugar beet but also the date palm and the maple tree. It is an article of human food and a sweetener for other foods, especially beverages, confectioneries, and canned goods.

Chemically, sugar is the organic compound sucrose, one of the carbohydrates and the prime source of energy in the human diet. Sucrose furnishes 712 BTU (British thermal units) of energy per pound (394 kilocalories per kilogram) to maintain body temperature and activity. Though chemists designate a large group of related compounds as sugars, in this article sugar refers specifically to sucrose. (For details on physical and chemical properties of sucrose and other sugars, see the article CARBOHYDRATE.)

Early sugar making in India

Sugar making was known in India as early as 3000 BC. *Gaura*, an old Indian word for sugar, was apparently derived from the name *Gur*, an ancient monarchy in Bengal. Domestic noncentrifugal sugar, sugar that is made without centrifuging for home consumption, is now called *gur* in India and *goela* in Java. A crown made of sugarcane is described in the Atharvaveda, a sacred book of the Hindus, written about 800 BC or earlier, and the Greek general Nearchus, who accompanied Alexander the Great to India in the 4th century BC, tells of a reed that produced "honey" without the aid of bees.

Sugarcane is believed to have originated independently in the Solomon Islands, in the South Pacific, because of a local legend that the ancestors of the human race were generated from a cane stalk: one bud stretched and grew, abruptly changing into a man, while another bud stretched and grew, abruptly changing into a woman.

The methods of growing cane and making sugar were propagated from India east to Indochina and west to the Arabian countries and Europe. In Sanskrit, the ancient language of India, the word for sugar is *śārkara* or *śākkara*, originally sand or gravel. This became *sukkar* in Arabic, *sákharon* in Greek, *zucchero* in Italian, *sucre* in French, and sugar in English. The changes in the words suggest the route of the propagation of the art.

By the 8th century, sugarcane was grown and sugar made in Arabic Spain and southern France. Venice was long the chief centre of sugar refining in Europe, supplanted in the 16th century by Antwerp. In 1544 two refineries were established in London and in 1689 one in New York.

In 1493, on his second voyage, Christopher Columbus brought sugarcane to the New World. The success of sugar plantations in Santo Domingo led to its cultivation in other islands of the West Indies. In 1737 planting was begun in Louisiana, but not until 1791 did it succeed.

Discovery of beet sugar

The beet grew wild in parts of Asia and was cultivated at an early time in southern Europe as well as in Egypt. It was long used as a garden vegetable and cattle fodder. In 1747 Andreas Marggraf, a German chemist, proved that various kinds of beet root, which tasted sweet, contained sugar and that it could be extracted in crystalline form. No one, however, took particular interest in his announcement until Franz Karl Achard, one of his pupils, planted beets on his estate, Caulsdorf, near Berlin, in 1786, in the hope of finding a species especially rich in sugar. Frederick William III, the king of Prussia, interested himself in Achard's work and provided for the world's first beet-sugar factory, which began operations in 1802 at Cunern, Silesia. By 1810 it was evident that beet sugar could be the basis for a successful industry.

In France, Benjamin Delessert made a quantity of well-crystallized beet sugar in his small factory at Passy in 1811, for which he was awarded the Cross of the Legion of Honour by Napoleon.

In the United States, beet sugar was first produced in 1838 at Northampton, Massachusetts, by Edward Church and David Lee Child.

Sugar was initially dispensed medicinally as a sedative; as a food or food sweetener it was a scarce luxury until the 17th century. It remained expensive as late as 1736, when it was listed with precious stones among the wedding gifts of Maria Theresa, future queen of Hungary.

CANE-SUGAR PRODUCTION

Sugarcane. Sugarcane is a perennial grass of the genus *Saccharum*. Most of present-day commercial canes are the offsprings or hybrids of the species *Saccharum officinarum*, which was developed from a wild cane species *Saccharum robustom* and cultivated by natives of southern Pacific Islands.

Culture. The cane-growing areas of the world are located from 37° N (south Spain) to 30° S (Natal, South Africa), but most of the world's cane sugar is grown in subtropical and especially in tropical areas.

The sugarcane plant produces a number of stalks that reach from 10 to 24 feet high and bear sword-shaped, long leaves. The stalks are composed of many segments and at each joint there is a bud. When the cane becomes mature, a growing point at the upper end of the stalk develops into a slender arrow, bearing a tassel of tiny flowers.

Sugarcane is propagated primarily by the planting of cuttings. Sections of the stalk of immature cane used for planting are known as seed cane, or cane sets, and have two or more buds (eyes), usually three. Seed canes are planted in well-worked fields. Mechanical planters that open the furrow, fertilize, drop the seed cane, and cover it with soil, are widely used.

Seed canes are spaced from 4½ feet to 6 feet apart at densities from 4,000 to 10,000 per acre. Under favourable conditions, each bud germinates and produces a primary shoot. Root bands adjacent to each bud give rise to a large number of roots, and each young shoot develops its own root system. Tillering, or sprouting at the base of the plant, takes place; and each original seed cane develops into a number of growing canes, forming a stool. The plant crop is obtained from these stools.

Ratooning

Another method of cane propagation is by ratooning, in which, when the cane is harvested, a portion of stalk is left underground to give rise to a succeeding growth of cane, the ratoon or stubble crop. The ratooning process is usually repeated three times so that three economical crops are taken from one original planting. The yield of ratoon crops decreases after each cycle, and at the end of the last economical cycle all stumps are plowed out and the field is replanted.

Sugarcane is grown in various kinds of soils, such as red volcanic soils and alluvial soils of rivers. The ideal soil is a mixture of sand, silt, and clay particles, with a measure of organic material. The land is plowed and left to weather for a time, before subsoiling, (stirring up the subsoil) is carried out.

Sugarcane demands a well-drained soil, so that its roots have access to air and can absorb it. Drains, either surface, underground, or both are provided, according to the topographic conditions of cane fields.

To attain good yields, sugarcane requires 80 to 90 inches of rain (2,000–2,300 millimetres) during the growing period. When precipitation is deficient, irrigation, either by spraying or by applying water in furrows, can make up for the deficiency.

Fertilizers are applied to sugarcane from the beginning of planting through the whole growth cycle, but not during the ripening period. Optimum amounts of fertilizers (nitrogen, phosphorus, and potassium) vary greatly with soil types, climatic conditions, and the kind and length of the growing cycle.

To secure a good cane crop, weeds in the cane fields must be attacked until the cane stools develop a good canopy, which checks weed growth. Weeding, still largely manual, is done with a hoe, though mechanical cane weeders with attached rakes have been developed. Chemical weed killers and herbicides are widely used.

Maturation

The growth period for cane crops varies considerably according to the region: 8–9 months in Louisiana; 15 months in Australia and Taiwan; 18–22 months in Hawaii, South Africa, and Peru.

The lowest temperature for good cane-plant growth is about 68° F (20° C). Continuous cooler temperature promotes the maturation of cane, as does withholding water. Harvesting and milling begins in the dry, relatively cool season of the year and lasts for five to six months.

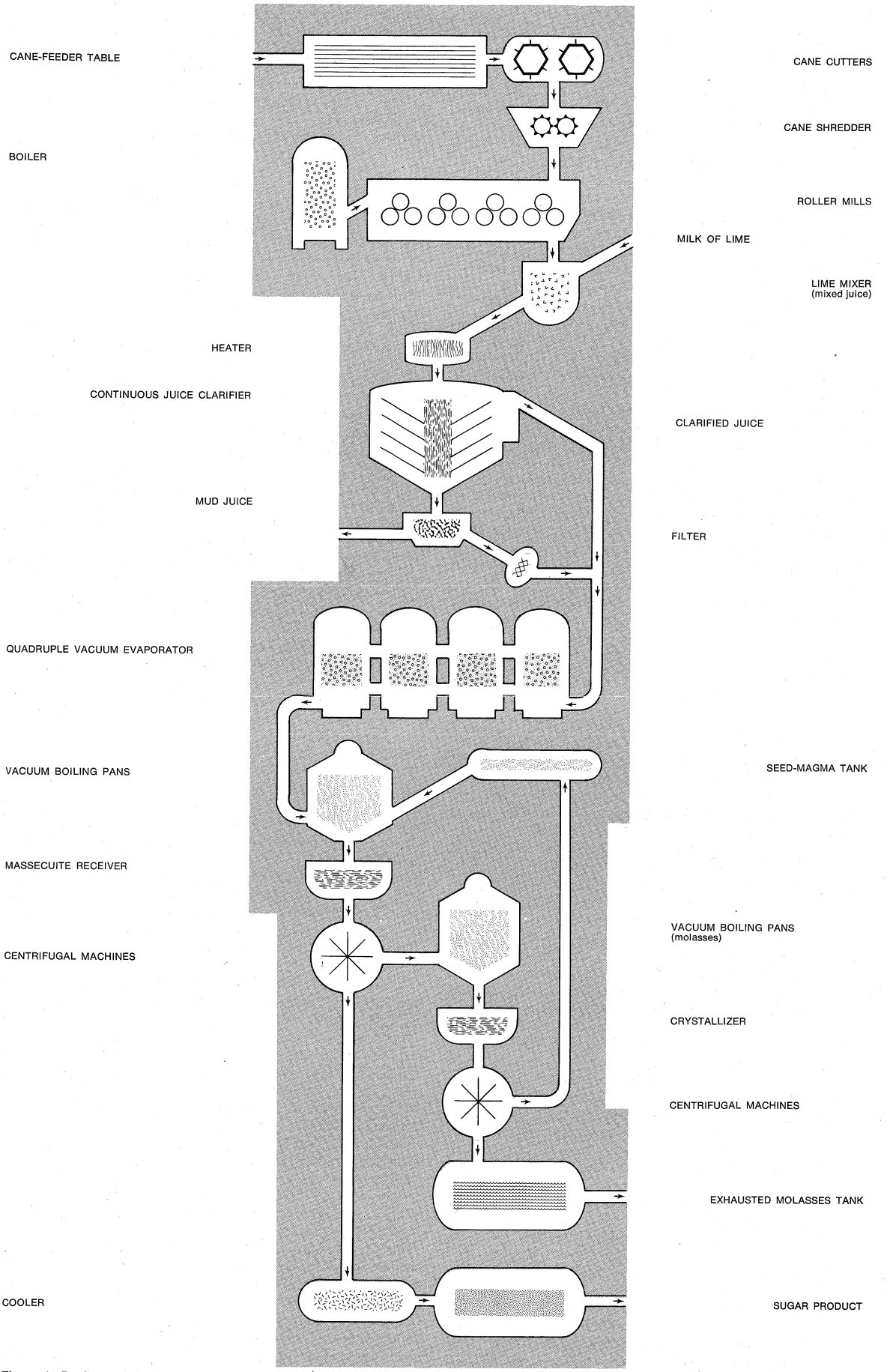


Figure 1: Basic steps in the raw-sugar manufacturing process.
Drawing by D. Meighan by courtesy of the Hatachi Zosen Co., Ltd., Tokyo

The mature cane is harvested by both manual and mechanical means. In Hawaii and Australia, mechanical harvesters are effectively used. Some are able to sever and discard the tops of erect crops and cut cane stalks, which are delivered into a bin trailer for transport to the mill by tractor or light railway wagon.

Diseases. The sugarcane plant is subject to many diseases. Sereh, a blackening and degeneration of the fanlike tops, is caused by an East Indian virus. Mosaic, mottling or spotting of foliage and sometimes curling, dwarfing, and narrowing of the leaves, may be caused by any of several viruses. Red rot (important in Java) is characterized by red patches within the canes, and is caused by the fungus *Physalospora tucumanensis.* Gumming disease (important in New South Wales, Australia) is characterized by gummosis, the pathological production of gummy exudates as a result of cell degeneration; it is caused by the bacterium *Xanthomonas vasculorum.* During the period 1880–93 the world's sugarcane production was severely damaged by these diseases, resulting in a temporary decline in the industry. In 1926 the spread of mosaic disease in Louisiana resulted in a drastic drop in sugar production to only one-quarter of the normal yearly average. Epidemics of these diseases have been checked by replacing the susceptible varieties of cane with varieties resistant to the disease.

Fiji disease

Fiji disease, a virus disease first reported from the Fiji islands, is characterized by elongated white to brown swellings on the underside of the leaves, followed by stunting and death. Early in the 20th century, Fiji disease caused such a severe loss as to menace the very existence of the industry in the Fiji islands. Leaf scald also caused serious damage in Australia and the Fijis during the first three decades of this century. Leaf scald is a vascular disease caused by the bacterium *Xanthomonas albilineans,* characterized by creamy or grayish streaking and later withering of the leaves. After World War II, ratoon-stunting disease was first recorded in Queensland as a specific virus disease of sugarcane, particularly ratoon crops. Recently the yield losses due to this disease in many cane-raising countries have been greater than those due to any other diseases.

Eyespot, characterized by yellowish oval lesions on leaves and stems, is a disease caused by the fungus *Helminthosporium sacchari.*

Pests. Sugarcane is attacked and damaged by various insect pests that bore into and feed on the different parts of the plant. Control measures include biological control by parasites or predators, chemical control by insecticides, and the introduction of resistant cane varieties.

The moth borer, *Diatraea saccharalis,* which is widely distributed throughout cane-growing areas, is capable of causing extensive damage when out of control. The sugarcane leafhopper and the anomala grub yielded to biological control in Hawaii when other measures were unsuccessful. Various predator animals live on insects destructive to the sugarcane. In Queensland the bandicoot, an insectivorous marsupial, is a most diligent destroyer of white grubs.

The insect pest responsible for the greatest crop losses in Queensland is the grayback beetle in its larval stage. The most effective grub control is obtained by applying the insecticide benzene hexachloride, after the young plant cane has germinated and stooled. Cane is also protected effectively against wireworms by applying insecticides when cane sets are planted. Rats, which destroy part of the stalk, are controlled by poisoning and trapping.

Breeding. The purpose of sugarcane breeding is to produce new hybrid varieties that will be immune, or resistant, to diseases and insect pests and will increase the production of sugar per unit area, yielding canes of higher sugar content and better fabrication qualities.

The first task of breeding is to obtain new cane seeds by crossing selected parent varieties, and then to select seedlings from the new seeds. The crossing is effected by enclosing in a cloth lantern two tassels from two different cane varieties selected as a male and a female parent. The commercially superior varieties are not necessarily ideal parents. Many of the best varieties were bred from parents unsuitable for commercial use. The production of such cane seeds and seedlings by crossbreeding has been established in Java and Barbados since the 1880s.

A selected seedling is planted and tested in the fields, and usually takes up to ten years before being released as a new commercial variety.

Most present-day hybrid commercial varieties are directly descended from the Cheribon cane, a Javan noble cane, species *Saccharum officinarium.* Noble canes, which represent the highest development of the species, are characterized by thick, barrel-shaped internodes, or segments, and large, soft-rinded, juicy stalks, and high sugar content. Original noble canes were susceptible to some serious cane diseases; their hybridization with wild canes was successful because, although wild cane *Saccharum spontaneum* contains little sugar, it is immune to most diseases.

Basic process

Raw cane-sugar manufacture. About 85 percent of the cane sugar produced in the cane-raising areas is raw sugar for further processing; the balance is plantation white sugar, for direct domestic consumption. The raw-sugar manufacturing process is illustrated in Figure 1.

Sugar juice in the cane is pressed out by means of a set of three-roll mills. The pressed juice, known as mixed juice, is purified by heating and adding milk of lime. This mixture is then sent to a continuous juice clarifier that separates the clarified juice and the muddy juice. The former is sent to an evaporator and the latter to a continuous-rotary vacuum filter. The filtrates are added to the clarified juice and the mixture concentrated in the evaporator. This raw-syrup mixture is further concentrated in vacuum boiling pans until sugar crystallizes out of the liquor, forming a mixture known as massecuite. The crystals (raw sugar) are separated from the massecuite by centrifugal machines.

Cane processing. Sugarcane is first weighed and then carried to cane-feeder tables.

Conventional milling of cane

The milling of cane by crushing and grinding is a highly developed process whose purpose is to rupture the plant cells by pressure to release the sugar-bearing juice. Machinery of massive construction is used because of the tough nature of the material to be ground. Multiple mills arranged in triangular form are universal. Rapidly revolving cane knives cut the cane into chips to level and compact the bed of cane on the carrier which delivers the cane to a two-roller or three-roller crusher that separates out about 60 percent of the juice by breaking and crushing the stalks or chips. Between the cane knives and the first mill, many factories use shredders with revolving and swinging hammers, to reduce the cane pieces to a fibrous structure, thus facilitating juice extraction by the mills. Shredders of the hammer-mill type shred the cane, and then pressurized rollers extract the juice in a process called maceration, imbibition, or saturation.

The conventional milling system is relatively inefficient, and continuous diffusers (such as are used to extract sugar from beets) have been successfully introduced to many cane-sugar factories. An example of such a cane-diffusion system is shown in Figure 2.

By courtesy of Hitachi Zosen Co., Ltd., Tokyo

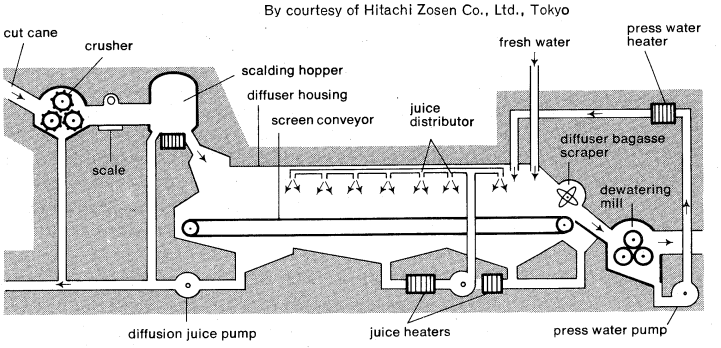


Figure 2: Cane-diffusion system.

The cane is cut in pieces as in a conventional milling plant by two sets of knives and is passed through a three-roller crusher, where about 60 percent of the juice is pressed out. The pressed cane, called crusher bagasse, is

conveyed into a continuous diffuser, in which sugar in the crusher bagasse is extracted in many stages by water percolation (lixiviation). Warm, fresh water is introduced at the bagasse outlet, and juice is drawn out after percolating through the bagasse at the bagasse inlet. This diffusion juice is mixed with crusher juice, and the mixture is sent to the clarification room. The diffuser bagasse contains much water, about 85 percent, and it is pressed by two sets of three-roller mills. The press water, after clarification, is used as lixiviation water for the diffuser, along with fresh water. The quantity of fresh water used in the diffuser is about 30 percent of the cane weight, and that of mixed juice, known as draft, is about equal to the cane weight. Lixiviation in the diffuser is carried out at close to 170° F (77° C) to inhibit bacterial growth.

Juice processing. The extracted, mixed juice is heated to about 220° F (105° C) and milk of lime (suspension of calcium hydroxide) added. The mixture then goes to a continuous juice clarifier. From here, clear juice is drawn off the surface of each tray and sent to the evaporator, while settled juice mud is scraped out and sent to mud receivers, continuous-rotary vacuum filters.

The invention of three basic machines in the 19th century made sugar a major industry; vacuum boiling pans (1813), centrifugal machines (1837), and effect evaporators (1843).

Effect evaporator

The effect evaporator shown in Figure 1 is a quadruple evaporator. Each of its four vessels has a set of heating tubes called calandrias. Heated in the tubes, the juice boils up with water vapour bubbles into the upper-vapour chamber. Only the juice comes down through a central downtake and goes again to the heating tubes, becoming more concentrated.

The calandria of the first evaporating vessel is heated with steam, that of the second with the vapour from the first. Vapour from the last vessel is drawn out by a vacuum pump and condensed in a water-cooled condenser. Pressure differences between the connecting vessels cause the juice to move to succeeding vessels when valves of the connecting tubes are opened.

Steam is supplied only to the first calandria and one pound of steam can produce about one pound of vapour by heating the juice to its boiling point; at the same time, one pound of vapour thus generated can also do the same work in the second vessel. If four evaporation vessels are used and one pound of steam is supplied, about four pounds of water can be evaporated from the juice. This is an effective and economical method of steam usage.

Raw-sugar factories may use triple, quadruple, or quintuple evaporators; quadruple evaporators are most usual.

The juice concentrated in the last vessel is drawn out by a syrup pump. This is raw syrup of about 65 percent solid and is sent to boiling pans.

Vacuum boiling pan

The structure of a vacuum boiling pan is shown in Figure 3. Near the bottom is a heating calandria. Modern pans have mechanical stirrers in the centre downtakes to insure good circulation of massecuite. A large vapour chamber above the calandria is connected to a vacuum pump via a cooling condenser so that a moderate vacuum can be maintained inside the pan during boiling.

To begin boiling, the pan is evacuated, raw syrup is drawn in up to the height of the upper tube-plate of the calandria, and steam is let in. The syrup begins to boil, becoming more and more concentrated. Seed crystals, small crystals of sugar, are added to the raw syrup as starters to help the crystallization of sugar from the syrup. Sucrose from the syrup begins to deposit on the seed crystals, causing them to grow. The growth of the seed crystals is maintained with continuous feeding of raw syrup until the massecuite volume reaches about 80 percent of the pan volume. During boiling, the sucrose crystals grow from about 0.012 inch (0.3 millimetre) to 0.039 inch (1.0 millimetre) in length. After the massecuite is made a little harder, the supply of steam is stopped and pan vacuum is cut off to the atmospheric pressure. The massecuite is then dropped into receivers for centrifugals.

In the centrifugal machine, the massecuite is separated into raw sugar and molasses; this molasses is boiled to produce a second massecuite, which is separated into a second raw sugar and a second molasses. The second molasses is also boiled and made to yield a third sugar, which is used as seed crystals for first and second massecuite boilings. The third molasses is exhausted molasses and is known as cane blackstrap.

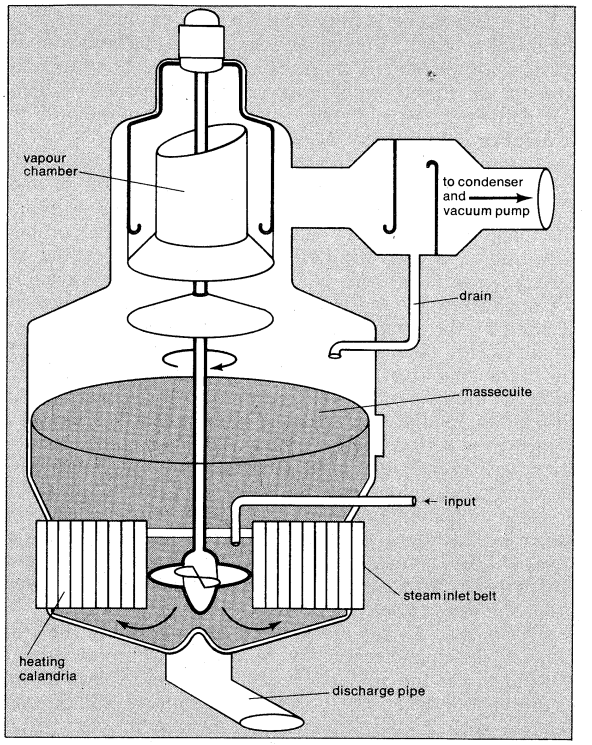


Figure 3: Principal structural features of the vacuum boiling pan with mechanical agitation.
By courtesy of Hitachi Zosen Co., Ltd., Tokyo

Discoloration of sugar products is far less in a vacuum pan, where the surface temperature of boiling massecuite is about 140° F (60° C), than in an open pan, where the temperature of the massecuite is over 250° F (121° C). For this reason, the invention of vacuum boiling pans was revolutionary in sugar manufacturing.

First and second massecuites are centrifuged immediately after discharge; the third massecuite is cooled to induce more dissolved sucrose to crystallize out.

Centrifugal machine

The centrifugal machine consists of a cylindrical basket to receive the massecuite; the basket is pierced with numerous holes to allow the molasses to escape and is lined with a screen of perforated sheet metal. An electric motor rotates the basket via a vertical shaft fixed at the bottom. The massecuite is fed into the slowly rotating basket until centrifugal action forces it to creep up to the edge, and then the basket is rotated at full speed (1,000–1,600 revolutions per minute). After a few minutes, the rotation is slowed almost to zero and the sugar inside the basket wall is scraped down onto a belt conveyer. After passing through coolers, where large conglomerates are removed by classifying screens, the raw sugar is stored in bulk.

In many modern cane-sugar factories, the operations are controlled automatically. At the milling station, the quantity of cane loaded to the mill is controlled by the motor speed of the cane carrier, which in turn is regulated according to the load change of the cane knives' motor and the depth of cane charged to the crusher. The control of evaporator work is effected by correspondence between the quantity of clarified juice and the steam supply and by the maintenance of a definite raw-syrup concentration.

Plantation white sugar (mill white sugar). Plantation white sugar is manufactured in south Asian countries for their own direct consumption. There are two methods of manufacture: impregnation with sulfur dioxide, sulfitation, and carbonation.

In sulfitation factories, mixed juice from the mill is heated to 165° F (74° C) and sent to tanks where the

gas is injected at the same time that milk of lime is added. After heating to 220° F (105° C), juice saturated with sulfur dioxide is sent to continuous clarifiers. The clarified juice is concentrated to raw syrup, which is further saturated with sulfur dioxide. The first massecuite is obtained from this syrup, and on centrifuging it yields white sugar.

In carbonation factories, mixed juice is heated to 131° F (55° C) and sent to carbonation tanks, where carbon dioxide is injected in the juice at the same time that milk of lime is added. Calcium carbonate precipitates are filtered off and the filtrate is further saturated with carbon dioxide and filtered again.

Sulfur dioxide is then injected into the carbonation juice, and the juice is concentrated to raw syrup; the syrup is further saturated with sulfur dioxide and filtered. White-sugar massecuite is obtained from this syrup.

BEET-SUGAR PRODUCTION

Sugar beet. Sugar beet, *Beta vulgaris, Saccharifera Alefeld* variety, is a biennial plant of the Chenopodiaceae family. It has long been grown as a summer crop in relatively cool northern parts of the temperate zones of the world, and recently, it has been grown as a winter crop in the southern parts of the temperate zones: South America, Africa, the Near and Middle East, and south Europe.

Culture. The growing period from sowing to harvesting is 170–200 days. A good yield of beet roots is obtained when the climate has been mild throughout the growing period. A good sucrose content in the beet roots is secured when the last period of growth has been cold. In the case of a winter crop, the ripening period is in the warm season; and ripening of the beet is promoted by withholding water to the beet.

Sugar beet requires a well-distributed precipitation of about 24 inches (610 millimetres) from January to October. If the precipitation is deficient, the crop must be irrigated.

Sugar beet is grown in various soils, ranging from sandy loam to heavy clay. An ideal soil is loam rich in humus, deep and homogeneous, having appropriate adhesion and mild moisture-holding capacity.

The seed bed is prepared by deep plowing after the preceding crop is harvested in autumn, and early the next spring it is shallow plowed, stamped, and harrowed.

Fertilizers are applied to sugar beets from the beginning of sowing through the entire growth period. Nitrogen fertilizer increases the weight of beet roots but delays ripening. Potash is well absorbed by sugar beet, increasing the root weight; but, again, if too much potash is absorbed, ripening is slowed down and, in addition, magnesium deficiency occurs. The absorption of phosphate is less than that of nitrogen and potash, but it increases beet-root weight and accelerates ripening.

Seed balls

Seed balls, or fruits, containing many seeds are formed at maturity by the sugar beet flowers. Balls containing single seeds were first obtained in the U.S.S.R. from atypical beets. Commercial production of single-seed balls has also been developed in the United States. Most beet seeds now used are from seed balls containing a single seed.

With single seeds, the quantity required has been reduced to less than three pounds (1.4 kilograms) per acre instead of the 20 pounds (9 kilograms) or more of whole-seed balls formerly required. When single seeds are planted with proper spacing, the seedlings grow at intervals that obviate the great labour expenditure for thinning and weeding required for seedlings from seed balls with many seeds. Recently, beet seeds have been grown in mild climates by the winter-annual method in which sowing is done in late summer or early autumn; the plants are dormant during the winter and produce seed the following season.

In most areas seeds may be sown during the latter part of March or in April. Before sowing, beet seeds are treated with disinfectants for black root disease. Precision drills sow monogerm seeds of well-classified sizes to a depth of ¾–1½ inches (2–4 centimetres) spaced 2½–3 inches (6–8 centimetres) apart. Rows are 20–22 inches (50–56 centimetres) apart. Fertilizers are applied simultaneously with seeds, and after covering, herbicides are applied by spray. The germination of the seeds occurs about ten days after sowing.

When weeds appear in the interrows they are removed so that fertilizers, air, and rainwater are more accessible to the beet roots.

Thinning is performed in the four- to six-leaf stage to keep the distance between plants at 8–10 inches (20–25 centimetres), leaving a final stand of seedlings averaging 100 to 150 per 100 feet (30 metres) of row.

Harvesting sugar beet

Sugar-beet harvesting usually starts late in September or early October and is performed rapidly so as to finish before the soil freezes. There are two methods of harvesting. In the Pommritzer method, the topping and the lifting of the roots are performed by two separate machines. In the other method, the two operations are carried out by one machine. The harvested beet roots are carried directly to the factories in trailer bins.

The sugar-beet crop is produced every four to six years; several other kinds of crops are also raised during this period. Sugar beet is usually planted after corn or wheat, in order to lessen the damage due to *Rhizoctonia* root rot or nematodes.

Diseases and pests. The beet plant is subject to many diseases and insect pests. Black root rot, a fungus disease characterized by lesions in the stem near the soil surface, and *cercospora* leaf spot, a fungus infection in which the leaves become greenish-yellow and root weight and sugar content are reduced, are most serious and can cause great damage if not controlled. Precautions must also be taken against damage by worms, beetles, and nematodes.

Breeding. Disease-resistant beets of higher sucrose content and heavier root weight are constantly sought. Sugar beet is a cross-pollinated plant and commercial varieties are hybrid plants. Superior polyploid varieties (varieties having multiple sets of chromosomes) have been developed in Europe. In the United States, breeding has been based on hybridization.

Beet-sugar manufacture. Harvested beets are carried as directly as possible to the factory yards where they are weighed and dropped into flumes to be carried by water to the factory. Dirt and extraneous matter adhering to the roots are removed during the fluming.

Washing and slicing

Preliminary processing. In the factory, beet roots are lifted by a wheel or pump to a beet washer, a tank overflowing with wash water and fitted with a horizontal shaft and propeller paddles. Washed beets are picked up from the tank and sent to a bunker over the beet slicer.

In Europe, suspended beet slicers with horizontally rotating slicing disks are widely used. They rotate at about 1,000–1,500 revolutions per minute. Washed beets are fed into a hopper and onto the rotating disk by gravity. Sliced-beet pieces, called cossettes, drop onto belt conveyors leading to the diffuser via a belt weigher.

The operating mechanism of a continuous beet diffuser is shown in Figure 4. Beet cossettes are fed to a hopper at the lower end of an inclined diffuser and carried upward by two scroll conveyors. Fresh hot water and press water from the diffuser-pulp press are added at the upper end of the apparatus. The cossettes are repeatedly squeezed lightly and loosened between two scrolls of the conveyors. Sucrose-rich juice accumulates at the lowest end of the apparatus, where it is drawn off, becoming diffusion juice, or raw juice.

The pulp is removed from the diffuser end and sent to presses, where pressed pulp and press water are obtained. The pressed pulp is then sent to dryers.

In order to inhibit bacterial growth in the diffusion apparatus, sucrose extraction is carried out at 160°–170° F 71°–77° C); and, usually, Formalin is added to the returned press water.

Preliming

The raw juice is sent to a tank where it is mixed with milk of lime which coagulates impurities. This process, called preliming, is indispensable for securing juice of good settling properties and filterability. The prelimed juice is then heated to 170°–180° F (76°–82° C) and again mixed with milk of lime in a main-liming tank. Preliming, main-liming, and carbonation are all continuous.

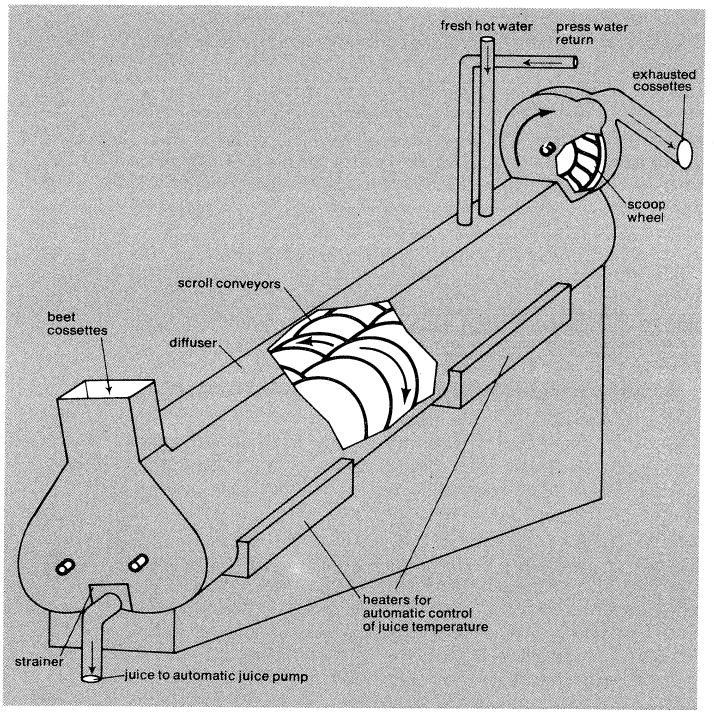


Figure 4: Continuous beet-diffusion system.
By courtesy of De Danske Sukkerfabrikker

The limed juice is further heated to 180°–190° F (82°–88° C) and carbonated in two or three tanks by the injection of carbon dioxide from the tank bottoms. Carbonated juice overflows to succeeding tanks. After the end point of carbon dioxide saturation is reached at the last tank outlet, the carbonated juice, along with the calcium carbonate that has precipitated out, is sent to a continuous mud-settling tank, or thickener, where clarified juice and muddy juice are separated. The thickened, muddy juice is filtered by continuous rotary vacuum filters, and the mixture of filtrate and clarified juice from the thickener is filtered, preferably by ceramic filters. The filtrate is heated to 210° F (99° C) and carbonated again.

The second carbonated juice is filtered by ceramic filters or pressure-leaf filters and the filtrate sent to a tank where sulfur dioxide is injected into it. This juice is then concentrated in a quadruple or triple-effect evaporator to a thick juice. Pressure evaporation is carried out in the first vessel of the effect at 15.5 pounds per square inch, and extra vapour is bled from the three vessels for heating and boiling. The thick juice is mixed with remelted liquors and filtered, the filtrate becoming the standard liquor.

Raw beet sugar. The thick juice is fed into boiling pans, and, after seeding, the seed crystals grow on the continuously fed juice and on the first molasses. As with cane sugar, the massecuite thus obtained is separated by centrifuge into raw sugar and molasses. This molasses is again boiled to produce the second massecuite, which is separated into the second raw sugar and molasses. The second molasses is considered exhausted.

White sugar manufacture. White sugar is produced by the three-boiling system from standard liquor, which consists of thick juice, the second sugar remelt, and the wash molasses from the first-massecuite centrifuging. The second massecuite is boiled from the first molasses, the third sugar remelt, and the wash molasses from the second-massecuite purging. The third massecuite is boiled from the second molasses to obtain the third sugar and exhausted molasses.

The Steffen process for recovering sucrose from exhausted molasses has been adopted in the U.S. and in Japan. This molasses contains about 50 percent sucrose. Finely powdered quicklime is added to diluted molasses, precipitating saccharate of lime that is used for liming beet juice in place of milk of lime. The disadvantage of this process is that raffinose is precipitated along with sucrose.

Use of ion-exchange resins

Many modern factories use ion-exchange resins that react chemically to de-ash and decolorize the thin juice, resulting in reduced quantity of molasses and improved quality of the sugar product.

The operations in modern beet-sugar factories are controlled automatically. At the diffusion station, variables are controlled according to the weight of the cossettes, as estimated at the belt weigher; at the liming station, the raw-juice quantity determines the quantity of lime added. Carbonation is controlled by setting the carbon dioxide valves according to the acidity of the juice at the last tank outlet. The evaporator control is effected by the correspondence between quantity of thin juice and the steam supply, which is adjusted according to the Brix, or percent by weight of sucrose in the thick juice. For sugar boiling, purging, and the boiler-house operations, controls are the same as for raw cane-sugar manufacture.

SUGAR REFINING

Most cane sugar is consumed in the form of refined sugars, and sugar refineries are located in or near big cities. Modern refineries are accessible to bulk, raw-sugar carriers.

Raw sugars. Of primary concern to the sugar processors is obtaining raw sugars of good refining qualities: (1) uniform grain sizes for ease of washing, called affining; (2) less colour and better filterability of washed, or affined, sugar; (3) lower ash content for higher sugar recovery.

Raw sugar is unloaded from sugar carriers at the factory storage house or sugar silos via an automatic weigher. The necessary quantity of raw sugar is carried to a continuous mingler, where it is mixed with warm syrup forming a continuous flow of magma, a thin, pasty mixture of raw sugar and syrup. The magma removes the molasses film that adheres around raw-sugar crystals and is fed into centrifugal machines for separation into affined sugar and affination syrup.

Affined sugar is dissolved at 150° F (66° C) in hot, sweet water with some fresh, hot water added to obtain raw liquor of about 65° Brix. Purification of raw liquor is most efficient and economical when a combination of liquor-carbonation and bone-ash decoloration methods are used. In the carbonation process, milk of lime is added to the continuous flow of raw liquor from the melter and the mixture fed into the carbonation tanks, where carbon dioxide is blown in from below. Carbonated liquor overflows to a succeeding carbonator tank where the reaction is completed. The liquor carbonation is carried out continuously. Carbonation produces decoloration of 50–60 percent, and the calcium carbonate precipitates that form are a good filter aid.

The carbonation slurry, or thin mixture, thus obtained is heated to 175° F (80° C) and filtered without added filter aid. The filtrate is further decolorized by percolation through deep beds of bone ash, tribasic calcium phosphate, which acts as a clarifying agent. A fine liquor is thus obtained and sent to the boiling house. The fine liquor is boiled to obtain sugar and molasses; then the molasses is boiled again to obtain the second sugar yield, and so on. The boiling system is complicated because the purity of fine liquor is higher than 98 percent, and at least six or seven stages of boiling are required until the molasses is exhausted; three stages of boiling are needed to desugar the affination syrup. A mixture of first, second, third, and sometimes fourth sugars is sold as white granulated sugar. Yellow or brown soft sugars are produced from low-grade massecuites. Refined products are dried, classified, and stored in sugar bins, from which they are drawn out for packaging or sending to the market in bulk.

Bone-ash decoloration

The installation for bone-ash decoloration is expensive, and many factories have adopted a combination method using carbonation and activated carbon for decoloration. Many Japanese refineries use ion-exchange resins for final decoloration of the liquor.

Automatic controls operate in the factory: the quantity of raw sugar fed into the mingler is measured continuously on a belt weigher and the required amount of affination syrup added according to the raw-sugar weight. Affined sugar is sent to the melter, where raw liquor of a constant

concentration is made automatically by adding controlled amounts of hot water. The liquor carbonation is controlled as for beet-sugar manufacture; the filtration end point of carbonation slurry is detected automatically by pressure measurements. Timers control the sequence of the bone-ash decoloration operation. Control of the boiling and purging operations is the same as for raw-cane-sugar manufacture.

Sugar types and grades. Most of the sugar consumed in the world is in the form of white granulated sugar from refineries, beet-sugar factories, or cane-mill white-sugar factories. These are classified according to grain size as extra-coarse, coarse, standard, fine, and extra-fine granulated. The grain size of extra-coarse granulated is larger than 0.04 inch (1.0 millimetre); that of extra-fine granulated is less than 0.01 inch (0.25 millimetre). Bakers' special (U.S.), or caster sugar (England), is produced by boiling very fine crystals; powdered sugar or icing sugar is produced by milling, with cornstarch added to prevent caking.

About 1,000,000 tons of soft white sugar is manufactured annually in Japanese refineries. White sugar massecuite of fine grain size, about 0.005 inch (0.125 millimetre), is purged, and the remaining sugar in the centrifugal baskets is sprayed with colourless inverted syrup. Soft yellow sugar and soft brown sugar are obtained from low-grade massecuites in the refineries. Low-grade sugars retain the reducing sugars (levulose and dextrose) of the massecuites.

Coffee sugar (England) consists of large brown crystals with characteristic flavour, produced as granulated sugar in several successive boilings from refinery liquor of appropriate colour. Candy sugar consists of very large white sugar crystals, obtained by slow crystallization from very high purity liquor; it is used mainly by the brewing industry.

Liquid sucrose and invert sugar

More than 2,000,000 tons of liquid sugar is produced in the U.S. chiefly for manufacturing purposes. There are two superior-grade liquid sugars: liquid sucrose and liquid invert. (Invert consists of equal parts of glucose and levulose obtained by the inversion of sucrose.) Liquid sucrose is manufactured by adjusting the concentration of fine liquor in refineries to 67.2 percent sucrose containing a maximum of 0.20 percent invert sugar. For liquid invert, the sucrose in the fine liquor is inverted more than half by passing through an ion-exchange resin column and adjusting the concentration to 77 percent solid.

Golden-coloured liquid sugar is produced for domestic use by processing selected refinery syrups, partially inverted.

Cube sugar in the form of regular hexahedrons or rectangles is produced by molding and pressing selected granulated sugar with sugar liquor for cementing the crystals and then drying.

In 1967, world production of noncentrifugal sugar in the form of brown lump sugar was 12,000,000 tons; most of this was produced in Asia.

About 170,000 tons of palm sugar in the form of lumps is produced in Asian tropical countries from the juice of coconut, palmyra, and date palm, obtained by tapping their trunks or inflorescences. Maple sugar, syrup or lumps, is produced in Canada and the northern U.S. from the juice of sugar-maple trees; the lower part of their trunks is tapped in early spring.

ECONOMICS

Production and consumption statistics. World production of raw sugar amounted to 73,710,000 metric tons, excluding noncentrifugal sugar, in 1968, of which cane sugar accounted for 40,960,000 and beet sugar 32,750,000 metric tons. World production of noncentrifugal sugar in 1966–67 amounted to 11,980,000 metric tons.

The principal cane-sugar producing countries are as follows (the figures in parentheses are tons of sugar produced in 1970): Cuba (8,530,000), Brazil (5,680,000), Australia (2,610,000), India 4,710,000), United States (2,440,000), Mexico (2,890,000), China (mainland) (2,400,000), South Africa (1,970,000), Philippines (2,200,000), and Argentina (1,160,000).

Main beet-sugar producing countries are as follows: U.S.S.R. (10,780,000), United States (3,060,000), France (2,530,000), West Germany (2,170,000), Poland (2,040,000), Italy (1,420,000), United Kingdom (1,040,000), Czechoslovakia (930,000), The Netherlands (780,000), and Turkey (771,400).

The world consumption of centrifugal sugar in 1968 amounted to 74,360,000 tons. Consumption is increasing yearly at a rate of 2–4 percent. The rate fluctuates because of the price. Per capita yearly consumption for the world was 43 pounds in 1968. Per capita consumption in pounds (0.454 kilogram) per year for ten countries are: The Netherlands (154), Ireland (141), Australia (126), Denmark (116), United Kingdom (114), Switzerland (114), United States (111), Israel (108), Soviet Union (105), and Canada (102).

World sugar trade. About 70 percent of world sugar production is consumed domestically. The balance is consumed in international trade: exports to the United States under its quota arrangement, exports under the Commonwealth Sugar Agreement, Cuban exports to socialist countries, exports from the U.S.S.R. to other socialist countries, and free-market sugar for trade. Except for free-market sugar, trades are carried out at preferential prices negotiated between the countries concerned.

The quantity of free-market sugar was about 9,400,000 tons in 1969; world production was about 68,000,000 tons. Free-market sugar is traded according to contracts between the countries concerned; the price is usually based on spot quotations on recognized exchanges, such as the New York Coffee and Sugar Exchange or the United Terminal Sugar Association (London).

The price of free-market sugar has fluctuated. In 1962 the price rose to 12.6 U.S. cents per pound because of the temporary depletion of sugar stock and sudden worsening of American-Cuban relations. In 1967 the price fell to 1.23 U.S. cents because of a surplus.

International Sugar Agreement

In order to stabilize sugar production and the free-market sugar price, the International Sugar Agreement was concluded in 1969, regulating the export quota for each member country and setting a floor price (3.25 U.S. cents per pound, or 0.454 kilogram) and a ceiling price (6.5 cents). This is the third post-World War II international agreement (1953, 1958, 1969) concluded under the auspices of the United Nations Conference on Trade and Development.

The retail price of refined sugar varies greatly in different countries, principally because of different amounts of custom levied on raw sugar and excise on refined sugar products. In 1969 the price was 3.8 U.S. cents per pound in Mauritius, 5.1 cents in Jordan, 7.6 cents in Singapore, 8.8 cents in Great Britain, 8.8 cents in Canada, 12.2 cents in the U.S., and 16.4 cents in Japan.

SUGAR BY-PRODUCTS AND SUGAR CHEMICALS

Sucrose esters

Methods of making useful sucrose derivatives have long been developed by the International Sugar Research Foundation, Incorporated, New York; sucrose esters are manufactured by a Japanese licensee for use as surface-active agents in food products.

Cane molasses. The world cane-molasses production was 14,000,000 tons in 1968–69. The major use is in livestock feed. Molasses contains 55 percent sugar and 2.8 percent protein and is a cheap source of readily available energy. Six to 10 percent by weight is added to dairy-cattle feeds. The second important use is for making alcoholic drinks, such as rum in Caribbean areas. A recently developed and interesting use for molasses is in manufacturing monosodium glutamate by fermentation in Japan and Taiwan; the diluted molasses, with urea added, is fermented by *Micrococcus glutamicus*.

Bagasse fibre. Bagasse, 24–27 percent of the cane, remains after the juice is pressed out. It is ordinarily burned as boiler fuel; but bagasse may be used for making soft or hard building board. Approximately 550,000 tons of bagasse pulp was manufactured in 1967 by 36 bagasse-pulp mills in the world; the bagasse is depithed; that is, the soft, centre tissue of the plant stems is removed. The pulp is made from such a depithed bagasse by sulfate process.

Cane wax. Cane wax is extracted from the dried filter cake of raw-sugar factories; naphtha, toluene, or alcohol is used as the extraction solvent. After decoloration, cane wax is used in manufacturing polishes, cosmetics, and paper coatings.

Beet pulp. Beet pulp is fed to livestock; wet pulp may be used soon after delivery from factories, but most of it is dried in oil-fired, direct-rotary driers and stored. Dried pulp contains about 9 percent protein.

Beet molasses. Beet molasses is also fed to livestock and may be added to beet pulp. Monosodium glutamate can be manufactured from beet molasses; it contains pyrrolidone carboxylic acid, which is easily hydrolyzed; that is, chemically split with water added, to form glutamic acid. In Europe beet molasses is used for manufacturing citric acid by fermentation.

SUGAR SUBSTITUTES AND NON-NUTRITIVE SWEETENERS

Sugar substitutes. Much corn syrup and honey are used as substitutes for sugar; in the United States a small quantity of sorghum syrup is used for the same purpose (see CEREALS AND OTHER STARCH PRODUCTS).

Corn syrup is a product of the hydrolysis of cornstarch; that is, the chemical splitting of cornstarch with addition of water.

Honey is inverted cane-sugar syrup with flower flavours, which is accumulated by bees during the blossom season.

Sorghum is the syrup made by concentrating sorghum-stalk juice.

Artificial, non-nutritive sweeteners. Artificial sweeteners were used originally as sugar substitutes for diabetics and others forced to restrict their intake of sugar, but recently they have been used for making soft drinks and other foods for people who wish to restrict their intake of calories to control their weight. Saccharin, cyclamates, and Dulcin, or Sucrol, are such artificial sweeteners.

Saccharin was discovered in 1879 by C. Fahlberg and Ira Remsen of Johns Hopkins University. It was first sold in 1884 in New York as sodium or calcium salt. It has hundreds of times the sweetening power of cane sugar.

Sodium cyclamate, or sodium cyclohexylsulfamate, was discovered in 1937 by L.F. Audrieth and M. Sveda at the University of Illinois. Many times sweeter than cane sugar, it was used as a sweetener from 1950 until 1969. Its general use has been prohibited in the United States, the United Kingdom, and Canada, because of evidence that it might cause cancer.

Dulcin, or Sucrol, is no longer used because of its toxic effects.

Austria has prohibited the use of artificial sweeteners in foods since 1939, with the exception of saccharin for certain products.

BIBLIOGRAPHY. H.C. PRINSEN GEERLIGS, *The World's Cane Sugar Industry, Past and Present* (1912); N. DEERR, *The History of Sugar*, 2 vol. (1949–50), a detailed and authentic history; U.S. BEET SUGAR ASSOCIATION, *The Beet Sugar Story*, 3rd ed. rev. (1959); A.C. BARNES, *The Sugar Cane* (1964), a detailed description of world-wide cane-growing techniques; N.J. KING, R.W. MUNGOMERY, and C.G. HUGHES, *Manual of Cane-Growing*, new ed. (1965), an authentic manual for cane growers in Queensland; J.P. MARTIN, E.V. ABBOTT, and C.G. HUGHES, (eds.), *Sugar-Cane Diseases of the World* (1961), and J.R. WILLIAMS, *et al.*, *Pests of Sugar Cane* (1969), two detailed and authoritative works for the professional; G.C. STEVENSON, *Genetics and Breeding of Sugar Cane* (1965), a practical guidebook; T. YAMANE, "Cane Sugar Handbook" (1966), a book showing world-wide technical data (in Japanese); G.L. SPENCER and G.P. MEADE, *Cane Sugar Handbook*, 9th ed. (1963), a standard technical work; U.S. DEPARTMENT OF AGRICULTURE, *Sugar Beet Culture in North Central States*, rev. ed., Farmers' Bull. no. 2060 (1964); R.A. MCGINNIS (ed.), *Beet Sugar Technology* (1951); VEREIN DER ZUCKERINDUSTRIE, *Technologie des Zuckers* (1968), a detailed technical book written by German authorities.

(T.Y.)

Suger of Saint-Denis

While the abbot Suger of Saint-Denis is important in the early history of France, it is for his role in the history of art that he is best remembered. Under his supervision and direction the abbey church of Saint-Denis (the patron saint of France) was rebuilt from a crumbling and decaying condition. It was a period that saw the emergence of new ideas in theology and philosophy, which the church of Saint-Denis fully reflects. It is believed that Suger was the inspiration behind many of the architectural innovations employed at Saint-Denis, which include an original use of the pointed (rather than round) arch and the ribbed vault and extensive use of stained glass, including a rose window in front. These features were instrumental in the development of a new style of architecture, which became known as Gothic.

Archives Photographiques, Paris



Suger, detail of a stained glass window, 12th century. In the abbey church of Saint-Denis, France.

Suger was born of peasant parents, probably somewhere near Paris, in 1081. As a child he showed unusual intelligence and was brought to the nearby abbey of Saint-Denis to be educated by the monks. His closest friend and schoolmate at the abbey was a boy his age, Louis Capet. This boy became King Louis VI in 1108. Suger became secretary to the abbot Adam of Saint-Denis and a close adviser to the king.

As secretary to the abbot Adam, Suger made various diplomatic missions to Henry Beauclerc of Normandy, who was also King Henry I of England and the son of William the Conqueror. Suger was greatly impressed by the strong, orderly administration of the Norman ruler, which contrasted with the chaotic feudalism in France.

The importance of Saint-Denis

Saint-Denis, shrine of the saint who had supposedly brought Christianity to Gaul, was an object of great veneration. Suger saw its destiny and that of the French crown as permanently related. He believed that by stressing and enlarging the king's role as vassal of Saint-Denis he could unite the king and his nobles under an idea they could mutually believe in. Suger also saw that the king could, and should be, protector of the peasants and the middle class.

In 1122 Suger was elected abbot of Saint-Denis. He had a chance shortly thereafter to test his idea of the cementing power of the symbolic idea of Saint-Denis. In 1124 the Holy Roman emperor Henry V invaded lands ruled by King Louis VI. The King rode into battle carrying the Oriflamme, the banner of Saint-Denis, which normally rested in the church, along with the relics of the saint. As a result of his (and Suger's) appeal to the nobility's veneration for the saint, he was followed by a larger army of nobles than had ever previously pledged their allegiance to him, or to his father. The army of Louis and the Oriflamme were so formidable that Henry V retreated without battle.

Although Suger was not an ascetic but a reasonable and humane man in a time of violent extremes, he guided the monks of Saint-Denis back into a life of greater piety and religious observation than they had known under the abbot Adam. In Adam's administration the monks had acquired a notorious reputation for behaving in an excessively secular fashion. Suger corrected this situation at the urging of Bernard, abbot of Clairvaux, later St. Bernard. Suger cooperated in this and many other matters with Bernard, who was a close adviser to the Pope and the greatest spiritual leader in Europe at the time. He could be a powerful enemy or ally, and Suger chose to be his ally.

After King Louis's death in 1137, his successor, Louis VII, rejected Suger's role as principal adviser, and Suger concentrated all of his efforts for the next five years on completing the rebuilding of the church of Saint-Denis. But in 1142 Louis seized lands belonging to his most powerful vassal, Thibaut, count of Champagne. Civil war resulted. The support of the powerful Thibaut had always been vital to the strength of the French monarchy, and the young king was making war ferociously and irrationally. Suger stepped in as an active adviser to Louis VII, as he had always done with his father, and negotiated a peace treaty between Thibaut and Louis. The treaty was signed at the dedication ceremony of the church of Saint-Denis, an architectural marvel.

As penance for the many lives that he had taken during the war with Thibaut, Bernard of Clairvaux urged Louis VII to lead a crusade to free the Holy Land from the Muslims. Suger was strongly opposed to this and tried unsuccessfully to change the King's mind. For the first time Suger stood in opposition to the wishes of the weak, young king as well as those of Bernard and the Pope.

Suger's regency

On June 11, 1147, Louis and Eleanor departed on the Second Crusade. Louis left his crown with Abbot Suger, who was appointed regent in his absence. The Crusade was a disastrous loss, but at home Suger governed well, despite the great financial drain on the funds at his disposal. He devised new and fairer means of taxation, passed laws preventing deforestation, and suppressed a revolt by a group of nobles who planned to make Robert, count of Dreux, brother of Louis VII, king in his absence. When, in 1149, Louis returned from the Crusade, many believed that Suger would not return the crown, but they were proved wrong.

In 1150 Suger himself, with Bernard, made plans for another crusade. But in 1150, before it was begun, Suger fell ill with malaria. He died on January 13, 1151.

BIBLIOGRAPHY. ANNE ROCKWELL, *Glass, Stones and Crown: The Abbe Suger and the Building of St. Denis* (1968), a brief, simply written biography of the Abbot Suger, focussing upon his career as architect and planner of the abbey church of Saint-Denis; ERWIN PANOFSKY (ed. and trans.), *Abbot Suger on the Abbey Church of St. Denis and Its Art Treasures* (1946), the English translation of Suger's Latin work, describing the building of the church of Saint-Denis, accompanied by a perceptive commentary and biographical introduction; AMY KELLY, *Eleanor of Aquitaine and the Four Kings* (1950), a readable biography of Eleanor of Aquitaine, her times, and the people, including Abbot Suger, who played an important part in her life.

(A.F.R.)

Suicide

The term suicide means the act or instance of taking one's own life voluntarily and intentionally but the term can also be applied to any person who attempts to take, or has a tendency to take, his own life. It also refers to the behaviour of individuals and groups, even of nations, that may bring about their own self-destruction. The term partial suicide has been used for self-mutilation, and chronic suicide for habitual behaviour patterns injurious to life, such as alcohol and drug addiction. This article is concerned with suicide as the act of "taking one's own life." This definition needs further clarification, however, as it does not say anything about the outcome. It is now customary to distinguish between "suicide"—that is, the fatal suicide—and "attempted suicide," or the nonfatal suicidal act.

PROBLEMS OF DEFINITION AND CLASSIFICATION

Question of intention

The statement that a person has committed suicide is a judgment about the cause of death based on the inference of self-destructive intention. If the intention was clear and unambiguous, if it had been communicated to others in writing or in some other way, and if the method of self-destruction employed left no doubt about the suicidal nature of the act, it would be easy to judge a death as having been suicidal. But these conditions do not obtain for the majority of suicidal acts, fatal or nonfatal. Self-destructive intentions are often difficult to establish retrospectively; in case of survival they are often denied, either because they were not clear and unambiguous, or because the truth is deliberately withheld. Suicide notes are left behind in only a minority of cases. The method of self-damage preferred in many countries today—namely, self-poisoning—does not exclude the possibility of accidental death. In his classic study, *Suicide* (1897), Émile Durkheim, the French pioneer of sociological research on suicide, rejected ascertainment of self-destructive intention as the indispensable criterion. He put the emphasis on the anticipation of the outcome. He implied that no one would carry out an act he knew to be fatal unless he intended to die. Methods employed to achieve self-destruction have changed since Durkheim, but his views on the difficulty of establishing intention directly still hold true today.

The majority of suicidal acts are nonfatal and need to be differentiated from nonsuicidal acts of self-damage. The outcome of most suicidal acts, including many fatal ones, is uncertain. The following definition seems to be the most serviceable: "A suicidal act is any deliberate act of self-damage which the person committing the act cannot be sure to survive." The question is simply whether the person committing the act of self-damage was likely to be aware of its potential dangerousness. Thus, an act of self-poisoning carried out by a person with expert knowledge about the effects of poisons will be viewed differently from a similar act committed by a person without such knowledge.

The definition of what constitutes a suicidal act is fundamental to suicide statistics. Statistical data based on different concepts and definitions of suicide are not strictly comparable.

Medicolegal definitions of suicide. In almost all countries every death suspected of being due to violence, including suicide, is subject to a statutory procedure. Suicide can be defined operationally as a death declared suicidal by the legal authority entrusted with this judgment, but there are no generally accepted criteria on which it is based. In the past, suicide used to be a criminal offense in many countries. This applied to England and Wales until the Suicide Act of 1961 repealed this legislation, which still exists in some countries once under British sovereignty. The legal approach prevails in the ascertainment of suicide. This means that positive evidence of suicidal intention is usually insisted upon, unless the method employed makes a suicide verdict inescapable. As a result, many cases that from the medicopsychological point of view would be classed as suicide are registered either as undetermined or accidental deaths because the information available did not meet legal requirements.

Suicide and undetermined or accidental deaths

Classification of suicidal acts. The division of suicidal acts according to outcome is essential, if only because the nonfatal act—that is, the suicidal attempt—calls for preventive measures concerning individuals and gives rise to problems not posed by the fatal act.

The best known classification of suicides is the one that was proposed by Durkheim and that is still used by some sociologists and anthropologists. The criterion of this classification was the disturbance in the relationship between society and the individual. Durkheim distinguished three types: (1) Egoistic suicide resulted whenever abnormal individualism had weakened society's control and reduced the person's immunity against the "collective suicidal inclination"; in this instance, the individual lacked concern for the community with which he was inadequately involved. (2) Altruistic suicide was due to an excessive sense of duty to the community. (3) Anomic suicide was due to anomie; that is, to society's failure to control and regulate the behaviour of individuals. Egoism and anomie may reinforce each other. Durkheim's classification has not been found useful by doctors and others actually dealing with suicidal individuals, however.

A great variety of acts of self-damage are subsumed under the term suicidal attempts. Numerous proposals for their classification have been made, but none of them has found general acceptance. They have been divided into "serious" and "nonserious" from the medical and psychiatric points of view. Suicidal "gestures" have been distinguished from "genuine" attempts, the "impulsive" acts

from deliberate ones, "psychotic" from the "nonpsychotic," the "rational" from the nonrational, and so on. Suicidal attempts apparently aimed solely at influencing others have been described as "manipulative" and harmless to life. The conceptions of death and the orientation toward death held by the person committing the act have been made the basis of classification. The term suicidal attempt has been rejected by some authors, and new terms have been proposed instead. Rating scales for "dangerousness," "lethality," "suicidality" have been suggested. All of these proposals have remained tentative, and none of them has been adequately tested for its usefulness. Classifications based on psychological criteria have to take account of the complexity of motivations and the coexistence of conflicting tendencies underlying suicidal acts.

SUICIDE IN VARIOUS SOCIETIES

Premodern societies. There would seem to be no society in which suicides do not occur. Statements to the contrary were based on travellers' tales rather than on systematic observations. Attitudes toward suicide differ in various societies and are linked with ideologies of death and afterlife.

Taboos and beliefs relating to suicide

Recent anthropological studies carried out among African tribes have revealed that their frequency of suicide was about the same as that of European countries with rather low suicide rates and that suicide among them was regarded as evil. Physical contact with the body of a suicide is feared to have disastrous effects, such as illness or suicide among one's kin. The tree on which a person hanged himself is felled and burned. The ancestors have to be placated by sacrifices. The place where the suicide happened is believed to be a haunt of evil spirits. Suicide is dreaded in the community, and a threat of suicide is sometimes used to exert pressure on the family.

Motives underlying suicide in premodern Asian societies are complex and include self-punishment, revenge, and rehabilitation in the eyes of the community. Violation of taboos and loss of status are common motives. Loss of status played a part in suttee, the more or less voluntary suicide of the widow in India, now outlawed. For the Hindu, death is not final but, rather, one incident in a long series of existences. There are timely and untimely deaths, suicide usually belonging to the latter category and leading to an earthbound ghostly existence. Similar beliefs still exist in modern society.

In Tikopia, a small island in the western Pacific where pagan ideology still lingers, the attitude toward suicide is one of mild disapprobation. The gods are believed to receive the souls of the dead but not of those who hang themselves. Their souls wander about until their ancestral spirits have found them. The spirits do not object, however, if a man commits suicide by going off to sea in a canoe, or a woman by swimming out to sea. These methods of suicide are admired. The Christian Tikopians believe that the soul of a suicide does not go to paradise but to satan.

Whereas Brahminism tolerates suicide, the attitude of Buddhism is ambiguous, though it encourages suicide under certain conditions, such as in the service of religion and country.

Ancient and medieval European societies. In ancient Greece and Rome, attitudes toward suicide varied among tolerance, condemnation, and admiration. Some philosophers saw in suicide the ideal way of gaining freedom from suffering. Others, among whom were Pythagoras and Plato, disapproved of it for reasons similar to those later advanced by Christianity (see below). Condemnation of suicide in principle was not incompatible with admiration for special cases. Cicero, for instance, rejected suicide but expressed admiration for Cato, who killed himself rather than live under Caesar's dictatorship. The hardening of the attitude against suicide in the late Roman Empire was partly because of the high incidence of suicide among slaves, who thus deprived their owners of valuable property.

The Old Testament mentions only four cases of suicide, the best known being those of Samson and King Saul, both of whom killed themselves to avoid torture or death by the enemy. Neither the Old nor the New Testament condemns suicide explicitly, in contrast to the Qur'ān, which declares it a crime worse than homicide.

The Christian Church in the Middle Ages adopted a similar attitude and condemned suicide as a form of murder and as interference with the prerogative of God. It was declared to be an act inspired by diabolical possession. The body of the suicide was refused a Christian burial and sometimes exposed to humiliating indignities as a deterrent to others. Acts of this type were still perpetrated in Europe at the beginning of the 19th century.

The attitude of the Christian churches has remained one of condemnation, but the sanctions are not equally severe in all denominations. The funeral rites accorded to the suicide differ more or less from those who have died a natural death. Exemption from sanctions on the ground of mental disturbance is more readily granted than in the past. Orthodox Jewish law, too, condemns suicide.

Modern societies. Public attitudes toward suicide are still as ambivalent and divided as in ancient Greece and Rome. There have been apologists for suicide even among the clergy. Eminent philosophers and writers have proclaimed man's right to dispose of his own life, whereas others have maintained its condemnation with equal conviction.

Decline of punitive attitudes toward suicides

In the Middle Ages and in modern times up to the 20th century, society has used first the canonic and later the criminal law in its fight against suicide. Following the French Revolution of 1789, antisuicide legislation was repealed in European countries, England being the last to follow suit in 1961. The change in the legal status of suicide has had no adverse influence on the suicide rate.

The "permissiveness" of modern society, which implies greater tolerance of deviant behaviour, may be partly responsible for the increase in suicidal acts, especially of self-poisoning. Society's attitude toward suicidal behaviour is probably less moralistic and punitive today than it used to be a generation ago. There is a greater readiness to understand rather than to condemn. Nevertheless, the tendency to conceal suicidal acts still persists, but the reasons for this tendency may not be quite the same as they were when a suicidal act was looked upon as a crime. Today a sense of guilt felt by those who could have prevented it may be the main reason for hushing it up.

Suicide in Japan. The custom of seppuku (hara-kiri), or ritual self-embowelling, used to be reserved for the nobility and members of the military caste. It has been regarded as a form of altruistic suicide, but it probably was in the majority of cases a form of sanctioned punishment whereby the individual chose to die at his own hand rather than be beheaded. Compulsory hara-kiri was made illegal in 1868. Voluntary hara-kiri as expiation for defeat or as self-sacrifice was highly praised and occurs occasionally even today. Although it has always been statistically insignificant, it still reflects the attitude of part of Japanese society that suicide is a demonstrative act of self-sacrifice.

The methods of suicide used in Japan today are similar to those employed elsewhere. This was probably always the case. The Japanese suicide rate used to be high, though not the highest among those reported by the World Health Organization (see Table 1). Recently the Japanese rate has declined considerably, but the significance of this reduction remains to be explained. There has at this same time been a marked increase of the suicide rate in the age groups below 30. A similar, though less marked, increase among the young has been observed in other industrialized countries. The excessive increase in the incidence of suicide among Japanese youth has been attributed to peculiarities of Japanese society, which has a strong sense of hierarchy and makes it difficult to express aggression against authority. Japanese youth is believed to suffer more than other young people from conflicts arising from a need for dependence, the wish for self-assertion, and a profound sense of insecurity.

High-risk groups. The elderly, the sick, the mentally ill (especially those suffering from depressive illness), alcoholics, soldiers, and policemen present above-average suicidal risks. In modern society new high-risk groups

Table 1: Death Rates from Suicide*
(1952–54; 1961–63)

	both sexes		male		female		excess male (1961–63)†
	1952–54	1961–63	1952–54	1961–63	1952–54	1961–63	
Australia	14.9	19.6	21.9	27.0	7.7	12.2	221
Austria	29.9	28.3	43.0	42.0	19.2	16.9	249
Czechoslovakia	...	28.2	...	40.9	...	16.5	248
Denmark	31.9	24.2	43.6	32.4	20.6	16.2	200
England and Wales	13.8	15.1	18.5	18.3	9.5	12.2	150
Finland	25.8	29.0	43.6	47.7	9.9	12.3	388
France	20.3	20.7	32.6	32.3	9.4	10.0	323
Germany (Federal Republic of)	23.6	24.1	33.7	33.3	15.1	16.2	206
Hungary	...	33.9	...	48.9	...	20.3	241
Israel (Jewish population)	...	10.1	...	11.9	...	8.2	145
Italy	6.4	7.1	11.9	10.2	4.8	4.2	243
Japan	31.4	7.1	38.8	29.0	24.5	20.6	141
Norway	9.8	10.0	15.1	15.7	4.6	4.5	349
Scotland	7.5	11.4	10.7	15.2	4.6	8.1	188
Sweden	23.4	21.7	35.2	32.0	10.7	11.6	276
Switzerland	28.8	23.3	44.9	33.9	14.3	13.2	257
United States	14.1	15.6	22.8	24.0	5.9	7.7	312

*Yearly average rates for each sex and both sexes, from 15 years of age per 100,000 population.
†Number of male suicides for every 100 female suicides.
Source: Public Health Paper no. 35, Geneva: WHO, *Prevention of Suicide*, 1968.

Increases in young suicides

have come to the fore, the best known being physicians, students of universities and allied institutions, and drug addicts. The excessive suicide rates among students have been related to the general increase of aggressive behaviour among the young and to the special competitive position of the student in modern society.

Voluntary euthanasia and suicide. The steady increase in the number of the aged in modern society and the ability of medicine to prolong life have led to demands for legalization of voluntary euthanasia. A bill that would have enabled individuals suffering from incurable illness to have their lives terminated painlessly was introduced but defeated in the British Parliament in 1970. Its advocates argued that voluntary euthanasia was "suicide by proxy" and that it required only a modification of the English Suicide Act of 1961, which legalized suicide but made aiding and abetting suicide a criminal offense. The idea of legalized euthanasia has many adherents among the general public but is strongly opposed by the representatives of the religions and by the practitioners of medicine who refuse to accept the role of executioners. It has also been argued that the relationship of euthanasia to suicide is not as close as it seems because a degree of uncertainty of outcome is inherent in most suicidal acts.

STATISTICS

Suicide rates. The suicide rate is the number of suicides per 100,000 population from 15 years of age, suicide being extremely rare below that age. Table 1 illustrates the way in which suicide rates are tabulated and how their fluctuations are recorded. Yearly averages are compared because the data from a single year can be misleading. In all the countries listed in the Table the male suicide rate was higher than the female, but the degree of the difference varied greatly. In several countries it was smaller in the later than in the earlier of the two periods, a trend that has been observed since the beginning of this century.

Decline of overall suicide rates

The correlations between the suicide rates and certain social characteristics of populations have been studied by sociologists. The suicide rates have been found to be positively correlated with the following factors: male sex, increasing age, widowhood, single and divorced states, childlessness, high density of population, residence in big towns, a high standard of living, economic crisis, drug and alcohol consumption, history of a broken home in childhood, mental disorder, and physical illness. Among factors inversely related to the suicide rates are female sex, youth, rural occupation, religious devoutness, the marriage state, a large number of children, membership of the lower socioeconomic classes, war. These correlations are not ubiquitous nor are they immutable. In a study carried out in Nigeria, for instance, unmarried men were found to be less liable to suicide than married men. In the United States, the suicide rates among higher socioeconomic classes were not found to be in excess of those of the lower income groups, although blacks overall have lower rates than whites. The studies resulting in these findings, however, were made in limited areas of the countries concerned, and their results cannot be generalized.

Validity of statistical data

Recently, the validity of the statistical data concerning suicide and the comparability of national suicide rates have been questioned. A worldwide survey carried out with the help of a questionnaire revealed great differences in the methods and criteria of certification of suicide. For national suicide rates to be comparable, not only the methods of ascertainment ought to be equally reliable but also the population census. In England and Wales, for instance, an autopsy is obligatory in every case in which a coroner's inquest is carried out. The legal procedures in Scotland are different, and autopsy is not obligatory. This means that the suicide rates of those two parts of the United Kingdom are not strictly comparable. In any community the social pressures against the categorization of a death as suicide are considerable. Most families want to avoid such a declaration. The need for caution in the evaluation of national differences has now been officially recognized. The introduction to the suicide mortality statistics of the World Health Organization for 1968 contains the following warning:

> The true incidence of suicide is hard to ascertain. Varying methods of certifying causes of death, different registration and coding procedures, and other factors affect the extent and completeness of coverage, making international comparisons impracticable.

Recently, even the comparability of regional suicide rates in the same country has been called into question. A comparative study of drug deaths handled by 19 coroners' offices in the United States and Canada revealed striking differences in the proportions allocated to categories of suicidal, accidental, and undetermined deaths, respectively. Similar differences were found in England when the ratios between suicide verdicts and open verdicts given by coroners of four neighbouring large cities were compared. When a Los Angeles coroner was joined by behavioral scientists who carried out a "psychological autopsy," the proportion of "undetermined" deaths was reduced significantly.

It would be a mistake to dismiss the statistical method in suicide research out of hand. But it is in need of critical scrutiny, and the comparability of data ought to be checked before conclusions about the significance of differences are drawn. Uniformity of search procedures and of criteria of categorization will not be easy to achieve, but they are worth striving for.

Difficulty of measuring attempted suicides

Attempted suicide. Whereas the legal requirements concerning death by other than natural causes assure the registration of at least a large proportion of suicides, the incidence of attempted suicides is much more difficult to establish. Not only is there no machinery for their registration but there appears to be a collusion on the part of all the parties concerned—the survivors, their families, the community at large, and even medical people—to keep quiet about them whenever possible. Although in most countries attempted suicide is no longer a punishable offense, it is still frowned upon by society and causes guilt feelings to those closely involved.

To establish the approximate incidence of suicidal attempts, it is not enough to obtain the number of hospital admissions following acts of self-damage in a defined population because not all of the cases are admitted to hospital. For a survey to be comprehensive, the cooperation of doctors practicing in the area has to be sought. A study of this type carried out in Sheffield, Yorkshire (population 500,000), revealed that in 1960 and 1961 the number of attempted suicides was about ten times as high as the number of suicides registered in the same period. A survey carried out in Los Angeles had a similar result. A Swiss survey suggested a ratio of 15 to one of nonfatal to fatal suicidal acts. These surveys did not include suicidal attempts in which no doctors intervened. There is reason to believe that their number is not negligible.

Comparison of an unselected sample of suicides with one of attempted suicides shows remarkable differences with regard to sex and age. The peak age for suicides has been found to lie between 55 and 64, for attempted suicides between 24 and 44. More men than women kill themselves, but more women than men attempt suicide.

Methods of suicide and attempted suicide. Tables 2 and 3 show the differences in the methods preferred in England and Wales and the United States as well as the change in their use over a decade.

Table 2: Methods of Suicide, United States*

	male		female	
	1955	1965	1955	1965
Analgesics and narcotics	29	59	145	312
Other solid or liquid poison	31	27	90	60
Domestic gas	14	6	35	9
Other gases	95	113	57	87
Hanging and strangulation	208	158	238	124
Drowning	29	20	72	43
Firearms and explosives	525	546	252	240
Cutting and piercing instruments	28	19	32	19
Jumping from high places	26	31	50	48
Other agents	15	20	31	59

*Per 1,000 deaths.
Source: *World Health Statistics Report 21*, No. 6, Geneva: WHO (1968).

In England and Wales, 53 percent of the men and 75 percent of the women had poisoned themselves in 1955; in 1965 the figures were 75 percent and 85 percent, respectively. Domestic coal gas claimed the highest number of victims. The shift from the latter to narcotic poisoning was very conspicuous. In the United States, the trend toward self-poisoning was only slight, firearms continuing to rule supreme. Domestic gas is not universally available in the United States, but firearms are.

Table 3: Methods of Suicide, England and Wales*

	male		female	
	1955	1965	1955	1965
Analgesics and narcotics	79	246	182	486
Other solid or liquid poison	29	24	22	19
Domestic gas	412	335	553	322
Other gases	13	48	1	7
Hanging and strangulation	187	134	64	55
Drowning	92	60	106	66
Firearms and explosives	71	63	5	4
Cutting and piercing instruments	53	21	11	5
Jumping from high places	19	18	29	17
Other agents	43	50	26	19

*Per 1,000 deaths.
Source: *World Health Statistics Report 21*, No. 6, Geneva: WHO (1968).

As is to be expected, the more lethal and violent methods are used less frequently in the nonfatal than in the fatal suicidal acts. In recent years hospital admissions for narcotic poisonings have increased steadily. In England and Wales their number trebled between 1957 and 1964. Similar developments have been reported from other European countries. At the same time, the suicide rates have not increased or have even declined. This apparent paradox can be explained by the very low lethality of narcotic poisoning under conditions of modern treatment. Also, coroners and their equivalents in other countries are more inclined to give an accident verdict or declare the cause of death undetermined in cases of narcotic poisoning than in other cases of self-inflicted death.

MEANINGS AND DYNAMICS OF SUICIDAL ACTS

It is difficult, if not impossible, for the individual to conceive of his nonexistence. Generally, religion reconciles him with death and even holds out the promise of an afterlife. In man's mind, death has usually been the end as well as a new beginning. Thus, suicide is experienced as an event in the life of the person who kills himself. It has various psychological and social meanings to those who commit it.

Suicide may be the manifestation of an extreme grief reaction and mean reunion with a deceased person. It may be associated with thoughts of a new and better existence, even of rebirth.

Suicide as an act of aggression

Suicide is an aggressive act in more than one sense: it is self-assault aiming at self-destruction, but it is also meant to destroy something that is, or ought to be, of value to others. It frequently stands for self-punishment but also for revenge.

As suicide is usually dreaded by the group to which a person belongs and by society at large, it lends itself to exhibitionistic tendencies. In the suicide fantasies of children as well as of adults, these and other meanings figure prominently.

A fatal suicidal act tends to cause grief reactions and guilt feelings on the part of those who may feel that they could have prevented it by caring and loving more than they did. If the act is nonfatal, it serves as an appeal for help and may give rise to efforts at reparation; it may lead to an improvement in the relationship to the person who attempted suicide. Conscious or unconscious expectation of these responses is one of the factors underlying suicidal acts.

Sigmund Freud emphasized the aggressive nature of suicide. He believed the urge to self-preservation to be so powerful that suicide could not be explained as an act of aggression against the self but rather that it was directed against another person with whom the individual had identified himself. The main contribution of psychoanalysis to the knowledge of suicide has been its insistence on looking behind the outward appearances of suicidal behaviour and searching for motivations of which the person who commits the suicidal act may not be aware. The psychoanalytic approach has also contributed to the understanding of suicide proneness. Whereas psychoanalysis viewed suicide as the result of intrapsychic conflicts, the Austrian psychiatrist Alfred Adler and his school stressed the importance of interpersonal factors and the effect of suicide on others. The lack of social relations makes life purposeless and leads to suicide.

In most suicidal acts the conflict between life-preserving and self-destructive tendencies can be discerned. People who commit suicidal acts do not wish either to die or to live, but they want to do both at the same time, usually the one more, or much more, than the other. The outcome of the suicidal act depends largely on the relative strength of these opposing tendencies.

The decline of the suicide rates in wartime has been related to the turning of aggression toward the common enemy and away from the self. Also the greater closeness of human relations in wartime tends to reduce the social

isolation that is regarded as the most important causative factor in suicide.

Suicide proneness. There is no situation and no experience to which suicide is the only possible reaction. This is why a suicidal act cannot be explained by the situation alone and why it has to be postulated that some people are more suicide prone than others. A broken home in childhood predisposes to suicidal acts, but there are additional personality factors.

Suicide and mental disorder

About one-third of the people who kill themselves have been found to have been suffering from mental illness requiring psychiatric treatment. Most of the rest have unstable and vulnerable personalities. Among the mental disorders, depressive illness carries the highest suicidal risk. A feeling of worthlessness and despair and a wish to die belong to the clinical picture of this illness, but even in this condition the urge to commit suicide varies. Other major mental disorders, such as schizophrenia, also have an above-average incidence of suicide. Abnormal personalities with aggressive tendencies show increased liability to suicide. The incidence of suicide among murderers has been found to be excessive.

Sociologists who rely on the statistical method alone have denied the importance of mental disorder in the causation of suicide. This is not surprising, because mental illness in the population cannot be quantified. It is nevertheless an established fact that the incidence of suicidal acts among the psychiatrist's clientele is excessive. This does not mean that every person who commits such an act is psychiatrically ill in the clinical sense, although even in the absence of mental illness a suicidal act is an abnormal reaction to stress, abnormal being defined statistically—that is, not in keeping with the reactions of the large majority of persons to comparable situations.

The recognition of the importance of psychopathology in the causation of suicide does not invalidate sociological theories. The motivations underlying suicidal acts carried out by the mentally ill are fundamentally the same as in the general population, although they are derived from disorders of mood or from such abnormal contents of thought as delusions.

Theories of suicide. *Sociological theories.* The best known theory is Durkheim's, according to which the suicidal tendency is the result of collective forces acting upon the individual. Its strength varies inversely with the degree of cohesion that exists in society. Durkheim's classification of suicides into three types, mentioned above, was derived from the basic theory. Although he rejected psychology and psychopathology, his classification was essentially an excursion into social psychology. His basic theory has been criticized as being untestable, but in modified form it has been advocated by some more recent scholars. These scholars believe that disruptions of social relations are the causative factor in suicide and that the greater the incidence of such disruptions in a population the higher the suicide rate. Most suicide victims have experienced a set of disrupted social relations that is not found in the history of nonvictims with equal frequency.

Suicide and social disruption and isolation

Other scholars have attempted to combine the sociological with the psychodynamic approach. They have made four basic assumptions: aggression is due to frustration; business cycles change the hierarchical social status of persons and groups; loss of status causes frustration; suicide is an aggressive response to frustration generated by the flow of economic forces.

These and allied sociological theories depend largely on the reliability and validity of statistical data.

Psychological and social–psychological theories. The Freudian and Adlerian propositions, based on study of individuals, belong here. It can be seen that Adler's ideas about the causation of suicide tally with the sociological theories about the role of social isolation.

The concept of hope has been made the basis of a theory of suicide. According to this theory, suicide occurs when the life outlook is one of despairing hopelessness. Whether or not a person will commit a suicidal act in such a situation depends on a personality factor called the sense of competence, which is inversely related to the likelihood of suicide.

Other theories. The familial occurrence of suicides has given rise to the hypothesis that suicidal behaviour may be an innate inherited behaviour pattern. Studies in identical twins one of whom had committed suicide failed to confirm this hypothesis. The tendency of suicides to occur in families must therefore be explained by the mental mechanism of "identification" with a suicidal person. Seasonal fluctuations of the suicide rates, with a peak in spring, have remained unexplained. They may be related to depressive illness, which also shows the highest incidence in spring.

Existential views

Despair and the idea of death hold a central position in Existential philosophy, which has exerted a considerable influence on many European and some American psychiatrists and anthropologists. A Swiss psychiatrist described in great detail the experiences of a patient whose suicide appeared to follow inevitably from her way of "being in the world." Another author, a lay psychotherapist strongly influenced by Existentialism, rejects the principle of suicide prevention as a medical prejudice against death and advocates suicide if death is needed by the soul. There is nothing new in these ideas.

Treatment and prevention. Suicide presents social and medical problems, and no single approach can be expected to succeed in substantially reducing its incidence. Early recognition and treatment of mental disorders are important. There has been much progress in the treatment of depressive illness. But even if the latter could be eliminated, a sizable suicide problem would remain.

Preventive measures are called for on several levels, as stated in *Prevention of Suicide* (WHO Public Health Papers No. 35, 1968): (1) Medical emergency services are required to prevent the fatal outcome of suicidal acts, especially self-poisoning. (2) Persons who have made a previous attempt are more likely to die through suicide than those who have not. This high-risk group needs special medical and social attention over long periods. (3) Recognition of suicidal intent and provision of support and treatment, if necessary, can prevent suicide. The majority of people who commit suicidal acts have given a previous warning that was ignored. (4) Measures against social isolation and the furtherance of satisfying social relations within and outside the family appear to provide the most effective prophylaxis against suicidal tendencies.

In the last two decades special centres and organizations for the prevention of suicide have been created in many countries. Most of them are not under medical direction, though all of them have medical consultants. The telephone is used as a means of communication between lonely and desperate persons in need of support and helpers available day and night. There is evidence that this kind of service may help to avert suicidal acts. The best known of these telephone services is that of the "Samaritans," founded by Chad Varah, an Anglican minister, in London in 1953. They aim at giving psychological first aid as well as support by befriending the lonely and desperate. In 1971 this organization had 118 branches in the United Kingdom and nine in Commonwealth countries. Similar organizations have sprung up in other countries.

SUICIDE IN THE FUTURE

Suicide is a peculiarly human problem. There is no evidence that what seems to be self-destructive behaviour in animals is associated with the wish to die. At some stage of evolution, man must have discovered that he can kill not only animals and fellowmen but also himself. It is difficult to imagine man in the future without this potential, which has sometimes been proclaimed as one of the basic human freedoms. Suicidal acts will continue to be committed. There are no prospects that an antisuicidal substance will be discovered in the foreseeable future.

There are certain aspects of present-day society that make a substantial decline in the incidence of suicidal acts unlikely. The steady increase in the number of the aged, who contribute the majority of suicides, will make it difficult to prevent a rise in the suicide rates, especially if the incidence of suicide among the young should continue to go up. The young are both more openly aggressive against society and more demonstrative today than they

were in the first half of the 20th century. Both of these behaviour features contribute to suicidal acts. Prosperity has not proved an antidote against them. It has in fact made possible the emergence of subcultures of alienation with a tendency to withdrawal from reality with the aid of drugs. These movements, though allegedly dedicated to the enjoyment of life, have a strong self-destructive component. The continued increase in population will make it more difficult for the underendowed to hold their own. Economic crises are likely to multiply and to take their toll of suicide victims.

The outlook regarding suicide would be very black indeed if it were not for the strong urge to self-preservation inherent in the individual and in society. The emergence of nonprofessional movements toward self-help in the struggle against suicide can be seen as a manifestation of this tendency.

BIBLIOGRAPHY. There has been a veritable explosion of suicide research and of planned efforts at suicide prevention involving the community as a whole since the end of World War II. The classic sociological approach as presented in EMILE DURKHEIM, *Le Suicide* (1897; Eng. trans., 1951), and brought up to date in JACK P. GIBBS and WALTER T. MARTIN, *Status Integration and Suicide* (1964), was supplemented by ecological studies, such as PETER SAINSBURY, *Suicide in London* (1955); and an early study in Chicago, RUTH S. CAVAN, *Suicide* (1928). KARL A. MENNINGER, *Man Against Himself* (1938), had prepared the ground for the introduction of psychodynamics into this area. LOUIS T. DUBLIN, *Suicide* (1963), is the best link between the classical statistical and the various modern approaches. ANDREW F. HENRY and JAMES F. SHORT, JR., *Suicide and Homicide* (1954), combines sociology and psychodynamics.

Research on attempted suicide has been in the forefront of interest since the publication of KARL G. DAHLGREN, *On Suicide and Attempted Suicide* (1945); ERWIN STENGEL and NANCY G. COOK, *Attempted Suicide* (1958); and NORMAN L. FARBEROW and EDWIN S. SHNEIDMAN (eds.), *The Cry for Help* (1961). The most recent survey of the multifarious research and preventive activities in this area is *Suicidal Behaviors*, ed. by H.L.P. RESNIK (1968). Stengel has specialized in the follow-up and the study of the social effects of suicidal attempts, a line of research presented by him in *Suicide and Attempted Suicide*, rev. ed. (1969); and by NILS RETTERSTOL in *Long-term Prognosis After Attempted Suicide* (1970). These workers owe a great deal to epidemiology, which has introduced systematic follow-up (cohort) methods into the study of abnormal behaviour. Modern suicide research has also questioned the basic premises of sociological pioneers. The inviolable status of the suicide rate has been challenged by J.D. DOUGLAS in *The Social Meanings of Suicide* (1967).

There has been progress in the anthropological approach also. PAUL BOHANNAN (ed.), *African Homicide and Suicide* (1960); RAYMOND W. FIRTH, *Suicide and Risk-Taking in Tikopia Society* (1961); and MAMORU IGA, "Japanese Adolescent Suicides and Social Structure," in *Essays in Self-destruction*, ed. by EDWIN S. SHNEIDMAN (1967), dispel old myths and break new ground. Suicide prevention is no longer the preserve of sociologists and psychiatrists. It has become a matter of public concern spearheaded by lay organizations, described in EDWARD CHAD VARAH (ed.), *The Samaritans: To Help Those Tempted to Suicide or Despair* (1965). Meanwhile, those who want to preserve and promote suicide as one of the fundamental human freedoms have not remained idle either. In *Euthanasia and the Right to Death*, ed. by A.B. DOWNING (1969), the case for a gentle death when desired is made by distinguished advocates of voluntary euthanasia.

(E.St.)

Sui Wen Ti

As emperor of China, Sui Wen Ti (in Pin-yin romanization Sui Wen Di; personal name Yang Chien), with his ministers, reunified that country after it had experienced 300 years of instability, reorganized the institutions of state and society, and established China's second great Imperial order—that of the Sui (581–618) and T'ang (618–907) dynasties.

Wen Ti was born in 541 into a powerful family that had held high office under the non-Chinese dynasties that controlled North and Central China in the period of fragmentation. His ancestors had married into prestigious non-Chinese clans, and the culture of such families was as mixed as their descent. Until he reached the age of 13, Wen Ti was brought up by a Buddhist nun. He then briefly attended the school maintained by the state for the education of sons of nobles and officials; it is doubtful that he acquired more than the mere rudiments of composition, a bit of history, and some maxims of Confucian morality, for young men of his class devoted themselves chiefly to horsemanship, the hunt, falconry, archery, and military exercises.

He received his first military appointment at 14 and rose rapidly in the service of the Yü-wen, the non-Chinese ruling house of the Northern Chou dynasty (557–581), who, with their military prowess, would soon control all of North China. Wen Ti held a command in the campaign against the dynasty which controlled the northern plain and a post in the administration of the conquered territory. He had done well for the Chou, and, as part of his reward, he was permitted to marry one of his daughters to the Chou crown prince. When the Emperor unexpectedly became ill and died at 36 and the Crown Prince's sanity became doubtful, Wen Ti, his wife, and their confidants decided to seize the throne. The summer of 580 was crucial, for rival contenders and Chou loyalists rose in many places. But, with luck, ruthlessness, superior military force, and discord among his rivals, Wen Ti prevailed. He assumed the Imperial title, held an audience on March 4, 581, and Sui dynasty was founded.

Seizure of the throne

Wen Ti surrounded himself with able men, mostly of mixed descent and mostly from backgrounds similar to his own. An early move was the building of a new capital on a new site southeast of the Han capital of Ch'ang-an; it was built on a scale unprecedented in Chinese history. The Sui evidently meant to replace the weak regimes of the age of disunion with strong centralized government, to unify China by eliminating the feeble "legitimate" Chinese regime at Nanking, and to reassert Chinese power throughout east Asia. Wen Ti moved into his half-built capital in 583, and he immediately set his grand design in motion. Centralization required drastic reforms on many levels—for example, the entrenched families that held local office by hereditary right had to be replaced by a bureaucracy answerable to the throne. The hereditary rights and the institutions that supported them were quickly abolished; a method of selecting new men by examination and recommendation was devised; appointment powers were vested in the Board of Civil Office (Chinese Li Pu); and the "rule of avoidance" was instituted, forbidding officials to serve in their native places. Wen Ti planned the conquest of the south with his usual care and attention to detail. The eight-pronged assault by land and water overwhelmed the southerners; the integration of this culturally different area into the Sui empire began and was greatly facilitated by the canal system that Wen Ti had begun. Beyond China proper Sui power was less easily asserted against the formidable empires of the western and the eastern Turks, but fortune and Sui intrigues brought success; the Turkish empires were weakened by internal rivalries, and by 603 the Sui had broken Turkish power in the areas most vital to them: Turkistan and Mongolia. A Sui attempt to administer Vietnam was a failure, but, toward the close of Wen Ti's reign, Korea and Japan were beginning to notice the new paramount power in east Asia.

Foreign affairs

In the year 601, when Wen Ti was 60, he had solid grounds for satisfaction: the empire was reunified and at peace; the people were productive, and the officials—carefully selected, frequently rotated, and under constant checks—collected taxes, saw that the granaries were filled, and carried out imperial orders at the local level. Looking about him in his spacious capital city, Wen Ti could see a large and increasing population, the opulent mansions of his nobles and ministers, temples, and thriving marketplaces. Moreover, the arrival of tribute missions reminded him that Sui power was being felt by neighbouring peoples. Yet, for all his accomplishments, Wen Ti was deeply unhappy. Henpecked by his aging wife, on bad terms with his sons, deprived of many of his life-long confidants by death or by his wife's jealousy of them, haunted by feelings of guilt and nameless fear, he turned against state Confucianism and ever more ardently to Buddhism. On his birthday in 601, he began an

eaborate empire-wide series of observances. Shrines were built in key cities and towns; then the Emperor himself sealed holy relics in jars, which delegations of eminent monks carried into the provinces. At a set time throughout the empire, the relics were simultaneously enshrined with appropriate ceremony. By this act of grandiose public piety, Wen Ti followed in the footsteps of the great 3rd-century-BC Indian emperor Aśoka, who was, like himself, a unifying emperor. At the time he assuaged his feelings of fear and guilt and laid in a great store of spiritual merit (*karman*) to see him through the lives to come.

Three years later—at the end of one of the great reigns in Chinese history—he fell ill and died, his end possibly hastened by the son who succeeded him.

BIBLIOGRAPHY. ARTHUR F. WRIGHT, "The Formation of Sui Ideology, 581–604," in JOHN K. FAIRBANK (ed.), *Chinese Thought and Institutions*, pp. 77–84 (1957), deals with Wen Ti's life and with his principal advisors.

(A.F.W.)

Sukarno

Sukarno, known by no other name but by a score of titles, including engineer, doctor, proclaimer of independence, and great leader of the revolution, dominated Indonesian national life and politics for 40 years. He did so with unfailing panache—from the time he emerged to sudden prominence in Dutch colonial Bandung (1927) as a dazzling young rake, firebrand orator, and freedom fighter until he was displaced by General Suharto after being implicated directly in a conspiracy to convert the Republic of Indonesia into a people's republic linked to Communist China.

Magnum



Sukarno, 1965.

Early life. Sukarno was born on June 6, 1901, in Surabaja, East Java, in the then Dutch East Indies, the only son of a poor Javanese schoolteacher, Raden Sukemi Sosrodihardjo, and his Balinese wife, Ida Njoman Rai. Originally named Kusnasosro, he was renamed Sukarno for better omen after a series of illnesses. Known to his childhood playmates as Djago (Cock, Champion) for his looks, spirits, and prowess, he was as an adult best known as Bung Karno (*bung*, "brother" or "comrade"), the revolutionary hero and architect of *merdeka* ("independence").

Sukarno spent long periods of his childhood with his grandparents in the village of Tulungagung, where he was exposed to the animism and mysticism of serene rural Java. There he became a lifelong devotee of *wajang*, the puppet shadow plays based on the Hindu epics, as animated and narrated by a master puppeteer, who could hold an audience spellbound through an entire night. As a youth of 15, Sukarno was sent to secondary school in Surabaja and to lodgings in the home of Omar Said Tjokroaminoto, a prominent civic and religious figure. Tjokroaminoto treated him as a cherished foster son and protégé, financed his further education, and eventually married him off at age 20 to his own 16-year-old daughter, Siti Utari.

Sukarno's education

As a student, Sukarno chose to excel mainly in languages. He mastered Javanese, Sundanese, Balinese, and modern Indonesian, which, in fact, he did much to create. He also acquired Arabic, which, as a Muslim, he learned by study of the Qur'ān; Dutch, the language of his education; German; French; English; and, later, Japanese. In Tjokroaminoto's home he came to meet, admire, and sometimes to understudy emerging leaders who spanned the rapidly widening national political spectrum, from feudal princelings to fugitive Communist conspirators. The eclectic syncretism of the Tjokroaminoto ménage, like the romance and mysticism of *wajang*, imprinted itself indelibly upon Sukarno's mind and personality. He was later to treat nation making as a heroic theatrical, in which the clash of irreconcilable men and ideas could be harmonized through sheer poetic magic—his own.

Endowed with commanding presence, radiant personality, mellifluous voice, vivid style, a photographic memory, and supreme self-confidence, Sukarno was obviously destined for greatness. In Bandung in 1927, where he had just acquired a degree in civil engineering, he found his true calling in oratory and politics. He soon revealed himself as a man of charisma and destiny.

In both early and later life, Sukarno's amours were almost as politically significant as his oratory. He divorced Siti in 1923 and married his well-to-do and newly divorced landlady, Inggit Garnisih, a woman of 34. He divorced the childless and aging Inggit in 1943 and married the beautiful young Fatmawati, by whom he had five children, including his eldest son, Guntur Sukarnaputra (born 1944). In 1954 he surreptitiously took a second wife, Hartini (as a Muslim, he was entitled to four wives), a dubious divorcée who bore him two children. In the early 1960s he also married Ratna Sari Dewi, a Japanese who bore him a daughter in 1967. Next was the comely young Harjati, then the yet younger and comelier Yurike Sanger. Sukarno also engaged in a lifelong series of liaisons featuring Dutch, Russian, American, Chinese, Japanese, Filipino, and Mexican beauties, among many others.

Imprisonment by the Dutch

Indonesian independence. For his challenge to colonialism Sukarno spent two years in a Dutch jail (1929–31) in Bandung and over eight years in exile (1933–42) on Flores and Sumatra. When the Japanese invaded the Indies in March 1942, he welcomed them as personal and national liberators. During World War II the Japanese made Sukarno their chief advisor and propagandist, their recruiter for labourers, soldiers, and prostitutes. Sukarno pressured the Japanese to grant Indonesia its independence and, on June 1, 1945, made the most famous of many celebrated speeches. In it he defined the Pantjasila, or Five Principles (nationalism, internationalism, democracy, social prosperity, and belief in God), still the sacrosanct state doctrine. When the collapse of Japan became imminent, Sukarno at first wavered. Then, after being kidnapped, intimidated, and persuaded by activist youths, he declared Indonesia's independence (August 17, 1945). As president of the shaky new republic, he fuelled a successful defiance of the Dutch, who, after two abortive "police actions" to regain control of their former colony, formally transferred sovereignty on December 27, 1949.

Life at the Sukarno court

From his revolutionary capital in Jogjakarta Sukarno returned in triumph to Djakarta on December 28, 1949. There he established himself, his collection of paintings, and his numerous retinue in the splendid palace of the Dutch governors general. He proceeded to preside urbanely over a spectacle that was at once diverting and disturbing. His increasingly numerous and outspoken critics maintained that Sukarno inspired no coherent programs of national organization and administration, rehabilitation, and development, such as were quite clearly necessary. He seemed instead to conduct a continuous series of formal and informal audiences and a nightly

soiree of receptions, banquets, music, dancing, movies, and *wajang*. Indonesian politics became increasingly frenzied, with Sukarno himself engaged in devious manoeuvres that made stabilization impossible. The Indonesian economy foundered while Sukarno encouraged the wildest of extravagances. To be sure, the nation scored impressive gains in health, education, and cultural self-awareness and self-expression. It achieved, in fact, what Sukarno himself most joyously sought and acclaimed as "national identity," an exhilarating sense of pride in being Indonesian. But it did so at a ruinous cost.

After "dreaming" in late 1956 of "burying" the feuding political parties in Indonesia and thus achieving national consensus and prosperity, Sukarno dismantled parliamentary democracy and destroyed free enterprise. He ordained "Guided Democracy" and "Guided Economy" for the achievement of MANIPOL-USDEK and RESOPIMNASAKOM—arcane acronyms symbolizing policies but signifying dictatorship.

Sukarno's personal and political excesses, as epitomized eventually by his neo-Marxist, crypto-Communist ideology and his infamous cabinet of 100 corrupt and cynical ministers, induced a continuous state of national crisis. Sukarno narrowly escaped recurrent attempts at assassination, the first in 1957. Regional insurrections broke out in Sumatra and Sulawesi in 1958. Inflation escalated the cost-of-living index from 100 in 1958 to 18,000 in 1965 and on up wildly to 600,000 in 1967. In 1963, after shouting repeatedly "To hell with your aid" (1950–65 total: $1,000,000,000 U.S.) Sukarno all but broke with the United States. After having exacted $1,000,000,000 U.S. in Soviet armaments and other items, he next affronted Moscow. On January 20, 1965, Indonesia formally withdrew from the United Nations because the latter supported Malaysia, which Sukarno had vowed to "crush" as "an imperialist plot of encirclement." Yet, until 1965, Sukarno was still able to stir the Indonesian masses to near-hysterical belligerency. Millions of Indonesians sang and shouted his slogans, "volunteered" for his causes, and acclaimed Sukarno as "Great Leader of the Revolution," "Lifetime President" (his official title), oracle and warrior of the NEFOS—his acronym for the "New Emerging Forces"—in violent conflict with NEKOLIM—the neocolonialism, capitalism, and imperialism of the "doomed" Western powers.

The coup of 1965. The nation was shocked and shaken out if its trance by a coup of September 30, 1965, in which Sukarno himself was clearly implicated. A clique of Communist and military conspirators caused six top army generals to be kidnapped, tortured, mutilated, and murdered. They seized a few key urban points and proclaimed a new revolutionary regime. General Suharto, the commander of the Djakarta garrison, swiftly reversed the coup. There ensued an oblique contest for power between Suharto and Sukarno while some 300,000 or more Communists and alleged Communists were being slaughtered. As the nation recoiled in horror, activist youths demanded the political demise of Sukarno, the Sukarnoists, and Sukarnoism and the total reform and reorganization of the state. On March 11, 1966, Sukarno was obliged to delegate wide powers to Suharto, who subsequently became acting president (March 1967) and then president (March 1968), as Sukarno sank into disgrace and dotage.

At 7:00 AM on June 21, 1970, aged 69, Sukarno died in the Central Military Hospital of his old kidney complaint and many complications. Suharto decreed a quick and quiet funeral. Nevertheless, at least 500,000 persons, including Inggit (aged 83), Fatmawati, Hartini, Dewi, their children, and virtually all of Djakarta's important personages, turned out to pay their last ambivalent respects. Next day another 200,000 assembled in Blitar, near Surabaja, for the official service followed by burial in a simple grave alongside that of his mother. The cult and ideology of Sukarnoism are now proscribed. The late Sukarno record is now deeply deplored. But the luminous Bung Karno of 1927 and 1949, the Bapak Republick Indonesia, survives the great leader of the 1960s, by then so sadly tarnished and diminished.

BIBLIOGRAPHY. BERNHARD DAHM, *Sukarnos Kampf um Indonesiens Unabhangigkeit* (1966; Eng. trans. rev. and updated, *Sukarno and the Struggle for Indonesian Independence*, 1969), deals with Sukarno's politics to 1950. See also WILLARD A. HANNA, "Sukarno: The Devolution of a Revolutionary," in *Eight Nation Makers* (1964); JOHN HUGHES, *Indonesian Upheaval* (1967); and *Sukarno: An Autobiography, As Told to Cindy Adams* (1965).

(W.A. Ha.)

Süleyman I the Magnificent

Süleyman I, known to the Turks as Kanuni, the Lawgiver, and to the Christian world as the Magnificent, sultan of the Ottoman Empire from 1520 to 1566, extended Turkish influence in both Europe and Asia; his reign witnessed the unfolding of all that was most characteristic and brilliant in Ottoman civilization. The only son of Sultan Selim I, Süleyman was born in November 1494 or April 1495. He became *sancak beyi* (governor) of Kaffa in the Crimea during the reign of his grandfather Bayezid II and of Manisa in western Asia Minor in the reign of Selim I.

Süleyman succeeded his father as sultan in September 1520 and began his reign with campaigns against the Christian powers in central Europe and the Mediterranean. Belgrade fell to him in 1521, and Rhodes, long under the rule of the Knights of St. John, in 1522. At Mohács, in August 1526, Süleyman broke the military strength of Hungary, the Hungarian king, Louis II, losing his life in the battle. The vacant throne of Hungary was now claimed by Ferdinand I, the Habsburg archduke of Austria, and by John (János Zápolya), who was *voivode* (lord) of Transylvania and the candidates of the "native" party opposed to the prospect of Habsburg rule. Süleyman agreed to recognize John as a vassal king of Hungary, and in 1529, hoping to remove at one blow all further intervention by the Habsburgs, he laid siege to Vienna. Difficulties of time and distance and of bad weather and lack of supplies, no less than the resistance of the Christians, forced the Sultan to raise the siege. The campaign was successful, however, in a more immediate sense, for John was to rule thereafter over most of Hungary until his death, in 1540. A second great campaign in 1532, notable for the brilliant Christian defense of Güns, ended as a mere foray into Austrian border territories. The Sultan, preoccupied with affairs in the East and convinced that Austria was not to be overcome at one stroke, granted a truce to the archduke Ferdinand in 1533.

By courtesy of the trustees of the British Museum



Süleyman I (centre), detail of an engraving of a panel by Pieter Coecke van Aelst showing a procession through Istanbul, 1533.

The death of John in 1540 and the prompt advance of Austrian forces once more into central Hungary drove Süleyman to modify profoundly the solution that he had imposed in the time of John. His campaigns of 1541 and 1543 led to the emergence of three distinct Hungarys—Habsburg Hungary in the extreme north and west; Ottoman Hungary along the middle Danube, a region under direct and permanent military occupation by the Ottomans and with its main centre at Buda; and Transylvania, a vassal state dependent on the Porte and in the hands of John Sigismund, the son of John Zápolya. Between 1543 and 1562 the war in Hungary continued, broken by truces and with few notable changes on either side; the most important was the Ottoman capture of the Banat of Temesvár (Timişoara) in 1532. After long negotiations a peace recognizing the status quo in Hungary was signed in 1562.

The campaigns against Persia

Süleyman waged three major campaigns against Persia. The first (1534–35) gave the Ottomans control over the region of Erzurum in eastern Asia Minor and also witnessed the Ottoman conquest of Iraq, a success that rounded off the achievements of Selim I. The second campaign (1548–49) brought much of the area around Lake Van under Ottoman rule, but the third (1554–55) served rather as a warning to the Ottomans of the difficulty of subduing the Ṣafavid state in Persia. The first formal peace between the Ottomans and the Ṣafavids was signed in 1555, but it offered no clear solution to the problems confronting the Ottoman sultan on his eastern frontier.

The naval strength of the Ottomans became formidable in the reign of Sultan Süleyman. Khayr ad-Dīn, known in the West as Barbarossa, became *kapudan* (admiral) of the Ottoman fleet and won a sea fight off Preveza, Greece (1538), against the combined fleets of Venice and Spain, which gave to the Ottomans the naval initiative in the Mediterranean until the Battle of Lepanto in 1571. Tripoli in North Africa fell to the Ottomans in 1551. A strong Spanish expedition against Tripoli was crushed at Jarbah (Djerba) in 1560, but the Ottomans failed to capture Malta from the Knights of St. John in 1565. Ottoman naval power was felt at this time even as far afield as India, where a fleet sent out from Egypt made an unsuccessful attempt in 1538 to take the town of Diu from the Portuguese.

The later years of Süleyman were troubled by conflict between his sons. Mustafa had become by 1553 a focus of disaffection in Asia Minor and was executed in that year on the order of the Sultan. There followed during 1559–61 a conflict between the princes Selim and Bayezid over the succession to the throne, which ended with the defeat and execution of Bayezid. Süleyman himself died during the night of September 5/6, 1566, while he was besieging the fortress of Szigetvár in Hungary.

Domestic achievements

Süleyman's reign was notable for more than military and naval exploits. He surrounded himself with administrators and statesmen of unusual ability, men such as his grand viziers (chief ministers) İbrahim, Rüstem, and Mehmed Sokollu. *'Ulama'* (specialists in Islāmic law), notably Abū al-Su'ūd (Hoca Çelebi) and Kemalpaşâzade, made the period memorable, as did the great Turkish poet Bâkî and the architect Sinan. Süleyman built strong fortresses to defend the places he took from the Christians and adorned the cities of the Islāmic world (*e.g.*, Mecca, Damascus, and Baghdad) with mosques, bridges, aqueducts, and other public works. In Istanbul, Süleyman's capital, the architect Sinan built several famous mosques for him, including the one bearing his name (Süleymaniye Cami), and others for his viziers Rüstem and Mehmed Sokollu. In general, Süleyman completed the task of transforming the previously Byzantine city of Constantinople into Istanbul, a worthy centre for a great Turkish and Islāmic empire.

Süleyman was, to his subjects, the "Lawgiver." The edicts bearing his name, though for the most part interpretative rather than original, nevertheless represent a sustained effort to adjust the complex structure of the Ottoman state to the circumstances of a new age. To his contemporaries in Europe Süleyman was "il magnifico," "the magnificent sultan." It was the Venetian Andre Navagero who paid him the rarest of compliments when he wrote that Süleyman, provided he were well informed, did injustice to no one.

BIBLIOGRAPHY. R.B. MERRIMAN, *Suleiman the Magnificent, 1520–1566* (1944, reprinted 1966); see also *The New Cambridge Modern History*, ch. 17 (1958).

(V.J.P.)

Sulfide Minerals

Sulfide minerals are a large and important group of compounds of great value to man. Although many sulfide minerals are known, their abundance is considerably less than that of some other mineral types, such as the silicate minerals and phosphate minerals (*qq.v.*). Included within the sulfide mineral group, although not sulfides in the chemical sense, are metallic arsenides, tellurides, antimonides, and selenides.

Sulfide minerals are the source of many common base and precious metals and, with the exception of a few metals such as iron, titanium, tin, and gold, are the ore minerals of most metals used in technology. They are the ore minerals of lead, zinc, copper, nickel, silver, antimony, bismuth, and others and also are the source of many less common metals such as cadmium, indium, rhenium, and selenium. These rarer metals occur in trace amounts in many common sulfides and are recovered in smelting and refining processes.

Most of the sulfide minerals are simple structurally and show high symmetry in their crystal forms. They have many of the properties of metals, such as metallic luster and electrical conductivity, characteristics that are attributed to metal–metal bonding in their structures.

The elements that occur as sulfide minerals have been termed chalcophile ("copper-loving") by Victor Goldschmidt. Naturally occurring sulfides are known for the elements: Silver, Ag; iron, Fe; lead, Pb; arsenic, As; rubidium, Ru; copper, Cu; cobalt, Co; mercury, Hg; antimony, Sb; tin, Sn; thallium, Tl; nickel, Ni; manganese, Mn; bismuth, Bi; molybdenum, Mo; gold, Au; zinc, Zn; calcium, Ca; platinum, Pt; tungsten, W; and cadmium, Cd. Elements that substitute for sulfur in the sulfide minerals are selenium, Se; tellurium, Te; As; Sb; and Bi. There are, in addition, other elements that, although not occurring as sulfide minerals, show a strong affinity for sulfide minerals and occur as important trace elements. Among these, gallium, Ga; indium, In; germanium, Ge; rhenium, Re; chromium, Cr; niobium, Nb; and tantalum, Ta, are important.

The geochemical environment in which sulfide minerals form is a broad one and ranges from high temperatures and pressures that exist at great depth in the earth's crust to the reducing environment at the surface in restricted basins or swamps. In addition, several sulfide minerals are known from meteorites (*q.v.*). Because of the chemically reduced state of sulfur in these minerals, they tend to be unstable in the oxidizing conditions present at the earth's surface. Air and water, probably aided by organic compounds and bacteria, act to decompose most sulfides within relatively short periods of geological time.

Except for dissemination in some sedimentary rocks, these minerals tend to occur in isolated concentrations. These concentrations make up mineral bodies such as veins or fracture fillings, pipelike bodies, replacements of pre-existing rocks in the shape of blankets and, where man has found it economical to extract the minerals, they are known as ore bodies. Compositionally, most mineral bodies tend to represent concentrations of a discrete metal sulfide or combinations of certain metals. Thus, one refers to the major copper sulfide concentrations of the southwestern U.S., of South Africa, and of the Kazakh S.S.R.; the lead-zinc sulfide concentrations of the Mississippi Valley, U.S., and of Morocco; and the deposits of silver-bearing minerals of western Mexico, to mention only a few.

This article treats the physical properties and chemical composition of the sulfide minerals, and their classification, genesis, and occurrence in nature. For further information on the economic importance and occurrence of sulfide minerals see ORE DEPOSITS; see also MINERALS

for the relation of sulfide minerals to other minerals, and GEOCHEMICAL EQUILIBRIA AT HIGH TEMPERATURES AND PRESSURES for a general discussion of phase equilibria of relevance.

CLASSIFICATION OF SULFIDE MINERALS

Unlike silicate minerals, for which a general classification based on crystal structure exists, sulfide minerals are classified basically according to their chemical compositions or formula type, and are subdivided within the type on the basis of their structures. Nearly all of the sulfide minerals have structural arrangements that belong to six basic types, four of which are important. These arrangements are close-packing combinations of metal and sulfur, governed by ionic size and charge.

Structural types

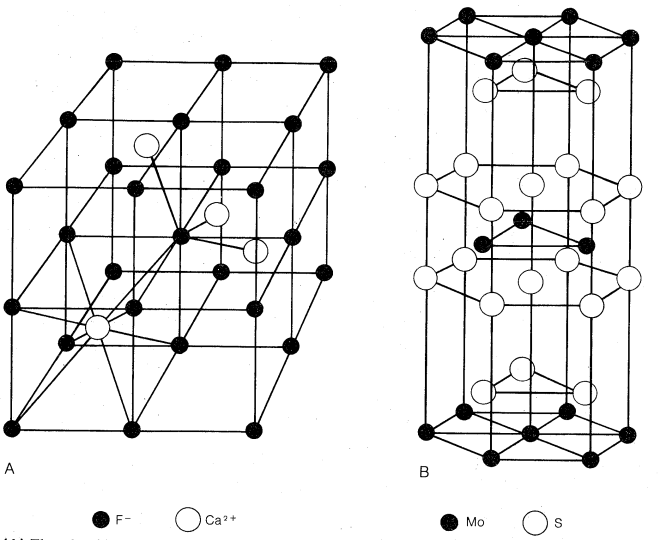
The simplest and most symmetrical of the four important structural types is the sodium chloride structure, in which each ion occupies a position within an octahedron composed of six oppositely charged neighbours. The most common sulfide mineral crystallizing in this manner is the ore mineral of lead, galena (PbS).

A type of packing that involves two sulfide ions in each of the octahedral positions in the sodium chloride structure is the pyrite structure. This is a high symmetry structure and is that of the common iron sulfide, pyrite (FeS_2). A group of more complex sulfide minerals also exhibit this structure. Ullmannite, (Ni, Co, Fe) (Sb,As,Bi)S, in which sulfide pairs in the pyrite structure are replaced by the B subgroup metal Sb, As, or Bi in combination with a single sulfide ion, is an example.

The second discrete structural type is that of sphalerite, ZnS in which each metal ion is surrounded by six oppositely charged ions arranged tetrahedrally. In addition to sphalerite, greenockite, CdS, the sulfide ore mineral of cadmium, alabandite, the sulfide of manganese, and the selenide and telluride of mercury all crystallize with this structure.

A third structural type is that of the mineral fluorite. In this structure, the metal cation is surrounded by eight anions, and each anion in turn is surrounded by four metal cations. The reverse of this structure, that is the metal cation surrounded by four anions and each anion surrounded by eight metal cations, is known as the antifluorite structure and is the arrangement of several of the more valuable precious metal tellurides and selenides, among which is hessite, Ag_2Te, silver telluride, the important ore mineral of silver.

Two structural arrangements that give rise to hexagonal symmetry are the wurtzite, ZnS, and molybdenite, MoS_2, structures. In the wurtzite structure, both metal and sulfide ions are surrounded by four of their opposite charge. The tetrahedral arrangement of ions gives rise to hexagonal symmetry and the second crystalline form of zinc sulfide. The other hexagonal structure is the molybdenite structure and consists of the metal ion coordinated with six sulfur ions and each sulfur ion coordinated with three metal ions. This particular structure is markedly layered and gives rise to the characteristic platelike cleavage in molybdenite.



(A) Fluorite (CaF_2) structure, each calcium surrounded by eight fluoride ions and each fluorine by four calcium ions. (B) Molybdenite (MoS_2) structure, composed of layers of molybdenum and sulfur atoms. The order of layers is -S-Mo-S-S-Mo-S-.

From W.S. Fyfe, *Geochemistry of Solids* (1964); reproduced by permission of McGraw-Hill Book Co., Inc.

One series of five closely related sulfide minerals crystallize in the spinel structure, common as an arrangement for oxide minerals, in which the oxygen or sulfur ions are close packed and the metal ions are in tetrahedral and octahedral coordination. In the series of cobalt and nickel sulfides (linnaeite series), nickel and cobalt occur in nearly continuously varying amounts in three named minerals between the two end members. The vacant sites and latitude for variation of composition in the spinel structure make this possible.

One mineral group—the argentite group—exists in the cesium chloride or body-centred cubic structure at elevated temperatures.

The structures described above are ideal ones; and when ionic charge and size are ideal, minerals that form in them give rise to crystal forms of high, usually isometric, symmetry. Certain combinations of metals and sulfur, however, can take the positions in many of the structures and by distorting or extending the ionic distances create typical crystal forms of lower symmetry. Examples of this are chalcopyrite, which through distortion of the sphalerite structure becomes tetragonal rather than isometric; and cinnabar, which through modifying the sodium chloride structure becomes hexagonal rather than isometric.

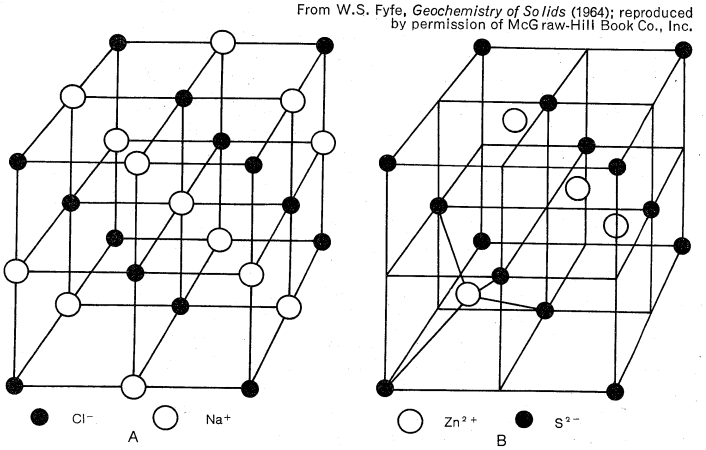
Chemical classification

The common means of classification of the sulfide minerals is that utilized in *Dana's System of Mineralogy*, 7th edition. The classification is based on chemical composition, denoted by the symbols A_mX_n; minerals are then further grouped within a given type on the basis of similar structural-crystallographic characteristics.

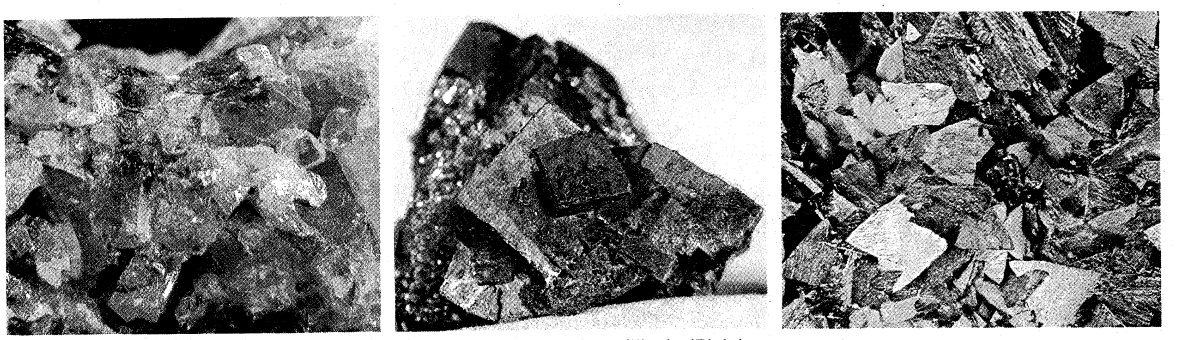
The notation A_mX_n refers to the chemical composition in a general way; A_m refers to the stoichiometric quantity of the metal or metals, and, X_n refers to the corresponding quantity of the cation, usually sulfur. Thus, pyrite, FeS_2, belongs to that group of minerals of the general composition AX_2, and argentite, Ag_2S, belongs to that group of general composition A_2X. This chemical classification usually indicates little of the structural or crystallographic nature of the minerals. Pyrite, FeS_2, for example, has completely different properties from molybdenite, MoS_2, with which it is classified on chemical grounds. Nonetheless, the chemical classification allows a generalized typing of minerals and within the types of the A_mX_n classification, sulfide minerals with similar structural properties can be classified according to groups.

AX sulfide mineral type. Within this formula type, which is further separated into six groups, are included

From W.S. Fyfe, *Geochemistry of Solids* (1964); reproduced by permission of McGraw-Hill Book Co., Inc.



(A) Sodium chloride (NaCl) structure, each ion at centre of regular octahedral group of oppositely charged neighbours. (B) Zinc blende (ZnS) structure, each ion in tetrahedral coordination.



(Left) Realgar from Getchell, Nevada. (Centre) Galena crystals from Galena, Illinois. (Right) Chalcopyrite crystals from Oklahoma.

By courtesy of (left, centre) the Field Museum of Natural History, Chicago, (right) Department of Earth Science, Washington University, St. Louis, Missouri; photographs, John H. Gerard—(left, centre) EB Inc.

many of the most common and widely sought ore minerals of the base metals. The groups of the AX type include the galena group, the sphalerite group, the chalcopyrite group, the wurtzite group, the niccolite group, and the covellite group. Minerals within each of the groups have nearly identical crystal class and the AX minerals have high symmetry.

Galena group. Minerals related to galena, PbS, to the selenides and tellurides of lead (clausthalite and altaite), and the rarer minerals alabandite, MnS, and oldhamite, CaS, make up the group. These minerals have the sodium chloride structure and similar crystal morphologies but otherwise exhibit many different properties. Their lustres range from the bright metallic lustre of galena to the dull and vitreous lustre of alabandite. All except oldhamite occur in crystals of high quality and cleavage is good to perfect. Except for oldhamite, which is known chiefly from meteorites, all occur as ore minerals formed at depth in the earth's crust.

Sphalerite group. Four minerals: sphalerite, ZnS; metacinnabar, HgS; tiemannite, HgSe; and coloradoite, HgTe, belong to this group. All crystallize with the sphalerite structure and belong to the isometric system. Except for sphalerite, all have metallic lustres. Metacinnabar and cinnabar constitute a major source of the metal mercury. Metacinnabar is a black metallic mineral, unlike cinnabar, which is red and tends to be earthy.

Sphalerite, the ore mineral of zinc, is one of the most unusual of all of the sulfide minerals because of its wide variety of colours and lustre. Known commonly by the term zinc blende, it is a common mineral that is commonly misidentified or not identified. Its colour ranges from nearly white, through the hues of yellow, red, green, blue, and brown, to black. The black variety is common and is known to the miner as blackjack, to the mineralogist by the varietal name marmatite. Sphalerite can include within its structure a variety of other metals as trace elements; indium, selenium, and germanium, and significant amounts of cadmium, iron, and manganese are typical. The recovery of wastes in the smelting of zinc ores is a process that yields considerable quantities of these rare elements for use by man. Much of the cadmium, indium, and selenium used in technology has been so derived.

Chalcopyrite group. Chalcopyrite, $CuFeS_2$, and stannite, Cu_2FeSnS_4, are the two principal members of this group of minerals. Both crystallize in structures that represent slight modifications of the sphalerite structure. Stannite is the rarer of the two and is known from tin veins. Chalcopyrite is one of the major ore minerals of copper. Both are metallic and where stannite occurs it is invariably associated with chalcopyrite.

Wurtzite group. Wurtzite, ZnS, is an unstable and thus rare form of sphalerite. It has many of the physical properties of sphalerite, such as resinous lustre and dark colour, and is known from many localities in the world. The second mineral assigned to this group is the rare mineral greenockite, CdS. Rarely crystalline, the mineral commonly occurs as coatings on sphalerite and in the oxidation products that remain after destruction of sphalerite by atmospheric processes. Its most diagnostic characteristic to the naked eye is a bright yellow-orange colour.

Niccolite group. Five minerals are typical of this group; four belong to the hexagonal system and one to the isometric. The most widespread and common of the five minerals is pyrrhotite, $Fe_{1-x}S$, a sulfur-deficient iron sulfide that occurs in geologic environments believed to represent very high temperatures of formation. Niccolite, NiAs, and breithauptite, NiSb, are compounds of nickel; niccolite is the most common.

Pentlandite, $(Fe,Ni)_9S_8$, is a common ore mineral of nickel and is found very closely associated, and finely intergrown with, pyrrhotite. The simple sulfide of nickel, millerite, NiS, is anomalous in its occurrence if it is contrasted with the other nickel minerals of this group. Most frequently it is found as a low-temperature mineral that has filled openings and has replaced calcite. It sometimes occurs as an alteration of other nickel sulfide minerals.

All minerals of this group are of metallic lustre, tend to be somewhat brittle, and have moderate hardness.

Covellite group. Covellite, CuS, and klockmannite, CuSe, are the two type minerals of this group. Both are hexagonal and possess good cleavage. Klockmannite is the rarer of the two. Covellite occurs as an oxidation product of other copper sulfides and is the bright, almost iridescent, blue mineral that usually forms much of the oxidized coating on chalcopyrite, chalcocite, bornite, and enargite.

A number of other minerals belong to this group, among them cinnabar, HgS, the principal ore mineral of mercury and realgar, AsS. Both of these minerals are formed in low-temperature, low-pressure environments such as those near hot springs. Because cinnabar decomposes readily at temperatures in excess of 580° C (1,080° F), and probably at temperatures of 150° C (300° F) in some solutions, it is believed to be closely related to the periphery of the volcanic-hotspring environment. Cinnabar is a red mineral of high specific gravity that most commonly occurs as very small crystals in earthy coatings or vein fillings. Realgar is unstable in the atmosphere and degenerates from a brilliant orange-red, translucent, crystalline aggregate to powdery material.

Other AX minerals of importance are two platinum minerals, cooperite and braggite, PtS, which are known from the high-temperature platinum deposits of Africa. Herzenbergite, SnS, is rare but known from the tin deposits of Bolivia, and empressite and muthmannite, both tellurides of silver, are rare minerals of this group.

A_2X sulfide mineral type. Minerals with this general formula include silver, copper, and gold tellurides, selenides, and sulfides. They have been separated into two crystallographic groups, namely those characterized by argentite, Ag_2S, that are isometric, and those typified by chalcocite, Cu_2S, that are orthorhombic. Most of the minerals of this type assume an isometric symmetry at elevated temperatures (chalcocite is hexagonal) and revert to a lower-symmetry form, usually orthorhombic, when cool. Structurally, most are body-centred cubic or antifluorite in their arrangement. All minerals of this type are metallic in lustre and soft.

Argentite group. Common and important minerals of this group are argentite; digenite, a sulfur-deficient copper sulfide, $Cu_{2-x}S$; hessite, Ag_2Te; and petzite, Ag_3AuTe_2.



(Left) Stibnite from Arnsberg, Westphalia, Germany. (Centre) Pyrite from Butte, Montana. (Right) Argentite from Freiberg, Saxony, Germany.
By courtesy of (left, right) the Field Museum of Natural History, Chicago, (centre) Joseph and Helen Guetterman collection; photographs, John H. Gerard—(left, right) EB Inc.

All are isometric and body-centred cubic. Argentite is one of the most important ore minerals of silver and occurs in low-temperature ore depositional environments. Digenite is a mineral associated with other copper minerals, particularly chalcocite. Difficult to identify, it probably has been overlooked in many early studies of copper minerals and erroneously identified as chalcocite. Hessite and petzite are commonly extracted with other sulfide minerals in many of the old gold and silver mining districts of North America, and are known from many telluride-bearing veins of Europe, Asia, and South America.

Chalcocite group. The minerals chalcocite, Cu_2S; stromeyerite, CuAgS; and acanthite, Ag_2S are all orthorhombic, and acanthite represents the low-temperature phase of argentite. Chalcocite is the copper mineral mined historically from the large copper deposits of western North America. In these deposits it occurred as a secondary mineral, derived from the destruction, by surface processes, of pyrite, chalcopyrite, and bornite. Chalcocite also is believed to have formed by primary processes acting within the earth's crust, but the occurrence of the mineral in its primary form is believed to be much rarer than its occurrence as a secondary mineral. Chalcocite occurs with copper sulfide deposits around the world and, although largely mined out from the "enriched blankets" that it formed, is still common and a mineral of major economic importance. Stromeyerite in small amounts occurs with other minerals of silver and copper in deposits of those metals.

A_3X sulfide mineral type. Two minerals typify this composition. They are dyscrasite, Ag_3Sb, and stibiopalladinite, Pd_3Sb. These are minerals of high specific gravity, of metallic lustre, and white to gray colour. Dyscrasite is an uncommon associate of some silver minerals and base-metal sulfide minerals in veins. Stibiopalladinite is an associate of the platinum deposits of South Africa. The contrast in conditions of formation of the deposits that contain these minerals is significant. The silver veins that contain dyscrasite are presumed to have formed under lower confining pressures and temperatures than the rocks that contain stibiopalladinite. In spite of their close chemical and structural characteristics, they represent somewhat extreme conditions with regard to their environment of formation.

AX_2 sulfide mineral type. A greater number of sulfide minerals belongs to this type than to any other. Groups consist of the structural types pyrite, FeS_2; cobaltite, CoAsS; loellingite, $FeAs_2$; arsenopyrite, FeAsS; molybdenite, MoS_2; and krennerite, $AuTe_2$. Virtually all of the minerals of this type are metallic to submetallic in lustre, and many are hard, although those of the krennerite group are soft. The structures of minerals of this type include the pyrite structure and its modifications and the layered structure of molybdenite and tungstenite.

Pyrite group. Approximately half a dozen minerals are known that belong to the pyrite group; among these, pyrite is the most widespread and significant. These minerals crystallize with the pyrite structure and all are isometric. Other minerals include the sulfides of ruthenium and platinum and the arsenides of iron and platinum.

Pyrite is the most abundant of all sulfide minerals and is the so-called fool's gold. It is present in amounts ranging from slight to overwhelming in virtually every deposit of sulfide minerals. Its bright, brassy lustre, and distinctive habits of crystal form and hardness allow its easy recognition under most circumstances. Its environment of formation ranges from the stagnant and reducing environment of modern and ancient swamps to the very high temperature and pressure environment of deeply derived sulfide ore deposits. Pyrite plays a major part in the processes that weather sulfide mineral deposits because of the ease with which it is decomposed into sulfuric acid and ferric sulfate by atmospheric processes.

Economically, pyrite is used in many parts of the world where the oxide minerals of iron are not abundant and as an ore of iron and in reduction of the mineral; sulfur is recovered and used to produce acid.

Another important mineral in this group is the compound sperrylite, $PtAs_2$. This mineral, originally found near Sudbury, Ontario, is known also from the platinum deposits of South Africa. Its chemical inertness and hardness contribute to its stability in the surface environment, where the mineral is sometimes found in placer deposits with gold and precious stones.

Cobaltite group. Three minerals, cobaltite, CoAsS; gersdorffite, NiAsS; and ullmannite, NiSbS, make up this group; all crystallize with the pyrite structure. In place of the paired sulfur ions in the structure, they contain the nickel or cobalt ion combined with paired As-S or Sb-S occupying the S-S position. All are isometric, all exhibit a metallic lustre, and all are of about equal hardness. Cobaltite is found in association with nickel minerals in primary veins or replacements. Gersdorffite and ullmannite are associates of other nickel minerals in primary deposits of minerals of that metal. Considerable latitude in the structure of these minerals allows for substitution of iron for cobalt and of cobalt and iron for nickel. In addition, bismuth and arsenic substitute for antimony in pairing with sulfur in ullmannite.

Löllingite group. Löllingite, $FeAs_2$; safflorite, (Co,Fe)As_2; and rammelsbergite and pararammelsbergite, $NiAs_2$, are characteristic minerals of this group. All crystallize in the orthorhombic system. Except for rammelsbergite, they occur as well-formed crystals. All have bright metallic lustres and tend to be white rather than gray in colour.

All occur in moderate-temperature, deep-seated, geological environments and all are associates of arsenic minerals. Safflorite is associated with other cobalt minerals, and rammelsbergite with other nickel minerals. Comparatively rare, they nonetheless make up minor amounts in the total mineral assemblage of ores of nickel and cobalt.

Arsenopyrite group. These are minerals of lower symmetry than the types discussed above, occurring as monoclinic or triclinic crystals. Except for arsenopyrite, FeAsS, and possibly glaucodot, (Co,Fe)AsS, they are comparatively rare minerals. Arsenopyrite is the most common arsenic mineral. As with the cobaltite group, there is considerable substitution for sulfur by arsenic and antimony, and rarely by bismuth. A considerable range of compositions exists between glaucodot and arsenopyrite, and cobalt and iron substitute readily within certain limits in the structure.

Arsenopyrite forms in a broad range of geological environments in which the sulfide minerals are formed. The antimonian iron arsenide, gudmuntite, and other members of the group are formed, however, under lower pressures and temperatures than those with which arsenopyrite is ordinarily associated.

Molybdenite group. Two minerals, molybdenite, MoS_2, and tungstenite, WS_2, make up a separate structural assemblage within this compositional type. These two minerals crystallize in the hexagonal system and assume the molybdenite structure. Both minerals are soft and sectile (capable of being cut by a knife with a smooth cut) and occur in scaly aggregates. Because of their structure, they tend to cleave perfectly into platy laminae. For this reason, they feel oily and as a result of this structure, molybdenite is commonly used as a lubricant; it exhibits much the same properties as graphite.

Molybdenite is the chief ore mineral of molybdenum and occurs as a late-stage mineral in processes giving rise to granites and associated end-member rocks. It occurs with quartz, gold, tungsten, and tin in molybdenum deposits and is an important ore mineral in many large low-grade copper deposits. Molybdenite is also a source of rhenium and germanium, both elements that occur as trace metals in the sulfide.

Tungstenite is extremely rare, known from only one locality in the U.S., Salt Lake County, Utah, where it occurs with other base- and precious-metal sulfides.

Krennerite group. Three comparatively rare but important minerals occur in this group, all tellurides of gold and silver: krennerite, $AuTe_2$; calaverite, $AuTe_2$; and sylvanite $(Ag,Au)Te_2$. These minerals exhibit low symmetry, crystallize in the orthorhombic and monoclinic systems, have high specific gravity, and are soft. They are of metallic lustre, white in colour, and are brittle rather than sectile or ductile.

The three minerals have been important historically in making up parts of the precious-metal-ore mineral suite mined extensively from veins in Europe and North America. They also occur in Australia and are found in low-temperature veins with other base-metal sulfides and precious-metal tellurides and sulfosalts.

AX_3 sulfide mineral type. One group of minerals is defined for this compositional type. A series of cobalt and nickel arsenides makes up the group, consisting of the mineral skutterudite, $(Co,Ni)As_3$, at one extreme and nickel-skutterudite, $(Ni,Co)As_3$, at the other. All of isometric crystalline form, they are difficult to distinguish from one another without benefit of chemical analysis. They are found in association with other cobalt and nickel minerals in primary deposits of those metals and are known from a number of localities in Europe and North America.

A_2X_3 sulfide mineral type. Minerals belonging to this type are sulfides of the B subgroup metals, As,Sb, and Bi, and one mineral is a selenide. They are separated into one group, the stibnite group, and two other unclassified minerals of dissimilar crystallographic characteristics. All tend to be quite soft. Although uncommon, most of them are important ore minerals of the metals they contain. They are formed in veins and replacements of low temperature of formation, and appear to be related to the volcanic environment.

Stibnite group. Stibnite, Sb_2S_3, the ore mineral of antimony, bismuthinite, Bi_2S_3, an ore mineral of bismuth, and guanajuatite, Bi_2Se_3 are the members of this group. All are metallic, are some shade of gray in colour, and tend to be sectile. Their crystal habit is prismatic although the habit of guanajuatite is fibrous. Stibnite is found with other low-temperature minerals such as cinnabar and marcasite. Bismuthinite, in contrast, is a product of high-temperature environments and is an associate of arsenopyrite; it is locally found associated with pegmatites.

Orpiment (As_2S_3) does not belong to the group, but, except for crystal character (monoclinic), it has most of the other properties of the stibnite group. Its lustre tends to be resinous and its colour is yellow. It occurs with realgar, from which it alters, and is derived from low-temperature and hot-spring environments.

A_3X_2 sulfide mineral type. Three minerals, including bornite, Cu_5FeS_4, an important ore mineral of copper, are representative of this type. The other two, maucherite, $Ni_{11}As_8$, and umangite, Cu_3Se_2, are rare. Bornite is the bluish, metallic, iridescent mineral of copper deposits known as "Peacock-copper" and can be recognized by the pinkish colour on a freshly broken surface that tarnishes rapidly to the bluish iridescence. Its environment of formation is broad from the standpoint of physical conditions, and it is known from low-temperature deposits as well as those formed at very high temperature and pressure. Almost as abundant as chalcopyrite, it occurs in copper and other base-metal deposits the world over.

AX_4 sulfide mineral type. This type, as with the AX_3 type, is composed of a series of cobalt-nickel sulfides in which there is more or less continuous composition between the end members, linnaeite, Co_3S_4, and polydymite, Ni_3S_4. Five minerals make up the end and intermediate members of the series. All have brilliant metallic lustres on fresh surface and all tarnish to some shade of red or violet. They crystallize in the isometric system in the spinel structure and are octahedral in form. The minerals occur with other sulfide minerals and with minerals of cobalt, nickel, and copper. They are rare.

A_4X_3 sulfide mineral type. Three uncommon minerals are assigned to this type. They are dimorphite, As_4S_3; rickardite, Cu_4Te_3; and weissite, Cu_5Te_3. Dimorphite is known only from fumaroles. The other two minerals are rare associates of other telluride minerals in such deposits.

A_mX_n m:n > 3:1 sulfide mineral type. Two groups of minerals make up this type. One, the tetradymite group, is a series of bismuthian telluro-sulfides, tellurides, and sulfides. The other is a group of copper arsenides.

Tetradymite group. Five minerals, all with similar properties, make up this group. They are tellurobismuthite, Bi_2Te_3; tetradymite, Bi_2Te_2S; grünlingite, Bi_4TeS_3; joseite, $Bi_3TeS(?)$; and wehrlite, $Bi_3Te_2(?)$. All are soft, laminar, metallic minerals with white or bright colour. Of the five, tetradymite is the most common, occurring worldwide with other tellurides, base-metal sulfides, and with gold in gold-quartz veins.

Copper arsenide group. Two distinct minerals and a third of doubtful validity belong to this group. Algodonite, Cu_6As, and domeykite, Cu_3As, are defined species, with algodonite crystallizing in the hexagonal system and domeykite forming in the isometric system. The doubtful species is whitneyite, an ill-defined copper arsenide that may belong to the series. Both algodonite and domeykite are gray in colour, bright on fresh surface, and dull when tarnished. Both minerals are known from copper deposits in North and South America and they are associates of the copper minerals in the deposits of the Upper Peninsula of Michigan, in the United States.

OCCURRENCE OF SULFIDE MINERALS IN NATURE

Sulfide minerals occur in all rock types. They may be syngenetic, that is formed at the same time and under the same conditions as those that formed the rock that encloses them. In other occurrences, they may be epigenetic, formed later than the rock that encloses them. Examples of syngenetic sulfides include pyrite in sedimentary rocks, such as shales or sandstones (*qq.v.*), chalcopyrite disseminated in felsic igneous rocks, and certain of the magmatic segregation deposits in which sulfide segregations occur conformably with rock structures.

Examples of epigenetic deposits of sulfide minerals are the common replacements of wallrock of veins by base- and precious-metal sulfides, the replacement of selected sedimentary beds by thick blankets of massive base-metal sulfides, and the deposition in open spaces of rocks of many sulfide mineral types.

Characteristic rock associations

On a worldwide basis, there is a rather consistent, although inadequately explained, association of certain metals and their primary minerals with certain rock types. Vein deposits of gold are commonly associated with diorites of igneous origin; nickel is commonly associated with norite, an ultramafic rock; lead and zinc extensively occur with sedimentary rocks, usually carbonates, and are related to igneous activity; silver commonly is associated as vein and replacement deposits with andesites of volcanic origin; and copper occurs with igneous rocks of diorite to quartz monzonite composition. No consistent rule, however, can be established to predict sulfide mineral type on the basis of associated rock type. Although a mineral deposit may be valuable because of one suite of sulfide

minerals, it also may contain an abundance of other minerals, many of them sulfide minerals, not valuable in the particular setting.

Sulfide minerals in ore deposits. Ore deposits (*q.v.*) range from simple mineralogy and principal value in one or two minerals of the same metal sought to large, district-wide distribution of numerous minerals that may be selectively extracted and processed for different elements. In many mining districts, sulfide and sulfosalt minerals are distinctly zoned. A classic example of such zoning is at Butte, Montana. An inner core of quartz-molybdenite is surrounded by a zone of copper sulfides. A zone of mixed copper and zinc ores surrounds the copper zone, and this mixed zone gives way at its outer edge to a peripheral zone in which are found ore minerals of zinc, manganese, and silver.

The common ore-minerals of most elements are sufficiently consistent worldwide that it is possible to list the significant minerals of most deposits of certain elements. See also NATIVE ELEMENTS.

Copper deposits. Principal sulfide minerals of copper deposits are chalcopyrite, bornite, and chalcocite. Lesser amounts of covellite and cuprian sulfosalt minerals occur, as do stromeyerite and some tellurides and selenides. Pyrite is a common associate and minor amounts of galena and sphalerite occur. Present in small amounts but nonetheless economically significant in many deposits of the Western Hemisphere is molybdenite.

Lead deposits. The chief ore mineral of lead is galena. Its common associates are sphalerite and pyrite. In addition to these minerals, lead deposits commonly are sources of silver, which may occur as a trace element in galena or in very fine-grained phases of silver tellurides, arsenides, or sulfides.

Zinc deposits. Sphalerite is the major ore mineral of zinc, and its common associates are galena and pyrite. Also extracted from these deposits is cadmium from the mineral greenockite.

Nickel-iron deposits. Pentlandite is the major ore mineral of nickel deposits. Pyrite is sometimes present, as are other nickel arsenides such as niccolite and maucherite.

Silver and gold deposits. Although native gold is the major mineral of gold deposits, and sulfosalts minerals are a major source of silver, there are some sulfide minerals important as ores. These are tellurides and selenides, together with the sulfide argentite. Among the ore minerals are hessite, petzite, empressite and muthmannite.

ORIGIN OF SULFIDE MINERALS

The formation of sulfide minerals, like all minerals, is a function of their environment. Because sulfides are deposited both in the subcrustal environment by processes acting within the earth and at the surface of the earth, it is necessary to consider the characteristics of both environments. In the earth's crust, temperatures are high and approach the melting point of silicate minerals. Pressures are likewise high, as a result of the confining pressure of overlying rocks. In contrast to the surface of the earth, water and oxygen concentrations in the crust are low. At the earth's surface, in contrast, temperatures and pressures are low, and oxygen, water, and organic material are plentiful. In each of these environments, sulfur and the metals with which it combines behave in manners that are predictable in a broad way, and it is clear that the environment exerts a profound influence upon the nature of the minerals deposited and their associations.

Subcrustal formation

Most sulfide minerals found at depth within the earth are the result of processes that give rise to modification of pre-existing rocks (metamorphism) or melting and recrystallization or rock formation by crystallization of a silicate melt. Sulfide minerals occur with the metamorphic and igneous rocks and result from the processes that form the rocks. Sulfur and metals can occur in the rocks that are modified; or, it is believed, they occur in a highly dispersed state as primordial material in rocks that lie beneath the crustal layer.

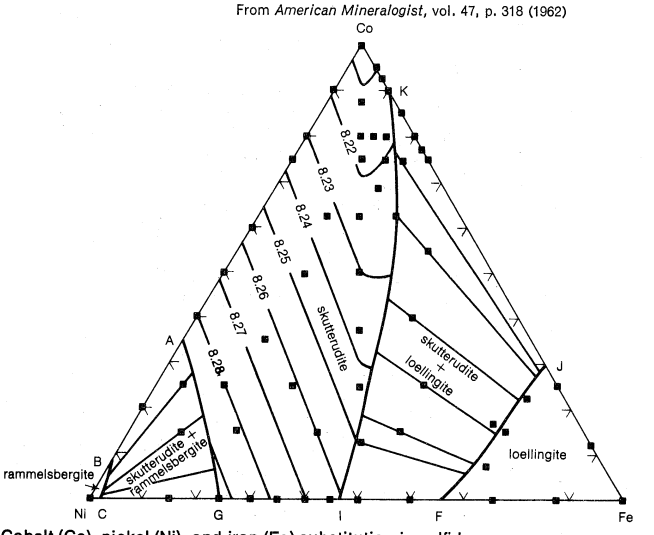
Material that gives rise to the igneous rocks in which sulfide minerals are found or with which they are closely associated is considered by many to have been molten at one time and to have been a silicate melt. During the cooling of this melt, sulfur and metals are progressively enriched in the uncrystallized portion because they cannot be incorporated in the rock-forming silicate minerals (*q.v.*) being formed. Along with water and gases such as carbon dioxide, sulfur and metals form a separate system that is immiscible with the silicate melt.

This system may be concentrated or dispersed within a cooling silicate mass. If highly dispersed, then sulfides eventually form as dispersed grains in the igneous rock. If concentrated at one or several locations above the silicate melt, the gas-sulfur-metal system may escape as the result of a geological accident such as faulting or fracturing or by encountering different rocks with increased permeability. The fluid moves down the pressure gradient, usually upward, where it loses heat to the surrounding rocks, changes composition by reaction with them, and deposits the sulfide minerals as vein deposits or replacements in the system.

Metals, derived in the destruction of other rocks, and sulfur, the product of the biogenic cycle, are incorporated in sedimentary rocks (*q.v.*). As these rocks become buried and are subjected to higher and higher temperatures, recrystallization of the sedimentary minerals takes place, and these minerals become the rock-forming silicate minerals of metamorphic rocks (*q.v.*). Similarly, metal and sulfur are excluded from these minerals and are subject to movement as an immiscible system down existing pressure gradients. Thus, metamorphism or diagenesis may act in the same way that melting does to produce a sulfide-water system.

A limiting condition upon the temperature of formation of any mineral is that of its melting temperature under the conditions in which it forms. Another limiting condition is the solubility of the sulfide in the gas or transporting phase that deposits it. Even if it is below its melting temperature, a sulfide mineral will not form if it is soluble in the system from which it would crystallize. A broad range of melting temperatures exists for the sulfide minerals, and in a general way this range is consistent with the inferred environment of their formation. Those minerals of high melting temperature are usually from high-temperature environments, although they also may be found in lower-temperature ones. An example is pyrite, which, with a melting temperature of 743° C (1,370° F), is believed to have formed at room temperatures in some environments. On the other hand, cinnabar, which has a melting temperature of 580° C (1,080° F) (approximate sublimation point of metacinnabar), is soluble in some

From *American Mineralogist*, vol. 47, p. 318 (1962)



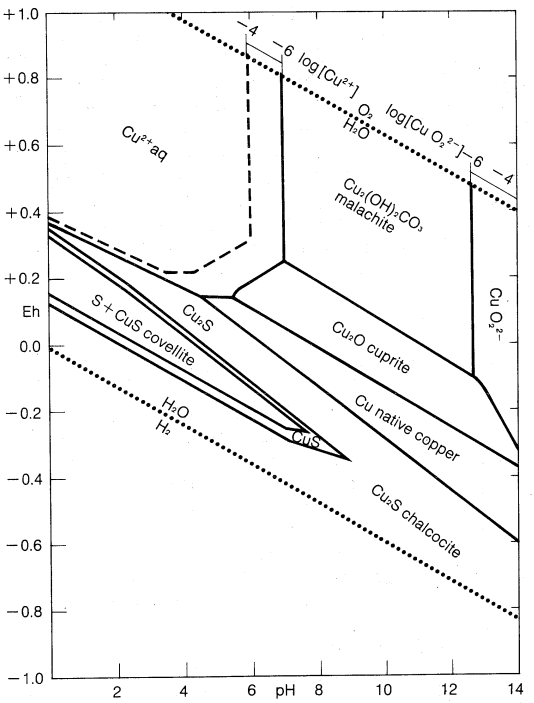
Cobalt (Co), nickel (Ni), and iron (Fe) substitution in sulfide minerals at 800° C and 25 atmospheres water vapour pressure. Tie lines connect pairs of coexisting mineral phases under these conditions in the stability fields of skutterudite plus loellingite (KIFJ), skutterudite (KIGA), and skutterudite plus rammelsbergite (ABCG). Compositions are in mole-percent Co–Ni–Fe, and squares indicate the total composition of samples (see text).

fluids at temperatures of 150° C (300° F); it is for this reason that it is believed to represent low temperatures of formation.

Mineral stability fields. Experimental work after 1920 resulted in a considerable body of information related to high-temperature equilibria in sulfide systems. The data are numerous and the system Co-Ni-Fe-As is representative. The diagram illustrates the variation in composition allowable within defined species of some systems and the corresponding variation in physical properties as substitution of other elements progresses. The diagram shows sulfide relationships at 800° C (1,500° F).

The behaviour of sulfides at low temperatures and pressures is defined by a different set of parameters, because the temperatures and pressures may be held fairly constant. Instead of temperature and pressure, the stability fields of minerals can be shown in terms of their relationship to concentration, oxidation potential (Eh), or oxidizing-reducing conditions, and to pH, or acidity-alkalinity. The diagram of the sulfur species shows the relationships of various copper minerals to their oxides, which form at higher conditions of Eh. The upper boundary of the native copper field, for example, shows the conditions under which cuprite will form in nature.

From R. Garrels and C. Christ, *Solutions, Minerals, and Equilibria* (1965); Harper & Row



Stability relations among some copper compounds in the system Cu-H_2O-O_2-S-CO_2 at 25° C and one atmosphere total pressure.

Paragenesis of sulfide minerals. The microscope allows close examination of sulfide mineral suites, and from this examination a sequence of formation may be determined. This sequence, termed paragenesis, allows further determination of the trends or directions that an ore-forming process takes. The consistency of these relationships of mineral formation and time is noteworthy.

SOME TYPE LOCALITIES OF SULFIDE MINERALS

Chalcopyrite. Kitzbühel, Tyrol, Austria; Freiberg, in Saxony; Rammelsberg, in the Harz Mountains; Río Tinto, in Spain; Braden, in Chile; in Cornwall, at Camborne and Carn Brea; at Ducktown, Tennessee; at Jerome, Arizona; and at the Noranda Mine, Quebec.

Sphalerite. Příbram, Czechoslovakia; Kapnik, Romania; in the Binnatal, Switzerland; Cumberland, at Alston Moor; at Ämmeberg, Sweden; Cananea, Mexico; Mississippi Valley, U.S.; and Joplin, Missouri.

Galena. At Tarnowitz, in Upper Silesia; Chemnitz (now Karl Marx Stadt) in East Germany; at Andreasberg in the Harz Mountains; in Cumberland at Weardale; in the Tri-State District, U.S.; from the Alta Mine, Wicks, Utah; and Inyo County, California, U.S.

Chalcocite. St. Just, St. Ives, and Camborne, in Cornwall; in Joachimsthal in East Germany; Bristol, Connecticut, U.S.; Butte, Montana, U.S.; and at the Kennecott Mine, Alaska.

Pyrrhotite. Sudbury, Ontario; in Transylvania, at Kisbanya; at Bottina, in Tuscany; at Falun, in Kopparberg, in Sweden; at Andreasberg, in the Harz Mountains; in Brazil at Minas Gerais; and from Ducktown, Tennessee.

Stibnite. in the Harz Mountains; in England, at Calston; Sarawak, Borneo; Puno, Peru; in the Coeur d'Alene district, Idaho, and Kern County, California.

Bismuthinite. Bolivia in the tin veins; Cumberland, England; Guanajuato, Mexico; at Schneeberg in Saxony; and at various localities in Colorado.

Pyrite. In the Tri-State District of U.S.; Bingham Canyon, U.S.; at Cornwall; in many locations in Colorado; at New Almaden, California; Elizabethtown, Ontario; and numerous localities throughout the world.

BIBLIOGRAPHY. J.D. DANA, *The System of Mineralogy*, 7th ed. entirely rewritten by C. PALACHE, H. BERMAN, and C. FRONDEL, vol. 1(1944), the best modern review of mineralogy, with extensive detail on the sulfide minerals; LINUS CARL PAULING, *Nature of the Chemical Bond and the Structure of Molecules and Crystals*, 3rd ed. (1960), a treatise on chemical bonding and structural chemistry of minerals; H.L. BARNES (ed.), *Geochemistry of Hydrothermal Ore Deposits* (1967), modern data on the chemistry of sulfide mineral formation.

(S.R.Ty.)

Sulfur Products and Production

Sulfur, a yellow, nonmetallic element, occurs in sedimentary and volcanic deposits and exists in several different forms, the most stable being orthorhombic sulfur, a crystalline form that melts at about 116° C (about 241° F). (The properties of sulfur are described in the article NATIVE ELEMENTS.) Liquid sulfur can be ignited and burns with a characteristic deep blue flame. Sulfur was discovered in prehistoric times and has been used ever since for making sulfur dioxide, sulfuric acid, carbon disulfide, and other compounds. *Sulfur* is Latin for "burning stone," which explains the archaic English term brimstone.

Sulfur has been widely used for religious ceremonies, to purify containers, and for bleaching cloth. The Romans also used it in warfare and medicine. In the Middle Ages, alchemists gave sulfur an important place. Many chemists long doubted the elemental nature of sulfur, but in 1777 Antoine Lavoisier classified it definitively as an element, and experimental confirmation was developed in 1809.

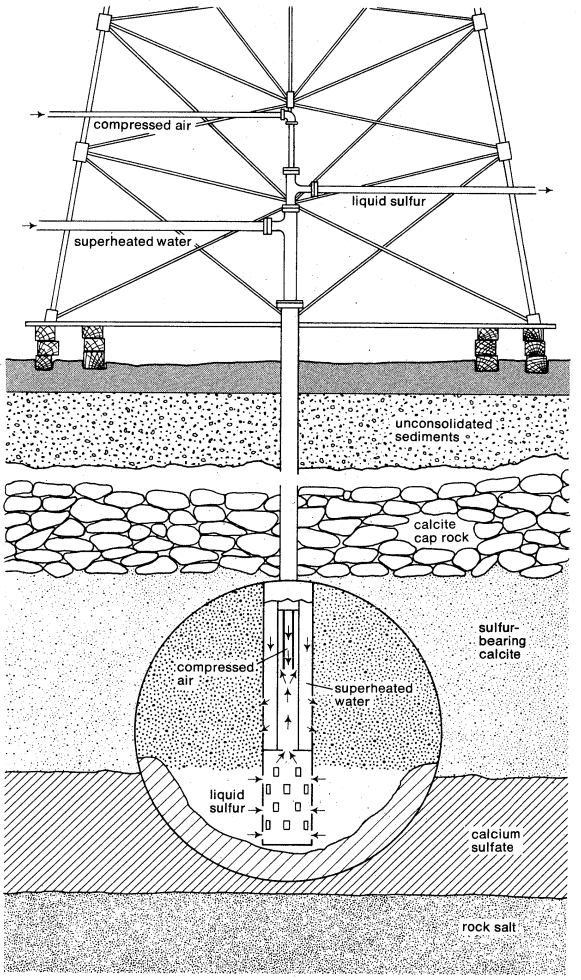
Occurrence. Sulfur, sulfides, and sulfates constitute about 0.06 percent of the Earth's crust. Most of the world's sulfur production is from elemental sulfur that forms the calcareous horizon of the cap rock of salt domes (*q.v.*) around the Gulf of Mexico and Tehuantepec, Mexico. In Sicily sulfur is produced from sedimentary deposits. Volcanic deposits are exploited in Japan, Indonesia, and South America, especially Chile. An increasingly important source of sulfur is natural gas, which in the western Canadian prairies, southwestern United States, and southern France contains up to 40 percent hydrogen sulfide. Vast amounts of sulfur compounds are contained in the Athabasca tar sands of Canada. In Scandinavia and some other countries, sulfur is produced from naturally occurring compounds of sulfur and iron such as pyrite, pyrrhotite, and chalcopyrite. Other important sulfur-containing minerals are galena (lead sulfide) and sphalerite (zinc sulfide). Gypsum, its anhydrite (calcium sulfate), and barite (barium sulfate) are the best known sulfates. Sulfur occurs also in organic substances.

Mining and recovery. In Sicily, where until the beginning of the 20th century most sulfur was produced, the element is refined by methods developed in antiquity. Native ore is piled in cylindrical kilns (calcaroni) and ignited. About one-third of the material burns, and the heat from the combustion melts the remainder, which is

cast in blocks. In a somewhat improved smelter, called a Gill furnace, hot gases are recirculated to preheat the ore. In Japan the prevailing recovery method is distillation from retorts.

The Frasch process

Since 1900, as the result of the so-called Frasch mining method developed by Herman Frasch, the United States has become the largest sulfur producer. This process, applied to salt domes, has made sulfur of a high purity available in large quantity and has helped establish sulfur as one of the four most important basic chemical commodities. Wells are drilled from 200 to 2,000 feet (60 to 600 metres) into the sulfur formation (see illustration).



Sulfur well.
Superheated water causes the sulfur to melt from the deposit and flow to the bottom. From there it is forced to the surface by air pressure.

The well is lined with a six-inch pipe in which a one-inch air pipe and a three-inch water pipe are concentrically placed. Superheated water, injected into the circular space between three- and six-inch pipes, penetrates the cap rock through holes on the bottom of the pipe. As the sulfur melts, it settles to the bottom of the deposit. From there it is pumped to the surface by applying air pressure through the central well pipe. Several wells operate under the ocean floor in the Gulf of Mexico. From these offshore platforms sulfur is pumped several miles through insulated pipes in liquid form along the ocean floor. Once on the surface, the sulfur is collected in reservoirs, or sumps, and from there transferred to vats or bins to solidify for storage and stockpiling. Vats may contain as much as 300,000 tons of sulfur.

In addition to sulfur obtained in elemental form, about 4,000,000 tons are recovered in the United States each year from natural gas, petroleum refinery gases, pyrites, and from smelter gases from the processing of copper, zinc, and lead ores. In most cases, sulfur is separated from other gases as hydrogen sulfide and then converted to elemental sulfur by the so-called Claus process, which involves partial burning of hydrogen sulfide to sulfur dioxide, with subsequent reaction between the two to yield sulfur.

Another important source of sulfur may lie in the large volume of sulfur dioxide emitted into the atmosphere by coal-fired steam power plants and other large coal-consuming facilities. In the early 1970s, techniques to collect this sulfur dioxide and convert it into usable sulfur were developed. If economically or politically feasible, such techniques could double sulfur production in the United States and provide numerous small sulfur sources throughout the world.

World production

World production and consumption of sulfur have increased steadily, except during periods of long recessions, since the inception of the Frasch process. In the early 1970s, about 12,300,000 tons of elemental sulfur were being produced annually in the United States and Mexico alone; France was producing about 1,900,000 tons and Canada about 4,900,000 tons as a by-product of natural gas production. Canadian production is rising rapidly; in most countries production is relatively steady. Poland, Scandinavia, the Soviet Union, Spain, Japan, Chile, and China produced smaller but significant quantities.

Refining and purification. Sulfur quality and chemical purity are reflected by its colour. Bright sulfur is brilliant yellow and 99.9 percent pure, or better. Dark sulfur contains 0.5 to 5 percent organic matter. The oldest production method, melting by heating ores, produces dull yellow or even dark sulfur that contains organic impurities. These are difficult to remove because they react slowly with sulfur. The Frasch process yields up to 99.9 percent sulfur. Clean Frasch sulfur is bright sulfur, as is by-product sulfur recovered through the Claus process.

Bulk quantities of the element are now shipped almost entirely as liquid. Japanese and Sicilian sulfur is sold in ingots, and some broken-rock brimstone is shipped in lumps.

Special forms of sulfur used in the rubber industry

For the rubber industry and other special uses, sulfur is transformed into the following special forms:

Flowers of sulfur is sulfur refined by sublimation—that is, condensation from the vapour phase without passage through the liquid phase. It contains 25 to 30 percent amorphous sulfur, which is insoluble in carbon disulfide. Its main uses are in rubber vulcanizing, the preparation of insecticides, fungicides, and pharmaceuticals.

Precipitated sulfur is precipitated from a sulfur compound such as thiosulfate or polysulfide solution and used primarily for pharmaceutical purposes.

Roll sulfur, refined sulfur cast into convenient shapes and sizes, is used for fumigation, burning, and bleaching. It is sometimes referred to as stick, or cannon, sulfur.

Commercial flour sulfur is prepared by milling (grinding) refined sulfur to a very fine particle size. It is used for vulcanizing rubber and in the manufacture of gunpowder and dyes.

Colloidal sulfur has such a small particle size that it remains suspended in water. It is prepared from thiosulfate and used in the manufacture of insecticides and pharmaceuticals.

Uses. Sulfur is one of the four basic raw materials of the chemical industry (the other three being coal, limestone, and salt). Eighty-five percent of all sulfur used is first converted into sulfuric acid, and nearly 80 percent of this is used in fertilizer production. Sulfur dioxide is used in leaching ores and in bleaching cloth and food products, and it is an important intermediate chemical in the chemical industry. Other important uses include inorganic pigments, cellulosic fibres, nonferrous metallurgy, pulp and paper, and petroleum refining. Many other products, such as cements, dyes, food preservatives, photographic materials, sugar, and textiles, depend on sulfur in various stages of manufacture.

Hydrogen sulfide can be formed by the reaction of hydrogen with liquid sulfur. It is a colourless gas, poisonous when inhaled, and has the odour of rotten eggs. Often used as a reducing agent, it is oxidized by chlorine, bromine, and iodine to sulfur—the excess halogen oxidizing

the sulfur to sulfuric acid—and by concentrated sulfuric acid to sulfur and sulfur dioxide. It can be converted to elemental sulfur with a good yield by combining with oxygen, a process increasingly used to produce sulfur from natural gas and oil. Some metallic sulfides, containing small amounts of heavy metal salts, are phosphorescent. Zinc sulfide is used in luminous paints and, combined with traces of a radium salt, has been used for hands and figures on luminous watch dials.

Carbon disulfide, formed by direct union of carbon and sulfur at high temperatures, is important commercially, primarily in the preparation of viscose rayon. Nitrogen tetrasulfide, formed by passing ammonia through carbon disulfide dissolved in benzene, is used as an insecticide and fungicide and may have future value as an ignition promoter in diesel fuels. Other sulfides have also been prepared, but their uses are minor.

The sodium salt of hyposulfite is used commercially and serves as a reducing agent in the dye industry. Sulfurous acid and its salts are excellent and cheap reducing agents and are also used as disinfectants and food preservatives. Calcium bisulfite and magnesium bisulfite are used to extract lignin from wood in preparing wood pulp. Thiosulfates are used extensively in photography.

Sulfur monochloride and sulfur dichloride are used to vulcanize rubber. The oxychlorides are useful reagents for the introduction of sulfoxides, sulfone, or chlorine groups into organic molecules. Chlorosulfonic acid, prepared by reacting hydrochloric acid directly with sulfur trioxide, reacts readily with water to form sulfuric and hydrochloric acids. It fumes in air and has found application as a smoke producer in military operations. Its most useful application is as a sulfonating agent in organic chemistry.

The aromatic sulfur compounds, usually nitrogen derivatives, find extensive use as accelerators in the vulcanization of rubber. The second type of reaction, oxidation of organic sulfur compounds, yields materials that, with the exception of the disulfides and the peroxides, are not analogous to oxygen compounds in structure. The most important of these compounds are the sulfuric acids, alkyl sulfates, and their derivatives, which are used extensively as medicinals (sulfa drugs), detergents, lubricants, and dyestuff intermediates.

Sulfur is also used for a variety of other purposes. Combined with sand or gravel, it yields acid-resistant, quick-setting cements that can be used for house construction, sewage pipes, and road surface repairs. Large-scale tests with sulfur foam indicate that it might become commercially important as an insulating material in the construction of houses and road beds. It is used as an insecticide and fungicide for plants and animals. Sulfur ointments are used in treating skin rashes and acne, and sulfur pills are still widely used medicinally as a tonic. A major use is as plant food, especially in alkaline soil.

Organic sulfur compounds are important plasticizers and are incorporated into polymers.

BIBLIOGRAPHY. W.N. TULLER, *The Sulphur Data Book* (1954), is a classical review. The FREEPORT SULPHUR COMPANY, *Freeport Sulphur Handbook* (1959); and TEXAS GULF SULPHUR COMPANY, *Properties of Sulphur* (1961), contain much useful information that is nowhere else available. B. MEYER (ed.), *Elemental Sulfur: Chemistry and Physics* (1965), contains a review of sulfur qualities produced and a chapter on products. A more recent account of this subject may be found in the *Kirk-Othmer Encyclopedia of Chemical Technology*, 2nd ed., vol. 19, pp. 337–364 (1969).

(C.B.M.)

Sulla, Lucius Cornelius

Lucius Cornelius Sulla, Roman general and dictator, carried out great constitutional reforms aimed at reinforcing the republican character of the Roman government. His reforms did not, however, prevent the extinction of the republic only a few decades after Sulla's death. He was born in 138 BC, the son of a politically unimportant patrician family. He early showed a taste for luxury and aspired to a political career, which he began in 107 BC, under the command of Gaius Marius, as a quaestor (financial magistrate) in Africa in the war against King Jugurtha of the Numidians. His spectacular capture of Jugurtha by trickery marked the start of his feud with Marius. Although Marius continued to use Sulla in the war against the invading Cimbri, in 103 BC his jealousy became obvious, and Sulla transferred to the service of Marius' co-commander.

Early career



Sulla, marble bust. In the Vatican Museum.
Alinari

After service as a Roman praetor (one of the chief magistrates) in 94 BC, Sulla fought in the Social War (90–89 BC), the struggle of Rome's Italian allies to obtain Roman citizenship. He became one of the two consuls—the highest office in the republic—in 88 and was placed in command of the war against King Mithradates VI of Pontus in Asia Minor. By his marriage—his fourth—to Caecilia Metella, in 88 BC, he formed important alliances. But Marius, through his alliance with the tribune Publius Sulpicius Rufus of the popular party, succeeded in being appointed commander instead. Sulla marched on Rome, Sulpicius was killed, and Marius escaped.

In 88 Sulla set off for Greece in charge of the war against Mithradates. By the spring of 87 most of Greece was in his power, and after a long siege he captured Athens in 86. Mithradates' general was pursued into Boeotia and finally defeated in two battles in 86.

At a meeting in 85 between Sulla and Mithradates at Dardanus on the Hellespont, the latter accepted a punitive treaty. Order was restored in Asia and Greece, and Mithradates became a vassal of the Romans again. In the summer of 83, Sulla, after a lengthy stay in Athens, returned to Brundisium in southern Italy with 40,000 men and enormous plunder.

During his absence, Sulla had been declared a public enemy by the ruling popular party. His laws were repealed, his house was destroyed, and his family and friends fled to join him in Greece. In 86 the former consul Lucius Valerius Flaccus was sent to replace Sulla in the Asian command. But Sulla's luck did not desert him; Flaccus was murdered by his lieutenant.

From Brundisium, Sulla began his march on Rome, joined by opponents of the popular regime. Through most of the ensuing civil war Sulla was opposed by the consuls Gnaeus Papirius Carbo and the younger Marius (whose father had died in 86). Sulla's victory of Colline Gate in the northern environs of Rome and the fall of Praeneste at the end of 82 ended the war, which was followed by massacres and proscriptions.

Sulla was appointed dictator under the Lex Valeria (Valerian law), which vested constituent, legislative, military, and judicial power in him, without, however, for the first time in Rome's history, limiting the duration of his dictatorship.

The state was reorganized and the new legislation enacted in 81, at the start of which year (January 27–28) Sulla celebrated his victory over Mithradates. In the speech delivered at the close of the ceremony, he chose for himself the name of Felix (Epaphroditos in Greek documents).

By his extensive program of constitutional reform he intended mainly to re-establish the supremacy of the Senate in the Roman state, and many of his reforms did indeed survive to the end of the republic. Of value were the increase of the number of courts for criminal trials; a new treason law, Lex Cornelia de Majestates, designed to prevent insurrection by provincial governors and army commanders; the requirement that the tribunes had to submit their legislative proposals to the Senate for approval; and various laws protecting citizens against excesses of judicial and executive organs. At the beginning of 79, Sulla resigned and withdrew to the neighbourhood of Puteoli (modern Pozzuoli) in Campania. This action caused a sensation in Rome; many different explanations have been given, starting with the classical writers. Most commonly accepted is the view that Sulla's resignation was an act of honesty by a man who had pledged to step down as soon as his reforms had been carried out. Henceforth a private citizen, he continued to write his memoirs. Active to his very last days, Sulla was struck down by a fever in the spring of 78. He left behind two children by Metella and a posthumous daughter by his fifth wife, Valeria.

Assessment

Sulla, a soldier and a politician, a dictator and a reformer, a man of contradictions in an age of contradictions, is the subject of contradictory opinions expressed by both classical and modern writers. The classical writers of Sulla's time or shortly thereafter found it difficult to form an opinion of him; they noted the discrepancy of the Sulla "who follows up good beginnings with evil deeds." Generally, their attitude was a negative one, with references to despotism, slavery, cruelty, and inhumanity, and the absence of any principle of good government. The opinions held by modern writers cover a spectrum ranging from Sulla the enigma (because of his resignation), to Sulla the monarch, to Sulla the honest reformer.

Sulla was the exponent of a decadent patriciate that tried everything in its power to save itself by instituting reforms that, while not without democratic aspects, lacked inner vitality. From the long-term perspective, Sulla's actions seem meaningless; but viewed in their historical context they are justified by the transitional character—both in its military and political aspect—of his age. Inspired by a glorious past, interpreting an extremely volatile present, and heralding a future faithful to tradition, Sulla played a historical role, conclusively shaping and epitomizing the republican ideal shortly before it became submerged. But he was mistaken about the significance of his reforms: he was a temporary dictator because he wanted no one else who might follow him to become a dictator for life; yet by his example he unwittingly paved the way for Julius Caesar.

BIBLIOGRAPHY. F. FROLICH in the *Pauly-Wissowa Real-encyclopädie*, vol. 7, col. 1522–1566 (1900), a useful, detailed, biographical article that explains Sulla's astonishing abdication; HUGH LAST and R. GARDNER, *The Cambridge Ancient History*, vol. 9, ch. 6 (1932), a positive evaluation of Sulla's reforms and of his organizational capacities, best demonstrated in his regulations for Asia and for "public justice"; CAROLINA LANZANI, *Lucio Cornelio Silla dittatore* (1936), an ample and accurate exposition of the events pertaining to the years 82–78; JEROME CARCOPINO, *Sylla, ou la monarchie manquée*, new ed. (1947), an explanation of Sulla's abdication as a result of his unsuccessful attempt to prevail as a monarch—this thesis has not had much of a following; ERNESTO VALGIGLIO, *Silla e la crisi repubblicana* (1956), an examination of the decade 88–79 of Cornelius Sulla, who is seen as an aristocrat faithful to the republican institutions, to which he tried to give a stability they could no longer have; U. LAFFI, "Il mito di Silla," *Athenaeum*, 45:177–213, 255–277 (1967), a study of the fortune of the Sullan regulation of the state, including evaluations of his work with respect to history and historiography.

(E.Va.)

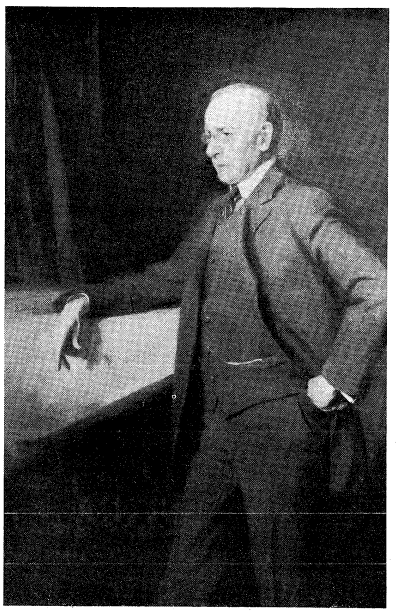
Sullivan, Louis

Louis Sullivan is regarded as the spiritual father of modern U.S. architecture and is particularly identified with the aesthetics of skyscraper design. He was a spokesman for the reform of architecture, an opponent of historical eclecticism, and did much to remake the image of the architect as a creative personality. His own designs are characterized by richness of ornament.

Early years in Boston

Louis Henry Sullivan was born in Boston, Massachusetts, on September 3, 1856, of Patrick, a dancing master, and Adrienne Françoise (List) Sullivan. His Irish-born father and Swiss-born mother had emigrated to the United States in 1847 and 1850, respectively, and were married in 1852. Their older son, Albert Walter, was born in 1854.

By courtesy of the Chicago Historical Society



Louis Sullivan, oil painting by Frank A. Werner, 1919. In the collection of the Chicago Historical Society.

Sullivan attended public schools in the Boston area and spent summers on his grandparents' farm in nearby South Reading. When his parents moved to Chicago in 1869, Sullivan stayed behind with his grandparents and later with neighbours, commuting to school in Boston.

In September 1872 he entered Massachusetts Institute of Technology, which had the first architectural school in the United States (founded 1865). Sullivan was an impatient architectural student and left at the end of the year with thoughts of studying at the École des Beaux-Arts in Paris or of being an apprentice to an architect. He discussed his ideas in New York with Richard Morris Hunt, one of the fashionable architects of the day and the first American to study architecture at the Beaux-Arts. Hunt suggested he work with the Philadelphia firm of Furness and Hewitt. Sullivan was hired, staying for several months until work dwindled in the economic panic of 1873. In November he left for Chicago and was soon employed in the architectural office of a prominent figure in the development of the style of the Chicago school, William Le Baron Jenney. The office foreman, John Edelmann, became his future friend.

The idea of studying in Paris persisted, however, and in July 1874 he sailed for Europe. Sullivan worked hard to pass the difficult entrance examinations for the Beaux-Arts, although after he was accepted he proved to be a restless and erratic student. He made a brief excursion to Florence and Rome. A romantic young man with sideburns, he affected a certain swagger in dress. During the single year he remained in Paris, he was attached to the atelier of the architect Émile Vaudremer.

Work in association with Adler. Back in Chicago in June 1875, Sullivan worked briefly as a draftsman for a

number of firms. One such job was for the recently formed firm of Johnston and Edelmann. It was John Edelmann who made the momentous introduction of Sullivan to his future partner Dankmar Adler. In 1879 Sullivan joined Adler's office and in May 1881, at the age of 24, became a partner in the firm of Adler and Sullivan, Architects. Their 14-year association produced more than 100 buildings, many of them landmarks in the history of U.S. architecture.

Collaboration with Adler

Sullivan's brilliance as a designer was complemented by Adler's business ability, his tact with clients, and his knowledge of technical matters, especially acoustics. After coming to Chicago in 1861, Adler had worked as a draftsman, and he returned to the city after serving in the Civil War. In 1871 he formed a successful partnership with Edward Burling that lasted until 1879. As an independent architect Adler designed Central Music Hall in Chicago (1879), which was the prototype of theatres later designed by the firm of Adler and Sullivan. Adler was a consultant on acoustics and in his later years was a writer on the technical and legal aspects of architecture.

Although Adler and Sullivan did substantial residential work, it was in their commercial work that they made their art-historic contribution. Most of their buildings were in Chicago, where the commercial expansion of the 1880s resulted in many important commissions.

The early years of the Adler and Sullivan practice did not result in buildings of lasting interest, however. It was the commission in 1886 to design the Auditorium Building in Chicago that marked the first period of Sullivan's design maturity. This project was a curious combination of a hotel and office block wrapped in a U-shape around an 3,982-seat auditorium for opera. Completed in December 1889, it is a ten-story-high building of granite and limestone with a 17-story tower. The noble arcaded exterior is very simple in profile, has little ornament, and owes much to the architect H.H. Richardson's design for the Marshall Field Wholesale Store, then recently completed in Chicago. The interior, however, is lavishly decorated with relief ornament and coloured stencilled patterns, all of which brought forth Sullivan's great talent for designing ornamentation. The interior of the auditorium (restored 1967) is particularly opulent and features gilded plasterwork and countless electric light bulbs. The decoration, which is dazzling and properly theatrical, owes nothing to historical eclecticism. The astonishingly effective acoustical design of the auditorium was the work of Adler, who was also responsible for all structural and mechanical aspects of the building.

Even before the auditorium proper was complete, the Adler and Sullivan firm moved to offices on the 16th floor of the tower, then the highest office suite in Chicago. It was there that the young Frank Lloyd Wright spent six years as apprentice to Sullivan. Wright left in 1893 after a quarrel with Sullivan, and it was not until 1914 that the friendship was renewed. He always acknowledged, however, the influence of Sullivan in shaping his work and ideas.

Influence of Sullivan on Wright

The ten-story Wainwright Building in St. Louis is the most important skyscraper designed by Sullivan. Unlike the Auditorium Building, the exterior walls of which are solid masonry and load bearing, it is of steel frame throughout, an idea advanced by William Le Baron Jenney in 1883–85 in Chicago. Jenney and others were unable to give visual expression to the height of a tall building and often resorted to unsuitable historical styles. Sullivan, however, took the problem in hand and made his design a "proud and soaring" unity. He gave his building a two-story base, above which the vertical elements are stressed and the horizontals, being recessed, are minimized. These vertical rhythms are capped by a deep decorative frieze (corresponding to the 10th floor) and a simple projecting cornice. The Guaranty (now Prudential) Building in Buffalo, New York, by Adler and Sullivan is similar to the Wainwright Building except that it has 16 stories and its surface is sheathed in decorative terra-cotta instead of red brick. Both buildings are among the best of Adler and Sullivan's work, the latter being the last work of their partnership.

The 1893 Columbian Exposition held in Chicago was a great disappointment to Louis Sullivan. The opportunity to design an international fair with imagination was passed over in favour of a loose adaptation of Classical architecture. The spectacle of an ensemble of these all-white buildings was an enormous success with the public. The Adler and Sullivan contribution was the Transportation Building, which stood apart and was painted in various strong colours as if in protest. It was a long, low arcaded building with a large polychromed archway entrance (the so-called "Golden Door"). Not all visitors were impressed by the neo-Roman grandeur of the fair. André Bouilhet, a delegate representing a Parisian decorative-arts union, praised the originality of the Transportation Building. Furthermore, he arranged for a small exhibit in Paris of Sullivan's work, including a plaster cast of the "Golden Door" and some photographs of his taller buildings. The exhibit later went to Russia and Finland. This European recognition, however, did not allay Sullivan's bitterness. He considered the Exposition a rejection of the progressive architecture that he saw developing in the Midwest. "The damage wrought by the World's Fair," he wrote, "will last for half a century from its date, if not longer. It has penetrated deep into the constitution of the American mind." It is with this event that Sullivan ended the *Autobiography of an Idea* (1924), his account of his career and his architectural theories.

Later work. The economic depression that began in 1893 severely curtailed commissions. With a lucrative offer as designer and agent for the Crane Elevator Company, Adler reluctantly decided in 1895 to withdraw from architecture. Sullivan reacted badly, accusing Adler of disloyalty. Adler's new job proved unsatisfactory, and he decided to return to architecture six months later. His offer to re-establish the firm was unwisely refused by Sullivan, and Adler opened his own office in another part of the Auditorium Building, where he practiced until his death in 1900.

In 1895, proud and optimistic, Sullivan began to practice by himself. His temperament was unsuited, however, to the handling of all of the phases of architectural practice. New work was very slow in coming. George Grant Elmslie, whom he had hired in 1889 at the age of 18, remained a loyal employee. Nevertheless, he was aware of Sullivan's shortcomings:

> He could be arrogant and unnecessarily decisive . . . prone to give advice where not needed, to good clients . . . he lost many jobs because he would not compromise his ideals, nor play fast and loose with vital conceptions of what was fitting for the purpose intended.

A simple tabulation of Sullivan's 20 commissions (many of them small) during the last 30 years of his life indicates the near collapse of his practice.

Among the few major commissions Sullivan received was the one for the Schlesinger & Mayer department store in Chicago, occupied by Carson Pirie Scott & Co. since 1904. Two connecting units were built between 1899 and 1904, and a third unit was added in 1906 by Daniel H. Burnham and Co., largely following Sullivan's original design. In contrast to the vertical emphasis of the Wainwright and Guaranty buildings, which are offices, the design for this department store stresses the horizontal. Particularly notable are the rectangular "Chicago windows"—each a large fixed pane flanked by moveable sash windows. The elegant simplicity of the upper floors is in contrast to the lavish decoration of the first two, which have windows that were treated as display windows, with the architectural decoration forming rich picture frames. This cast-iron ornament is based on a combination of geometric and stylized floral forms. Much of it is thought to have been designed by Elmslie, in emulation of Sullivan's style. In any case, the decoration of the Carson Pirie Scott Store, particularly the ornament over the main entrance, represents the height of Sullivan's achievement as a designer of architectural ornamentation. Only the decorative panels that once surrounded Sullivan's Gage Building in Chicago were a match for its decorative exuberance.

Commission for the Carson Pirie Scott Store

Greater plastic richness and a heightened subjectivity are clearly apparent in Sullivan's work after 1895. His 12-story Bayard (now Condict) Building in New York was richly embellished with molded terra-cotta and cast-iron ornament.

As his flourishing years with Adler became a provoking memory, Sullivan grew lonely and difficult. He became estranged from his brother Albert, who was a successful official of the Illinois Central railroad and who also lived in Chicago. The middle-aged Sullivan became something of a recluse, seeking solace in writing and in visits to his winter cottage in Ocean Springs, Mississippi. His marriage in 1899 to Margaret Davies Hattabough did little to bring lasting happiness; they were separated in 1906 and divorced in 1917 without having had children. With declining income, Sullivan moved to progressively cheaper hotels in an effort to economize. By 1909 a lack of commissions reduced him to desperate straits; he was forced to sell his library and household effects. Perhaps an equal loss was the departure that year of Elmslie, his assistant for 20 years, who went to join forces in Minneapolis with William Gray Purcell, an architect who had worked briefly for Sullivan in 1903.

Designs of banks in small towns

Particularly noteworthy projects undertaken in his last years were seven banks in a number of small Midwestern towns, beginning with the National Farmers' (now Security) Bank in Owatonna, Minnesota. Sullivan's work habits had become erratic, and it is known that this particular design is primarily the work of Elmslie, his last effort before leaving Sullivan's employ. The Owatonna bank has a simple cube form pierced on two sides by large arched windows. Its walls of red sandstone and brick, which convey a sense of security, are ornamented by bands of coloured mosaic and blue-green glazed terra-cotta. The balance between simple form and decoration in this structure has been much admired. The square interior was designed in harmony with the exterior: semicircular murals appear opposite the two arched windows.

Another attractive bank design is that of the Merchants' National Bank in Grinnell, Iowa (1914). Like the Owatonna bank, it has a relatively austere form, relieved by imaginative, intricate ornament. The facade is embellished with a spectacular decorative frame for the circular window above the entrance. Sullivan's last commission was the facade for the Krause Music Store in Chicago (1922).

Sullivan had to abandon his Auditorium tower suite in 1918 for a small second-floor office. In 1920 he had no office at all and was reduced to living in one bedroom, being supported by donations from friends. His workplace came to be a desk in the office of a Chicago terra-cotta company, where he was able to complete two significant projects: the writing of his *Autobiography* and the completion of 19 plates for *A System of Architectural Ornament According with a Philosophy of Man's Powers* (1924). He died on April 14, 1924, a week after he had received published copies of these two works. Sullivan was buried in Graceland Cemetery in Chicago, next to the graves of his parents and within sight of the Getty and Ryerson Tombs, which he had designed. Later, a modest stone was erected by friends. Much later, in 1946, the American Institute of Architects awarded him its Gold Medal.

Sullivan's writings

Assessment. Sullivan's importance lies in his writings as well as in his architectural achievements. Unfortunately, his prose style was sometimes obscure and verbose. He thought of himself as a philosopher-architect and assumed the role of teacher in *Kindergarten Chats*, which first appeared in 1901 as a series of articles in *Interstate Architect and Builder*. In 1905 he sought, unsuccessfully, a professorship in architecture at the University of Michigan. Sullivan's writings transcended the daily concerns of architectural practice and entered the realm of poetry. These writings, which are subjective and metaphorical, suggest directions for architecture, rather than explicit doctrines or programs. Sullivan himself warned of the danger of mechanical theories of art.

Sources of Sullivan's ideas have been traced to the mid-19th-century writings of two Americans, the sculptor Horatio Greenough and the essayist Ralph Waldo Emerson, as well as to the English naturalist Charles Darwin. Darwin's writings on evolution, particularly on organic growth, left their mark on European writers on architecture and, in turn, on Sullivan's own thinking. The French architect César-Denis Daly, for example, in an essay reprinted in a Chicago architectural journal, stated that

> each style of architecture . . . being born of the intellectual and moral forces of a human society . . . , has become naturally the expression of a certain civilization . . . The adoption by one age of a style . . . other than that which it has itself created, is hence in itself a false principle.

In a similar vein, the contemporary German architect Gottfried Semper wrote, "Style is the conformity of an art object with the circumstances of its origin and . . . its development." Out of such inquiries into the nature of style came Sullivan's own dictum "form follows function," a phrase that should not lead one to conclude that Sullivan believed that a design should be a mechanistic visual statement of utility. Rather, he believed that architecture must evolve from and express the environment in addition to expressing its particular function and its structural basis. It has been said that Sullivan was the first U.S. architect to think consciously of the relationship between architecture and civilization.

Importance of the skyscraper in his work

The skyscraper was central to both Sullivan's writing and practice, and it is on this subject that his thought is most concise. His pre-skyscraper commercial buildings in Chicago, such as the Rothschild Store and Troescher Building, show a conscious clarification and opening up of the facade. This simplification is carried into his "skyscrapers," the Wainwright and the Guaranty, which are conceived as "a single, germinal impulse or idea" that permeates "the mass and its every detail with the same spirit." The exceptional clarity of Sullivan's designs has lost some of its impact because contemporary architecture has in part absorbed his ideas. Sullivan considered it obvious that the design of a tall office building should follow the functions of the building and that, where the function does not change, the form should not change. Unfortunately, Sullivan's most dramatic skyscraper design, the Fraternity Temple (1891), intended for Chicago, was never built. This was to be a symmetrical structure with bold step-back forms and a soaring 35-story central tower.

In his designs Sullivan combined emphatically simple forms with rich decoration. For him ornament was integral and necessary to architecture. He believed

> that a decorated structure, harmoniously conceived, well considered, cannot be stripped of its system of ornament without destroying its individuality. It has been hitherto somewhat the fashion to speak of ornament . . . as a thing to be put on or omitted, as the case might be. I hold to the contrary—that the presence or absence of ornament should, certainly in serious work, be determined at the very beginnings of the design.

Sullivan's position on ornament has largely fallen out of favour because of the emphasis in postwar architecture on the almost total elimination of architectural ornament.

Sullivan was just as much a revolutionary in his ornament as he was in his use of plain surfaces and cubic forms. His ornament was not based on historical precedent but rather upon geometry and the stylized forms of nature. Although his early ornament has some links to that of the Gothic Revival style and to the Queen Anne style, his mature ornament, seen best in his works at the turn of the century, is indisputably his own. It stands as a curious yet unrelated parallel to Art Nouveau ornamentation in Europe. Crisp yet fluid, tightly constructed yet exuberant, these designs remind one of Sullivan's feeling that architecture should not only serve and express society but also illuminate the heart.

MAJOR WORKS

Rothschild Store, Chicago, Illinois (1880–81); Troescher Building, Chicago (1884); Auditorium Building, Chicago (1886–89); Wainwright Building, St. Louis (1890–91); Schiller Building, Chicago (1891–92); Transportation Build-

ing, Columbian Exposition, Chicago (1893; demolished); Guaranty (now Prudential) Building, Buffalo, New York (1894–95); Condict (formerly Bayard) Building, New York (1897–99); Gage Building, Chicago (1898–99); Carson Pirie Scott Store, Chicago (1899–1904); National Farmers' Bank (now Security Bank and Trust Company), Owatonna, Minnesota (1907–08).

BIBLIOGRAPHY. HUGH MORRISON, *Louis Sullivan: Prophet of Modern Architecture* (1935), is the standard biography, which also includes an evaluation of Sullivan's work. WILLARD CONNELY, *Louis Sullivan As He Lived: The Shaping of American Architecture* (1960), is a more personal account, although the full story of his later years will probably never be known. Sullivan's own *Autobiography of an Idea* (1924) and *Kindergarten Chats* (published serially in 1901–02), are indispensable for a grasp of Sullivan's architectural theory. The 1947 Wittenborn edition of the latter, *Kindergarten Chats and Other Writings* (rev. 1918), includes eight additional essays by Sullivan and a bibliography of writings by and about him. JOHN SZARKOWSKI, *The Idea of Louis Sullivan* (1956), contains dramatic photographs. ALBERT BUSH-BROWN, *Louis Sullivan* (1960), is a brief, interpretative account of Sullivan's contributions and is well illustrated.

(H.F.K.)

Sumerian Language

Sumerian is the oldest written language in existence. First attested about 3100 BC in southern Mesopotamia, it flourished during the 3rd millennium BC. About 2000 BC, Sumerian was replaced as a spoken language by Semitic Akkadian (Assyro-Babylonian) but continued in written usage almost to the end of the life of the Akkadian language, around the beginning of the Christian era. Sumerian never extended much beyond its original boundaries in southern Mesopotamia; the small number of its native speakers was entirely out of proportion to the tremendous importance and influence Sumerian exercised on the development of the Mesopotamian and other ancient civilizations in all their stages.

History. Four periods of Sumerian can be distinguished: Archaic Sumerian, Old or Classical Sumerian, New Sumerian, and Post-Sumerian.

Archaic Sumerian covered a period from about 3100 BC, when the first Sumerian records make their appearance, down to about 2500 BC. The earliest Sumerian writing is almost exclusively represented by texts of business and administrative character. There are also school texts in the form of simple exercises in writing signs and words. The Archaic Sumerian language is still very poorly understood, partly because of the difficulties surrounding the reading and interpretation of early Sumerian writing and partly because of the meagreness of sources.

Old, or Classical, Sumerian

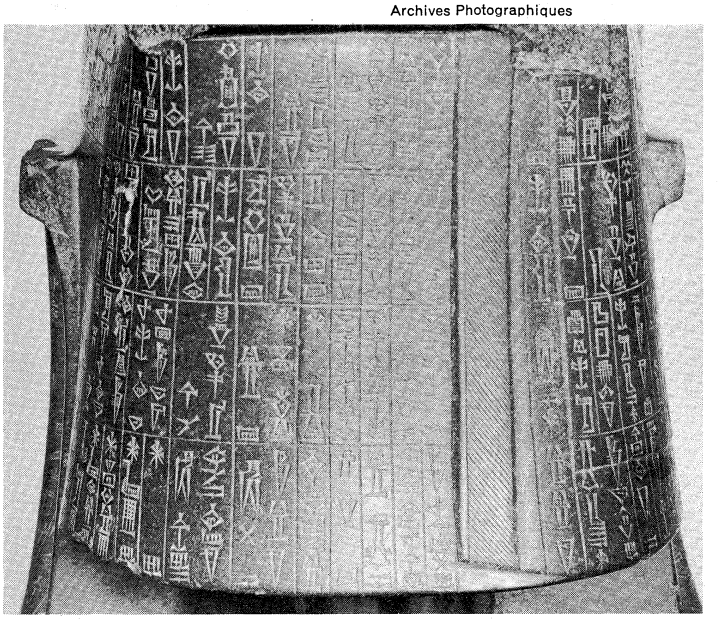
The Old, or Classical, period of Sumerian lasted from about 2500 to 2300 BC and is represented mainly by records of the early rulers of Lagash. The records are business, legal, and administrative texts, as well as royal and private inscriptions, mostly of votive character, in letters, both private and official, and in incantations. These sources are much more numerous than those of the preceding period, and the writing is explicit enough to make possible an adequate reconstruction of Sumerian grammar and vocabulary.

During the period of the Sargonic dynasty, the Semitic Akkadians took over the political hegemony of Babylonia, marking a definite setback in the progress of the Sumerian language. At this time the Akkadian language was used extensively throughout the entire area of the Akkadian empire, while the use of Sumerian gradually was limited to a small area in Sumer proper. After a brief revival during the 3rd dynasty of Ur, the New Sumerian period came to an end about 2000 BC, when new inroads of the Semitic peoples from the desert succeeded in destroying the 3rd dynasty of Ur and in establishing the Semitic dynasties of Isin, Larsa, and Babylon.

The period of the dynasties of Isin, Larsa, and Babylon is called the Old Babylonian period, after Babylon, which became the capital and the most important city in the country. During this time the Sumerians lost their political identity, and Sumerian gradually disappeared as a spoken language. It did, however, continue to be written to the very end of the use of cuneiform writing. This is the last stage of the Sumerian language, called Post-Sumerian.

In the early stages of the Post-Sumerian period the use of written Sumerian is extensively attested in legal and administrative texts, as well as in royal inscriptions, which are often bilingual, in Sumerian and Babylonian. Many Sumerian literary compositions, which came down from the older Sumerian periods by way of oral tradition, were recorded in writing for the first time in the Old Babylonian period. Many more were copied by industrious scribes from originals now lost. The rich Sumerian literature is represented by texts of varied nature, such as myths and epics, hymns and lamentations, rituals and incantations, and proverbs and the so-called wisdom compositions. For many centuries after the Old Babylonian period, the study of Sumerian continued in the Babylonian schools. As late as the 7th century BC, Ashurbanipal, one of the last rulers of Assyria, boasted of being able to read the difficult Sumerian language, and from an even later period, in Hellenistic times, there are some cuneiform tablets that show Sumerian words transcribed in Greek letters.

Archives Photographiques



Sumerian inscription, detail of a diorite statue of Gudea of Lagash, 22nd century BC. In the Louvre, Paris.

Rediscovery. Around the time of Christ, all knowledge of the Sumerian language disappeared along with that of cuneiform writing, and in the succeeding centuries even the name Sumer vanished from memory.

Unlike Assyria, Babylonia, and Egypt, whose histories and traditions are amply documented in biblical and classical sources, there was nothing to be found in non-Mesopotamian sources to make one even suspect the existence of the Sumerians in antiquity, let alone fully appreciate their important role in the history of early civilizations.

Decipherment of Sumerian cuneiform

When the decipherment of cuneiform writing was achieved in the early decades of the 19th century, three languages written in cuneiform were discovered: Semitic Babylonian, Indo-European Persian, and Elamite, of unknown linguistic affiliation. Only after the texts written in Babylonian had become better understood did scholars become aware of the existence of texts written in a language different from Babylonian. When the new language was discovered it was variously designated as Scythian, or even Akkadian (that is, by the very name now given to the Semitic language spoken in Babylonia and Assyria). It was only after knowledge of the new language had grown that it was given the correct name of Sumerian.

Characteristics. The linguistic affinity of Sumerian has not yet been successfully established. Ural-Altaic (which includes Turkish), Dravidian, Brahui, Bantu, and many

other groups of languages have been compared with Sumerian, but no theory has gained common acceptance. Sumerian is clearly an agglutinative language in that it preserves the word root intact while expressing various grammatical changes by adding on prefixes, infixes, and suffixes. The difference between nouns and verbs, as it exists in the Indo-European or Semitic languages, is unknown to Sumerian. The word *dug* alone means both "speech" and "to speak" in Sumerian, the difference between the noun and the verb being indicated by the syntax and by different affixes.

The distinctive sounds (phonemes) of Sumerian consisted of four vowels, *a, i, e, u,* and 16 consonants, *b, d, g,* ŋ, *h, k, l, m, n, p, r, s,* ś, š, *t, z.* In Classical Sumerian, the contrast between the consonants *b, d, g, z* and *p, t, k, s* was not between voiced (with vibrating vocal cords) and voiceless consonants (without vibrating vocal cords) but between consonants that were indifferent as to voice and those that were aspirated (pronounced with an accompanying audible puff of breath). The semivowels *y* and *w* functioned as vocalic glides.

In the noun, gender was not expressed. Plural number was indicated either by the suffixes *-me* (or *-me* + *esh*), *-hia,* and *-ene,* or by reduplication, as in *kur* + *kur* "mountains." The relational forms of the noun, corresponding approximately to the cases of the Latin declension, include: *-e* for the subject (nominative), *-a(k)* "of" (genitive), *-ra* and *-sh(e)* "to," "for" (dative), *-a* "in" (locative), *-ta* "from" (ablative), *-da* "with" (commitative).

The Sumerian verb

The Sumerian verb, with its concatenation of various prefixes, infixes, and suffixes, presents a very complicated picture. The elements connected with the verb follow a rigid order: modal elements, tempo elements, relational elements, causative elements, object elements, verbal root, subject elements, and intransitive present–future elements. In the preterite transitive active form, the order of object and subject elements is reversed. The verb can distinguish, in addition to person and number, transitivity and intransitivity, active and passive voice, and two tenses, present–future and preterite.

Several Sumerian dialects are known. Of these the most important are *eme-gir,* the official dialect of Sumerian, and *eme-SAL,* the dialect used often in the composition of hymns and incantations (see also WRITING, FORMS OF).

BIBLIOGRAPHY. ARNO POEBEL, *Grundzüge der sumerischen Grammatik* (1923), partly out of date, but still the only full grammar of Sumerian in all its stages; ADAM FALKENSTEIN, *Grammatik der Sprache Gudeas von Lagaš,* 2 vol. (1949–50), a very thorough grammar of the New Sumerian dialect, and *Das Sumerische* (1959), a very brief, but comprehensive survey of the Sumerian language; CYRIL J. GADD, *Sumerian Reading Book* (1924), outdated, but the only grammatical tool in English; SAMUEL N. KRAMER, *The Sumerians* (1963), provides a general introduction to Sumerian civilization. There is no vocabulary, not even a glossary, of Sumerian; only sign lists exist, such as ANTON DEIMEL, *Šumerisches Lexikon,* vol. 1, 3rd ed. (1947) and vol. 2–4 (1928–50).

(I.J.G.)

Sun

The dominant body of the solar system, the Sun is Earth's nearest star, eight light-minutes (93,000,000 miles) away from it. It is a typical star, approximately midway between the largest and the smallest, the brightest and the faintest stars known. On the basis of its temperature and radius (see below) it is called a G1 type dwarf star.

The Sun's energy comes from the conversion of hydrogen to helium at a temperature of about 15,000,000 degrees Kelvin (27,000,000° F), deep in its core where pressures are many thousand million (10^9) times that of the Earth's atmosphere at sea level and the density is 150 times that of water. The energy released in the core by nuclear processes slowly works its way to the surface, where it is radiated into space. The enormous energy output, of which the Earth intercepts only one part in 2,200,000,000, has been sufficiently constant over geologic time scales of hundreds of millions of years for complex life-forms, governed by cycles of day and night, by seasons, or by tides, to develop.

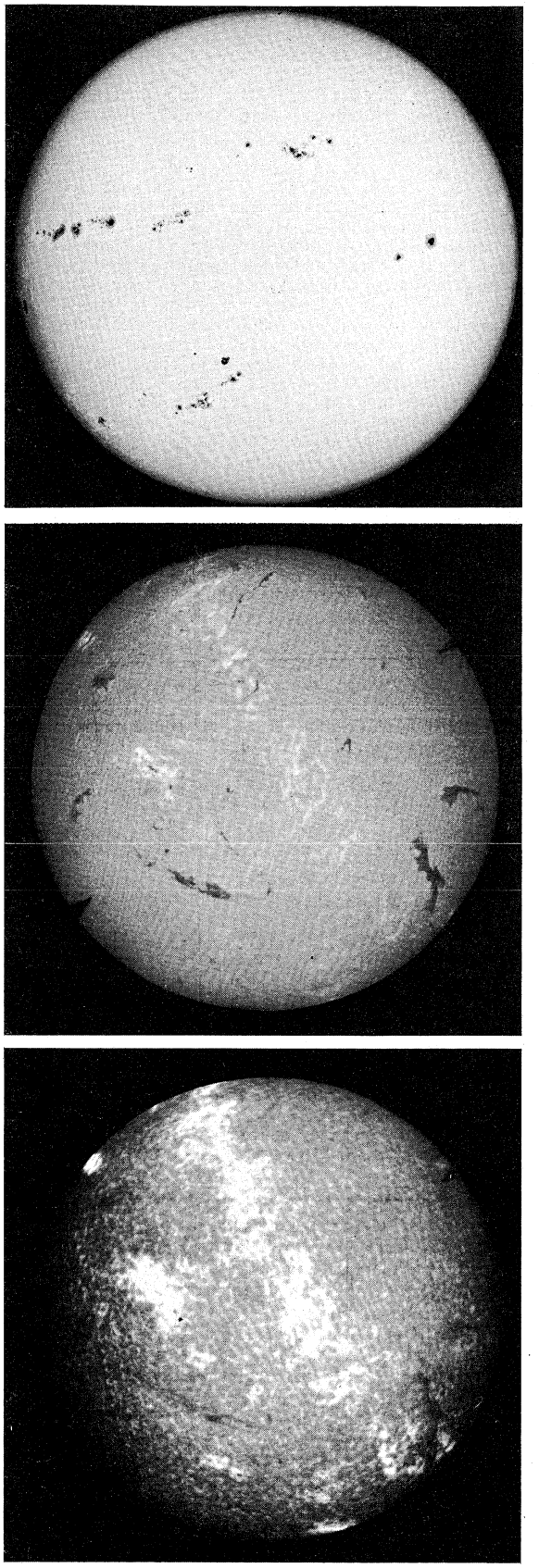


Figure 1: *Photographic studies of the Sun.*
(Top) Taken in ordinary light at sunspot maximum, December 21, 1957. (Centre) Taken in red hydrogen light (H-alpha), September 15, 1949. (Bottom) Calcium spectroheliogram taken September 15, 1949.
By courtesy of Hale Observatories

The visible surface of the Sun is the photosphere (Figure 1, top). Above it a layer 5,000 kilometres (3,100

miles) thick constitutes the inner atmosphere, or chromosphere, while above this lies the exceedingly tenuous high-temperature corona, which extends to the Earth and beyond. The Sun has a magnetic field; it rotates; it generates energy in its core. The outflow of energy from the deep interior through the surface, and the surface manifestations of the magnetic field and solar rotation combine to make a complex and fascinating picture of solar phenomena that can be observed in great detail. Their study constitutes modern solar astronomy.

The Sun's distance, defined as the semi-major axis of the Earth's orbit, the average of its greatest and least distances from Earth, is the astronomical unit, adopted by the International Astronomical Union as 149,600,000 kilometres, or 92,957,000 miles. Because the Earth's orbit is elliptical, its distance from the Sun varies during the year. It is closest to the Earth on January 3 when it is 147,103,000 kilometres (91,410,000 miles) and farthest from the Earth on July 4 when it is 152,106,000 kilometres (94,519,000 miles).

History of solar observations

For centuries the Sun was worshipped as a deity and was, therefore, not often the subject of physical study. Anaxagoras, the 5th-century-BC philosopher of Athens, supposed a great meteorite, which fell at Aegospotami in the daytime about 467 BC, to have come from the Sun, and he concluded that the Sun was a mass of red-hot iron larger than the Peloponnesus. With the invention of the telescope, Galileo Galilei, Johannes Fabricius, Christoph Scheiner, and Thomas Harriot almost simultaneously (1610–11) discovered sunspots (Figure 1), but it was the genius of Galileo that recognized their true nature as solar phenomena. Two centuries later, a German amateur astronomer, Samuel Heinrich Schwabe (1843), having diligently observed the Sun for 33 years, announced that the average number of spots varies cyclically with a period of about ten years. In 1852 the cyclic period was found to be 11.2 years and the possibility of there being still another period of 80 years was recognized. In 1858 it was noted that the sunspots of a new cycle appear at latitudes of ±30° and, as the cycle progresses, the spots break out closer to the Sun's equator so that the last spots of the cycle appear at about ±8°.

Table 1: Data for the Sun

Distance	149,600,000 km 92,957,000 mi
Mean angular diameter	31′59.3″
Diameter	1,392,000 km (864,950 mi) (109.3 times the Earth's diameter)
Surface area	6.087×10^{12} km² (2.35×10^{12} sq mi) (12,000 times the Earth's area)
Volume	1.412×10^{18} km³ (3.387×10^{17} cu mi) (1,306,000 times the Earth's volume)
Mass	1.99×10^{30} kilograms (2.19×10^{27} tons) (333,400 times the Earth's mass)
Mean density	1.41, *i.e.*, less than 1½ times that of water
Escape velocity to infinity from the Sun's surface	618 km per sec (384 mi per second)
Effective temperature	5,740° K
Total energy output	3.86×10^{33} ergs per second
Radiation from the Sun's surface	6.35×10^{10} ergs/cm²/sec
Brightness at the centre of the disk	8.23×10^{5} lamberts
Candlepower	3.17×10^{27} candles
Light flux	141,400 lux, at the Earth's mean distance, outside the atmosphere 103,000 lux, the average illumination from the Sun at the zenith
Period of rotation	
Equator	26.9 days (synodical)
Sunspot zone (16° N)	27.3 days (synodical)
Pole	31.1 days (synodical)
The Sun's interior	
Density	150 gm/cm³
Temperature	15,000,000° K*
Pressure	4×10^{17} dynes/cm² (4×10^{11} atmospheres)
The Sun as a star	
Spectral type	G1
Colour index	Mpg-Mpv = +0.45
Visual magnitude	Mpv = −26.86
Absolute visual magnitude	Mpv = +4.71
Velocity in space relative to nearby stars	19.4 km per sec in the direction of Lyra (Vega). The motion of the local stars of which the Sun is a member is about the galactic centre with a velocity of 220 km per sec

*Different calculations give results from 13,000,000°–25,000,000°.

Establishment of the first magnetic observatory

A German mathematician and astronomer, Carl Friedrich Gauss, established the first magnetic observatory in 1834 at Göttingen. Others soon followed. Small daily magnetic changes were shown to be related to the sunspot cycle in 1857. In 1904 intense magnetic storms were found to be associated with the passage of large sunspot groups over the central meridian.

Many important advances in solar astronomy have resulted from the construction of new telescopes and ancillary instrumentation of great sophistication. The physical and chemical nature of sunspots became known only after the development of the spectroscope. In 1870 dark bands were discovered in the spectrum of light radiated from sunspots, indicating the presence of molecular compounds; furthermore, it was postulated that the widening and splitting of some lines and the complete reversal of some dark lines to bright ones were caused by gaseous eruptions in the neighbourhood of sunspots—now called flares. The outflow of gases from the cooler dark centre of the spot (the umbra) to the edge of the penumbra, or outer part, at a speed of two kilometres per second (somewhat greater than one mile per second) was discovered spectroscopically in 1909. A reverse flow, inward, toward the centre of the spot, for the gases at the highest levels of the atmosphere above it, was found in 1913. The invention of the spectroheliograph by the U.S. astronomer George Ellery Hale in 1891 made possible the examination of the Sun in hydrogen light, calcium light, and the light radiated by ions or atoms of other elements, bringing out a wealth of detail in the chromosphere. In 1908 Hale looked for splitting and polarization of some atomic lines (the Zeeman effect) in spot spectra, but it required the large solar tower telescopes and pit spectrograph designed and constructed by Hale at Mount Wilson Observatory, California, now part of the Hale Observatories, to prove the magnetic character of sunspots. Between 1914 and 1924 Hale developed the laws of alternating sunspot polarity (see below). Although he also attempted to measure the Sun's general magnetic field, reliable measures were achieved only after the invention of the photoelectric magnetograph in 1948.

Joseph von Fraunhofer, a German physicist with greatly improved spectroscopes of his own construction, had visually found and mapped many (574) dark lines in the solar spectrum; to the strongest of these he assigned letters—*e.g.*, A, a, B, C, D—that are still used to identify the lines. The physical significance of the dark lines was explained by another German physicist, Gustav Robert Kirchhoff, in 1859. Kirchhoff conceived the idea of a hot Sun surrounded by cooler layers of vapour, which he called the reversing layers, in which the dark lines were supposed to have been formed. By comparison of laboratory spectra with the solar spectrum, the presence of eight elements in the Sun was later established. Henry Augustus Rowland, a U.S. physicist, published in 1897 a 40-foot photographic map of the solar spectrum that was of the highest quality and that allowed for the presence of 39 chemical elements in the Sun. Rowland's map extended to the ultraviolet limit caused by terrestrial ozone. This limit has only recently been surpassed by observations from space; now the spectrum has been mapped from the visible to less than one angstrom (10^{-10} metre). Toward red, the solar spectrum was recorded to 53,000 angstroms in 1888. Today, with new detectors, the limit is set only by terrestrial water vapour bands that obscure the infrared spectrum beyond 25 micrometres (25×10^{-6} metre). The complete spectrum can be observed only from very high altitudes, above the water-vapour absorption and the ozone absorption.

Attempts to measure the temperature of the Sun

The determination of the Sun's temperature has been one of the most difficult problems in solar astronomy. The English astronomer Sir John Herschel at the Cape of Good Hope and Claude-Servais-Mathias Pouillet in France (1837) observed that the vertical rays of the Sun, if totally absorbed, would raise the temperature of a layer of water 1.8 centimetres deep by 1° C per minute. The

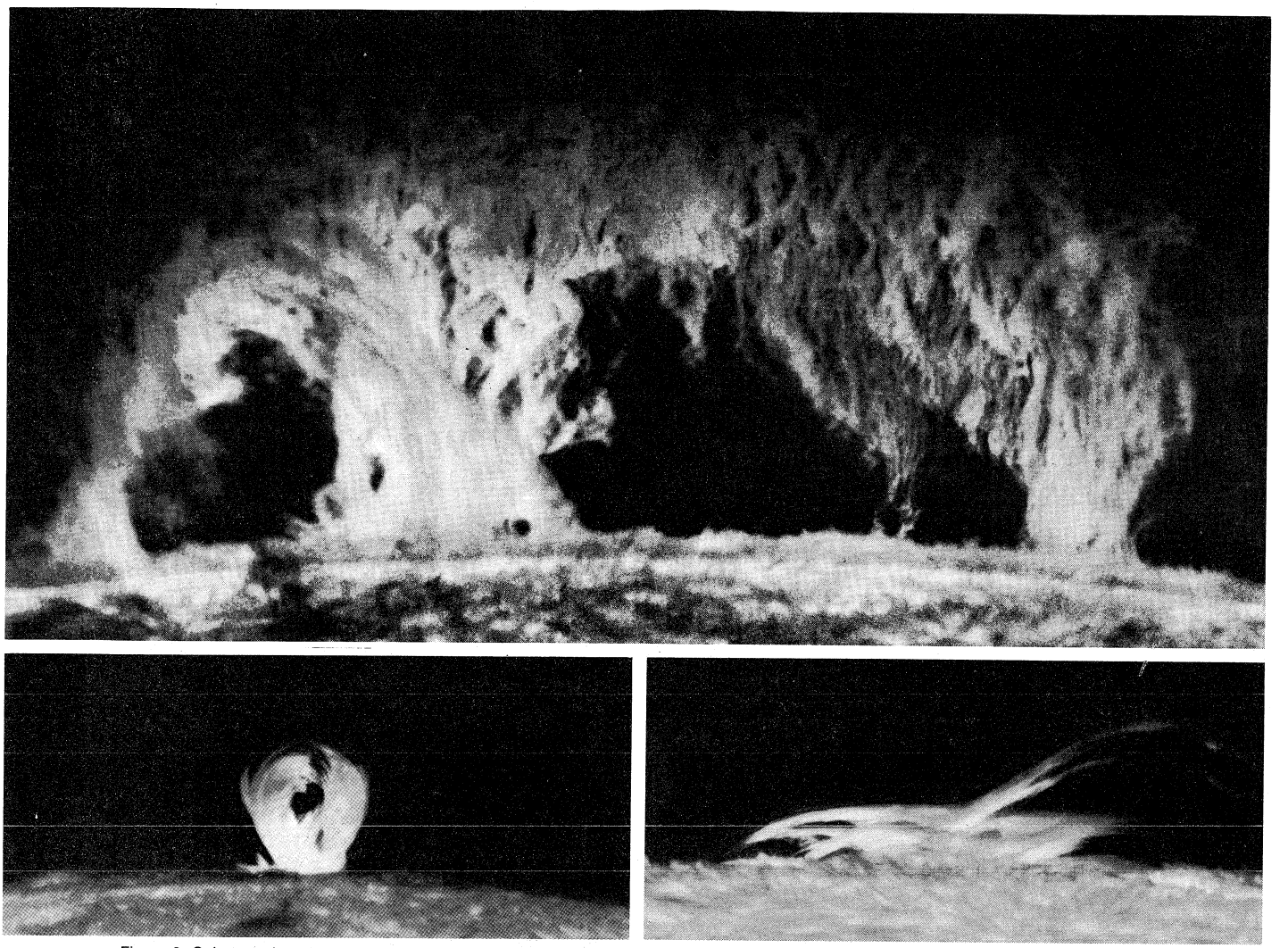


Figure 2: Solar prominences showing streamers, arches, loops, and other phenomena in the light of hydrogen (H-alpha). (Top) March 11, 1971. (Left) August 3, 1970. (Right) September 23, 1971.
By courtesy of Hale Observatories

measurement is in principle easy, but atmospheric absorption has been the uncertain factor. The problem was solved in 1881 and again in 1904, when it was shown that extrapolations must be made for solar radiation of different wavelengths, to determine what the intensity of each would be if there were no intervening air; and the values for all wavelengths must then be summed. A value of 1.96 calories per square centimetre per minute was obtained for the solar constant (the flux of total radiation received outside the Earth's atmosphere per unit area at mean Earth–Sun distance), that only recently has been improved by observations from balloons and high-flying aircraft. Early estimates of the Sun's temperature ranged from 1,461° C (2,660° F) to 10,000,000° C (18,000,000° F). Transforming the solar constant into the energy radiated by the Sun and thence into its effective temperature was impossible until the law that connects the temperature of a surface with the amount of heat radiated per second—the Stefan-Boltzmann law—was established about 1880.

The Moon, with its opaque disk of almost exactly the same apparent size as the Sun, and with a sharply defined edge, has been valuable to the solar astronomer at times of solar eclipse. In the Middle Ages, prominences on the Sun (see Figure 2) that became visible during eclipses were thought to be part of the Moon's atmosphere, mirages, clouds, or holes in the Moon. During an eclipse of July 1860, which crossed Spain, prominences were successfully photographed, and, by observing the motion of the Moon across them, it was demonstrated that these were solar. During an eclipse of August 1868, observers in India and in England, struck by the brilliance of the prominences, independently showed that they might be seen in broad daylight simply by opening the slit of a spectroscope and aiming it just outside the brilliant disk of the Sun's photosphere. The invention (1930) of the coronagraph made it possible for the first time to study the corona without an eclipse from high-altitude stations and on a nearly continuous basis. With a modified instrument, the K-coronameter (1950), which uses photoelectric detection, it became possible to view the corona through hazy skies at sea level, and measurements at very clear sites were greatly extended, to large distances from the limb of the Sun.

The mysterious coronal spectrum—at first attributed to the mythical element coronium—was discovered in 1941 to be made up of the so-called forbidden lines of highly ionized atoms of iron, nickel, and other elements. The lines are described as forbidden because they are almost impossible to produce in the laboratory.

Detection of radio waves

Radio waves from the Sun were first detected in 1942 on British radar sets. It was suggested that the bursts of radio energy were associated with an active sunspot visible on the surface of the Sun and, in particular, with the large solar flare of February 28. In the same year, a weak background radio emission, or noise, was detected from the "quiet" Sun; *i.e.*, the Sun observed when sunspots and related phenomena are at a minimum. The science of radio astronomy has since become a large and important field of solar research.

The modern concept of energy generation in the Sun's interior follows from Sir Arthur Stanley Eddington's suggestion that only through nuclear processes could sufficient energy be produced to have sustained the output of energy from the Sun throughout the known lifetime of the solar system. More details were provided in 1937 and 1938 when the energy production from nuclear reactions, capable of transforming hydrogen to helium, were calculated. Later experimental and theoretical work has given improved values for all the reactions considered possible for energy production.

The quiet Sun

PHOTOSPHERE

The visible luminous surface of the Sun is the photosphere or light sphere. When carefully examined through a telescope, it is seen to have a mottled appearance called granulation (Figure 3); the most conspicuous objects to be seen, however, are sunspots which form and dissipate and partake in the solar rotation. The photosphere radiates most of the Sun's energy; it comprises a thin but rather opaque layer approximately 400 kilometres (290 miles) thick, the temperature of which varies from 10,000° K (about 18,000° F) at the bottom to 4,200° K (about 7,560° F) at the top. A typical density in the photosphere is around 10^{-8} gram per cubic centimetre, about $\frac{1}{1,000}$ that of air at sea level.

By courtesy of Project Stratoscope of Princeton University sponsored by National Aeronautics and Space Administration, Office of Naval Research, and the National Science Foundation

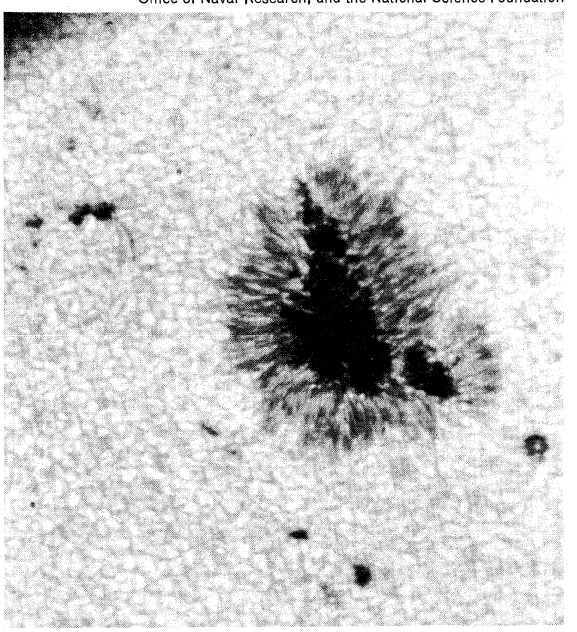


Figure 3: Granulation on the Sun showing a sunspot (centre right), which appears dark by contrast with the hotter regions surrounding it.

Granulation. About 4,000,000 granules, each one the top of a hot cloud of gas, cover the surface of the Sun. Because of their small size—300 to 1,450 kilometres, or 190–900 miles—they can be seen and photographed from the ground only when the Earth's atmosphere is steady and free of turbulence. Excellent photographs have been made with telescopes carried by balloons to heights of 20 kilometres—about 12 miles. The best of these show that granules are irregular, often polygonal, and that each is surrounded by a narrow, relatively dark, lane. They last about seven to ten minutes. A typical bright granule starts as a small grain the diameter of which grows to about 1,500 kilometres (930 miles) before it divides into several parts that spread and fade. Each granule rises with a speed of about 0.5 kilometre (about 0.3 mile) per second; as the energy radiates, the gas cools and descends rapidly in the dark intergranular lanes. A remarkable five-minute vertical oscillating motion of the granules was discovered in 1962. It has also been observed that after a bright granule forms, its upward movement sets the chromosphere into oscillation, which then continues for three to five periods after the disappearance of the exciting granule.

Limb darkening. The limb of the Sun (the visible edge) appears darker than the centre. The direct view downwards at the centre of the solar disk penetrates, on the average, into deeper, hotter layers of the photosphere; at the limb the higher, cooler layers are seen. Limb darkening (which exists for all stars) is due solely to a temperature gradient in the atmosphere, the deeper layers being hotter. Conversely, measures of the limb darkening lead to a determination of the temperature distribution with depth.

The Sun's temperature. The Sun emits energy over all wavelengths, from X-rays to radio waves. Approximately 40 percent of the emitted energy is in the visible portion of the spectrum, 50 percent is in the infrared, and the remainder is in the ultraviolet. Radiation escaping from the Sun's surface into space comes from different depths of the Sun's atmosphere, each with a different temperature. The boundary temperature is about 4,200° K (about 7,000° F), but the eye can see to depths where the temperature is 10,000° K; the Sun has, however, a well-defined effective temperature, namely, that of an equal sphere radiating a like amount of total energy in accordance with physical (blackbody) radiation laws. From the total energy radiated a temperature of about 5,700° K, roughly 10,000° F (see STAR) is found. Different (colour) temperatures are derived from different spectral regions. A colour temperature of 6,000° K (11,000° F) fits the whole visible spectrum reasonably well.

Solar energy

It is not easy to describe the enormous output of solar energy. If the Sun were enveloped in a shell of ice 40 feet thick, it would thaw its way out in one minute of time. If an ice bridge two miles in diameter could be built, spanning the immense distance of 93,000,000 miles from Sun to Earth, and, if by some means, the whole of the Sun's radiation could be concentrated upon this ice bridge, in one second it would be water and in seven seconds more it would be dissipated into vapour. Though the energy received at the Earth's surface amounts to over 4,000,000 horsepower per square mile, to harness it effectively is difficult. To achieve high temperatures requires concentrators or large parabolic collectors, such as those used in solar furnaces. Except for the very specialized experimental usage in heating homes, in cooking, in solar furnaces and solar-heat engines, which are subject to the variations of the weather and to day and night alternation of temperature, the present economic picture is not particularly favourable to the general utilization of solar energy.

The Sun's rotation and sphericity. Galileo gave the first proof of the Sun's rotation and shape. He noted the Sun's roundness and saw that the apparent motion of sunspots in parallel lines with diminished movement at the limbs could be explained if the spots were on the surface of a sphere rotating from east to west. Modern spectrographic observations of the Doppler shift (a change of wavelength with motion toward or away from the observer) of Fraunhofer lines show that the equator is rotating faster than the poles. Unlike some stars, which rotate rapidly and are strongly flattened oblate spheroids, the Sun is a slow rotator and is so nearly spherical that it is uncertain whether any flattening can be detected. A polar diameter 70 kilometres less than the equatorial diameter has been suggested by some observations. The ellipticity of the Sun has been of interest to scientists because the gravitational field of a nonspherical body affects planetary orbits in a way different from that of a sphere. It has been suggested that part of the perihelion advance (a gradual rotation of the point closest to the Sun, which is otherwise very satisfactorily explained by the theory of of relativity) of Mercury's orbit could be explained in this manner.

The Fraunhofer spectrum. The solar spectrum, within the wavelength interval 2950–10,000 angstroms, shows about 25,000 so-called Fraunhofer dark lines (see above *History of solar observations* and Figure 1, centre). The

identification of elements in the Sun is still carried out by the classical method comparing the solar lines with lines observed in the laboratory for both wavelength and intensity. Care must be taken to distinguish between solar lines and lines produced by water vapour, oxygen, and other constituents of the Earth's atmosphere.

Element identification

About 73 percent of the lines have been identified; 63 elements and 11 molecules are recorded. Iron contributes the greatest number (5,458) of identified lines. There are 1,453 lines from chromium, 1,344 from titanium, 856 from nickel, 388 from zirconium, 1,572 from solar cyanogen, and 1,477 from terrestrial water. For many elements, however, only a few lines are observed, for example: two from beryllium, two from silver, six from platinum, one from gold. The majority of the unidentified lines are faint and probably due to molecules and to the faint lines of the abundant elements in the Sun.

Modern techniques of spectrum analysis make use of the most powerful spectrographs that can be built. This makes it possible to study each spectrum line in detail for position, shape, depth and width.

A Fraunhofer line has considerable width because of the turbulent motions of the solar gases and collisions caused by neighbouring atoms. The resulting line profile, measuring the intensity against wavelength (or frequency), is that of an inverted bell-shaped curve near the line centre with broad, shallow wings, which, for strong lines, may extend many angstroms from the line centre.

The Fraunhofer spectrum has been extended to include lines in the far ultraviolet from rocket observations above the Earth's atmosphere. Within the far ultraviolet region, in the interval 3000–2097 angstroms, 7,146 lines are observed. Below 2100 angstroms, the intensity of the spectrum drops and by 1500 angstroms the radiation comes from the chromosphere and the absorption spectrum has turned into an emission spectrum. This emission spectrum contains the strong resonance lines of the elements first identified from weaker subordinate lines in the visible spectrum. Also, higher stages of ionization for a given element are observed in the far ultraviolet emission spectrum.

The Sun's infrared spectrum is dissected by numerous terrestrial water vapour absorption bands. Thus, at ground level the solar spectrum must be observed in the "windows" between the bands. To observe the entire infrared spectrum of the Sun, it is necessary to use a vehicle such as a high-flying aircraft, a stratospheric balloon, or a space vehicle that can carry the infrared spectrometer above all, or most of, the Earth's absorbing atmosphere. In the wavelength interval 11,984–25,578 angstroms, 1,786 solar lines and 6,911 telluric lines—*i.e.*, lines caused by substances in Earth's atmosphere—have been listed. Silicon, iron, carbon, carbon monoxide, and titanium contribute most of the solar lines, while water vapour, carbon dioxide, methane, and isotopic bands of these molecules make up most of the telluric lines.

The chemical composition of the Sun. The proportions (abundances) of the elements in the Sun can be determined directly only for the surface layers, the so-called solar atmosphere. In the interior, the abundances of the elements may be different because the elements are transformed there at different rates by nuclear reactions. The chemical composition of the solar atmosphere is thought to be representative of the composition of the Sun at the time of its formation, before such alterations occurred. This initial composition is relevant to the study of the evolution of the Sun and the stars. The chemical composition of the solar atmosphere can be determined in various ways, which depend on the types of observation available.

Solar cosmic rays are emitted from active regions of the Sun. Abundances of the following light elements have been derived from analysis of solar cosmic rays: helium, beryllium jointly with boron, carbon, nitrogen, oxygen, fluorine, and neon. Since the heavier elements are difficult to separate by nuclear counting methods, their abundances are generally given for groups, such as phosphorus to scandium and titanium to nickel.

Abundances derived from observations of the solar emission lines that originate in the higher chromosphere and the corona are called coronal abundances. The coronal-emission spectrum may be analyzed in two different ways, depending on whether the observations were secured on the disk or above the limb of the Sun. The far-ultraviolet-emission lines observed on the disk of the Sun are produced by atoms in many different stages of ionization. Analyses have been carried out for the more abundant atoms, such as helium, carbon, nitrogen, oxygen, neon, sodium, magnesium, aluminum, silicon, phosphorus, sulfur, calcium, and iron. The emission lines in the interval 3000–10,000 angstroms seen above the solar limb are a result of "forbidden" electron transitions within atoms; *i.e.*, transition almost impossible to obtain on Earth. Since the forbidden emission lines originate in the corona, and are thus extremely weak compared to the radiation from the body of the Sun, they can only be observed at the time of a total solar eclipse or by means of a coronagraph. The known forbidden emission lines are due to argon and calcium, and to atoms of the iron group, such as chromium, manganese, iron, cobalt, and nickel, all of them being highly ionized.

The most commonly used method of determining the abundance of elements in the solar atmosphere makes use of the solar absorption (Fraunhofer) spectrum, which extends from the near ultraviolet, through the visible, into the infrared wavelength regions. The majority of Fraunhofer lines are formed in the lowest layers of the solar atmosphere, that is, the photosphere and low chromosphere. The abundances derived from these lines are called photospheric abundances.

Photospheric abundances

The determination of photospheric abundances is based on a measurement of the intensity of the lines. With modern instruments a solar spectrum line can be recorded at very high resolution (Figure 4) and, consequently, the intensity distribution in the line (the line profile) can be obtained with very good precision. The line profile and the measurement of the total intensity absorbed by each line depend not only on the abundance and physical characteristics of the particular atom (or molecule) producing the line but also on the physical conditions (temperature, density) that prevail in the layers of line formation. Thus, in an analysis of the solar spectrum, both the physical state of the solar atmosphere and its chemical composition are determined at once.

In Table 2 the results of abundance determinations for the photospheric and the coronal abundances, both as determined in 1968, are shown. For the elements lithium,

By courtesy of Kitt Peak National Observatory, Tucson, Arizona

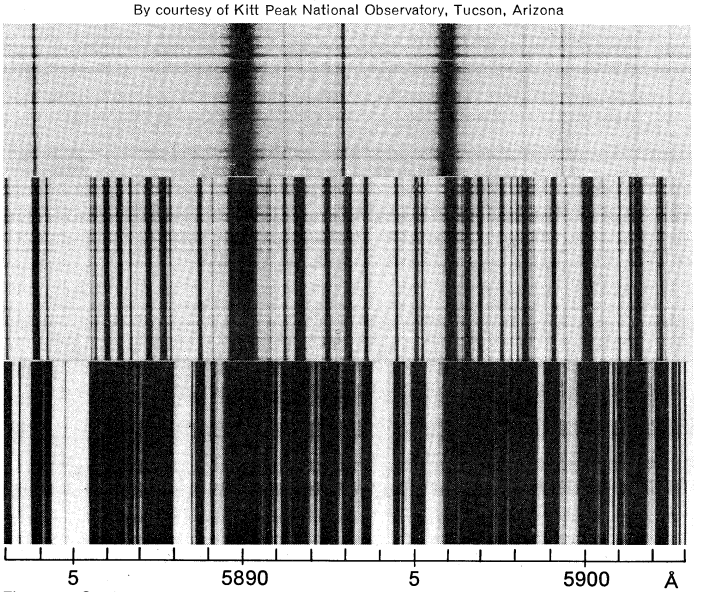


Figure 4: Sodium D lines observed at high Sun (top spectrum) and, through a greater amount of air, at low Sun (middle spectrum). The bottom spectrum shows water vapour bands for comparison. The strengthening of terrestrial water vapour lines at low Sun is most apparent.

Table 2: Solar Abundances
(log Σ relative to log Σ = 12 for hydrogen)

Z atomic number	element	log Σ photosphere	log Σ corona (forbidden lines)	log Σ corona (UV lines)
1	H	12.00	12.00	12.00
2	He	—	—	11.2
3	Li	1.0	—	—
4	Be	1.2	—	—
5	B	< 2.5	—	—
6	C	8.5	—	8.8
7	N	8.1	—	7.8
8	O	8.8	—	8.7
9	F	4.6	—	—
10	Ne	—	—	7.8
11	Na	6.3	—	—
12	Mg	7.4	—	7.6
13	Al	6.2	—	6.5
14	Si	7.2	—	7.8
15	P	5.3	—	—
16	S	7.3	7.0	7.3
18	Ar	—	7.3	—
19	K	4.7	5.8	—
20	Ca	6.0	6.8	—
21	Sc	2.9	—	—
22	Ti	4.8	—	—
23	V	4.2	—	—
24	Cr	5.0	6.0	—
25	Mn	4.9	5.8	—
26	Fe	7.6	7.8	7.8
27	Co	4.7	5.5	—
28	Ni	5.8	6.7	6.6
29	Cu	4.5	—	—
30	Zn	3.5	—	—
31	Ga	2.7	—	—
32	Ge	2.5	—	—
37	Rb	2.5	—	—
38	Sr	3.0	—	—
39	Y	3.2	—	—
40	Zr	2.7	—	—
41	Nb	2.3	—	—
42	Mo	2.3	—	—
44	Ru	1.8	—	—
45	Rh	1.4	—	—
46	Pd	1.6	—	—
47	Ag	0.8	—	—
48	Cd	1.5	—	—
49	In	1.5	—	—
50	Sn	1.5	—	—
51	Sb	1.9	—	—
56	Ba	2.1	—	—
57	La	2.0	—	—
58	Ce	1.8	—	—
59	Pr	1.5	—	—
60	Nd	1.9	—	—
62	Sm	1.6	—	—
63	Eu	1.0	—	—
64	Gd	1.1	—	—
66	Dy	1.0	—	—
70	Yb	1.5	—	—
82	Pb	1.8	—	—
83	Bi	< 0.8	—	—
90	Th	< 0.8	—	—
92	U	< 0.6	—	—

beryllium, boron, iron, lead, bismuth, thorium and uranium, more recent and more accurate photospheric abundance results are given here. The differences between photospheric and coronal abundances of certain elements are essentially due to the uncertainties in the abundance determinations, both from the Fraunhofer spectrum and the emission spectrum.

From Table 2 it can be seen that only a minute portion of the solar gas is made up of elements heavier than helium, that about 15 percent is helium, and that the bulk of the solar gas is constituted of hydrogen.

CHROMOSPHERE

The relatively transparent chromospheric layers form a complex transition zone about 5,000 kilometres, or 3,000 miles, thick between the low-temperature photosphere and the high-temperature corona. The chromosphere temperature increases outward from about 4,500° K (about 8,000° F) to 1,000,000° K (1,800,000° F), whereas the density decreases rather rapidly from about 10^{12} particles per cubic centimetre at the base to about 10^{9} particles per cubic centimetre at the top. The chromosphere is a region of great activity, visible at the time of a solar eclipse as a bright-coloured arc from which the "flash spectrum" originates. The low chromosphere is a



Figure 5: High-resolution photograph of sunspot in hydrogen (H-alpha) light shows bright plages and dark filamentary structures (prominences) on the Sun.
By courtesy of Lockheed Solar Observatory, Palo Alto, California

relatively homogeneous layer of gas, about 1,500 kilometres thick, from which emerges, geyser-like, a multitude of upward-moving jets called spicules. At about 8,000 kilometres the spicules stand apart as separate entities and it is thought that the hot coronal gas reaches down between them. The chromospheric gas is quite transparent to continuous radiation but opaque to light in the cores of strong Fraunhofer lines. As a result, if the Sun is viewed in the monochromatic light of a Fraunhofer line, its substance is visible. By selecting lines of different strengths the astronomer can choose the depth of observation; thus, the chromospheric layers can be figuratively peeled like an onion. Through equipment capable of isolating a spectrum line a fraction of an angstrom in width, the Sun presents a picture of fine and coarse structures (Figure 5). Much of the fine structure visible in the cores of metal lines is due to Doppler displacements from ascending and descending volumes of gas. The periodic character of this motion has been discussed in the section on granulation above. Photographs that are taken in the light of one particular wavelength, called the hydrogen-alpha line, show a very coarse mottling, which near the sunspots becomes more or less organized into filamentary whorls.

Flash spectrum. During a total solar eclipse there is a moment when the photosphere is just hidden by the

By courtesy of the Joint eclipse expedition: High Altitude Observatory, Joint Institute for Laboratory Astrophysics, and Air Force Cambridge Research Laboratories

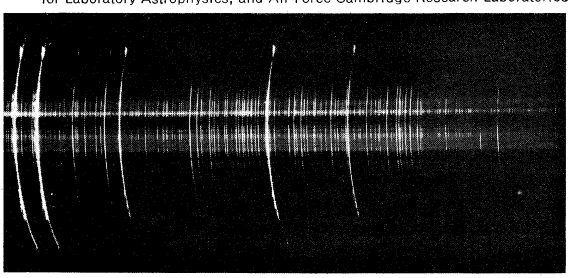


Figure 6: Flash spectrum of the Sun in ultraviolet light. The two bright spectral lines at left, showing prominence structure, are the H and K lines of ionized calcium. The bright line in the centre is a line of ionized titanium at 3685Å. Frequency increases from left to right, wavelength from right to left. Obtained at the solar eclipse of February 1962 in New Guinea.

Moon, leaving exposed the chromospheric layers, which appear as a thin, orange-red crescent along the edge of the Moon. The light from this crescent is found to have a bright-line spectrum, which, for a few seconds, "flashes" in and out of view as the Moon moves across the Sun's atmosphere, hence the name "flash spectrum" (Figure 6). The brightest lines and those which extend to the greatest heights are lines of hydrogen, helium, and calcium; when examined in detail the spectrum is nearly, but not quite, identical with the Fraunhofer spectrum, the differences reflecting the different physical state of the photospheric and chromospheric layers. Lower pressures and higher temperatures favour greater ionization and excitation. Characteristic flash spectra show an enhancement of high-excitation lines and the appearance of a great number of lines from singly ionized iron, titanium, nickel, chromium, and the rare earths.

Spicules. A 19th-century Italian astronomer, Angelo Secchi, described the chromospheric layers viewed at the limb as like a burning prairie. He had seen through his spectroscope a forest of superimposed jets, "the spicules," each lasting only a few minutes. Spicules appear as tiny, bright surges, less than 500 kilometres (300 miles) in diameter, moving upward with velocities of 20–30 kilometres per second to heights of 8,000 kilometres above the photospheric layers. They are often grouped in rosettes or brushlike shapes. They number about 100,000 and have an average lifetime of five to 15 minutes. On the basis of spectroscopic and radio studies, a spicule column is believed to have temperatures of 10,000° K (18,000° F) in the core and 50,000° K on the surface. The denser core material emits lines of calcium and hydrogen and radio waves of millimetre and centimetre wavelengths. The hot sheath is the source of the far-ultraviolet emission of highly ionized atoms and of the lines of neutral and singly ionized helium.

Supergranulation. Spectroheliograms made, for example, with calcium light show a coarse pattern in addition to smaller mottles, which are due to a network of cells about 30,000 kilometres in diameter called supergranules. In these cells an expansion of material takes place from centre to edge with a speed of about ½ kilometre per second. These supergranules, which number about 5,000 at any time, have been identified as deep convection cells with a lifetime comparable to a day. It is observed that the flow of material from centre to edge tends to collect the majority of spicules on the cell boundaries and to sweep up the magnetic field domains so that the cellular pattern is visible on magnetic maps as well as on spectroheliograms.

CORONA

The corona is a luminous, high-temperature, rarefied gas envelope of the Sun. Close to the Sun it is about $\frac{1}{1,000,000}$ as bright as the solar disk. Its brightness decreases rapidly with distance, so that at two solar radii the brightness is less than one percent of that of its innermost parts; hence, the feeble light is visible only at the time of a total eclipse or through a coronograph. The corona has a kinetic temperature of about 2,000,000° K (about 3,600,000° F) out to several solar radii. Because of its high temperature it evaporates continuously, producing an outward flow of electrically charged particles called the solar wind, which extends beyond the vicinity of the Earth. The highly ionized coronal gas has a density of the order of 5×10^8 atoms per cubic centimetre at its base; the density decreases with height, diminishing by a factor of 2.7 approximately every 50,000 kilometres (30,000 miles).

Form and structure. The corona owes its beauty to its form, which is made up of streamers, filaments and rays of pearly white light. Polar plumes, fanning out like iron filings about the poles of a magnet, suggest north and south magnetic poles for the Sun. At sunspot maximum the corona is globular in form with streamers in all directions, while the minimum phase is characterized by polar plumes and long equatorial streamers extending outward many solar radii.

Corona light has three components: the so-called *L*-, *K*-, and *F*-coronas. The *L*-corona, which contributes only about 1 percent of the coronal light, refers to the line spectrum emitted by the highly ionized atoms of the coronal gas, which in such a state is called a plasma. The light of the *K*-corona is caused by scattering of photospheric light by fast-moving free electrons near the Sun. The *K*-component is strongest in the inner corona where the electron density is highest. Since the electrons move at very great speeds (of the order of 6,000 kilometres per second), the photospheric Fraunhofer lines are here completely washed out by the scattering process and, consequently, the *K*-component reveals a continuous spectrum only. The *F*-component is caused by diffracted photospheric light from interplanetary dust particles. Its spectrum is that of the photosphere; *i.e.*, a continuous spectrum intersected by the Fraunhofer lines. The *F*-component is an extension of the zodiacal light into the region close to the Sun. The *L*-component and the *K*-component represent the true corona.

The *F*-, *K*-, and *L*-coronas

The corona may also be studied on radio heliograms, and on X-ray pictures made from above the Earth's atmosphere. The solar radio emission at wavelengths greater than one metre originates in the corona, and at the other end of the electromagnetic spectrum, any wavelength shorter than 200 angstroms can be emitted only by the corona. X-ray pictures show the regions of strongest coronal condensation. Comparison of observations over different parts of the electromagnetic spectrum show that the corona is very inhomogeneous and that its form, structure, and intensity change considerably with the level of activity on the surface of the Sun.

The quiet Sun emits radio waves. At metre wavelengths the emission comes from great heights where the electron density is less than 3×10^7 electrons per cubic centimetre—half a solar radius from the photosphere. At ten metres, the corresponding height is ten solar radii. Thus, the corona can be examined by the radio astronomer, layer by layer, in different radio wavelengths, just as the optical astronomer examines the chromosphere, layer by layer, with the spectroheliograph.

Physical properties. The most remarkable property of the corona is its high kinetic temperature, 1,500,000 to 2,500,000° K. This follows from many types of observation: (1) The coronal spectrum was shown in 1941 to be emitted by atoms of iron, nickel, calcium, and argon that had lost ten to 14 electrons. Such a degree of ionization can come about only by collisional ionization by energetic electrons driven by high temperatures in a low-density gas. (2) The coronal emission lines are much broader than the Fraunhofer lines and if this is interpreted as Doppler broadening by random motion of the emitting atoms, a temperature of about 2,000,000° K (3,600,000° F) is indicated. (3) The brightness of the Sun in radio wavelengths longer than 100 centimetres is about 1,000,000° K (1,800,000° F). (4) Finally, the density gradient of the electron gas has been shown to follow an exponential decrease with height and results in a temperature of 1,500,000° K.

It is generally believed that the energy source supplying the heating of the corona lies in the mass motions of granules and spicules in the photosphere and chromosphere. The interaction can take place by direct collision between the atoms of the spicule and those of the corona or by the transfer by shock waves (hydromagnetic, acoustic) of energy from below to the chromosphere and corona.

The solar wind. It was predicted in 1958 that the high temperature of the corona would cause some of it to boil off the Sun, and that the coronal particles would appear in the neighbourhood of the Earth as a solar wind moving with speeds of 300–1,000 kilometres (about 190–600 miles) per second. Though it had been suggested as early as 1951 that the behavior of certain comet tails could be explained by the interaction of a solar plasma from the Sun with the comet, not until satellites explored the space around the Earth was direct proof forthcoming. A space probe from the United States to Venus in 1962 established the wind's existence and showed that there is a steady stream of protons and electrons in the vicinity of

the Earth with a density of about one to ten particles per cubic centimetre and carrying with them an associated magnetic field ranging from two to 40 microgauss. Thus interplanetary space is filled with a fully ionized hot plasma of coronal origin. At solar minimum the average flow into space amounts to 10^{33} particles per second; this mass loss is thought to be supplied from below by the dissipation of spicules.

A changing pattern of active regions covers the surface of the Sun and proceeding from each is a positive or negative magnetic field, which is carried by the plasma. Because of the Sun's rotation and the steady outflow of material, the magnetic field lines trace curves in space—Archimedean spirals in the equatorial plane.

The active Sun

The active Sun is characterized by the growth and decay of magnetic fields together with all the associated phenomena observed at many different levels of the solar atmosphere: sunspots with associated very bright areas called faculae, plages (other bright areas not necessarily associated with sunspots), flares, surges, prominences, coronal condensations, and the emission of radio, X-ray, and cosmic-ray bursts (see Figure 1 and below).

CENTRES OF ACTIVITY

A centre of activity is an area formed by the arrival from below of a localized strong magnetic field at the Sun's surface. Its description includes all of the transient interrelated phenomena listed above. A few centres of activity may live 200 days or more (eight to ten solar rotations); most are short-lived lasting one to two weeks.

Life history of a typical centre of activity

The life history of a typical centre of activity is given below. In the first few days a bipolar magnetic region (one showing both north and south magnetic polarity) develops, without spots, followed later by bright areas known as faculae and coronal extensions projecting from them. The coronagraph shows stable ray and arch structures about 150,000 kilometres long above the region. After two to three days the most important (spot) phase starts with the formation of a small spot and a rapid extension of the fine lineal elements—small prominences—in the neighbourhood of the spot. A second spot forms shortly afterward, followed by the formation of numerous small spots between the larger bipolar pair. A few flares and flare surges take place. After about ten days the group of spots reaches its maximum size. Each spot appears to be embedded in a large, complex penumbral region of converging filaments. Flare activity reaches its maximum along with the associated phenomena of solar radio bursts, enhanced X-ray emission, and, rarely, cosmic rays from super flares. By 27 days—one solar rotation—the activity has decayed and all spots except the preceding spot have disappeared. Flare activity is greatly reduced, but the bright facular areas (extremely bright areas associated with sunspots) continue to increase in size. Next, a stable prominence filament forms on the poleward side of the region, which, after the lapse of 50 days, cuts the facular region in half, separating regions with opposite magnetic polarity. The prominence filament now extends for 100,000 kilometres (about 60,000 miles) and is oriented more or less east and west, drawn out by the Sun's differential rotation. At about the 80th day, the facular area breaks up, but the filament continues to lengthen. After 100 days the faculae have disappeared completely and the prominence filament reaches its maximum length. In the succeeding days it shortens and drifts poleward where it still may have an identity after 150–250 days.

SUNSPOTS

Physical nature. In appearance a sunspot consists of two parts: a dark core, the umbra, surrounded by a filamentary border called the penumbra. Single spots comprise about 40 percent of the total. Normally, they are in complex groups exhibiting great variety of form and shape, often covered with brilliant "bridges," giving the impression of great turmoil. The smallest spots of a group are mere pores without penumbra, a few hundreds of kilometres in diameter, whereas the largest spots exceed by several times the diameter of the Earth. Large spot groups may extend over one-sixth of the diameter of the Sun and are easily visible to the naked eye at sunset when the Sun's intensity is much reduced. At the Sun's limb (visible edge) a large spot may appear as a slight depression in the Sun's surface.

Sunspots are dark only in contrast to the surrounding photosphere, for, in fact, they are brighter and hotter than molten tungsten. From their observed brightness a spot temperature of 3,800° K (6,800° F) is obtained, as compared to 5,700° K for the photosphere. At spot temperatures, a number of chemical compounds are stable. The spectrum lines of many radicals and molecules, cyanogen, carbon monoxide and hydrides of nitrogen, carbon, oxygen, and magnesium, also known in the disk spectrum, strengthen greatly in the spot, indicating increased abundance; also lines of calcium, silicon and fluorine hydrides, and titanium and zirconium oxides show in the spectrum of a sunspot.

Magnetic property of sunspots

The predominant property of sunspots is their intense magnetic field. In general, the larger the spot, the stronger is the magnetic field. A small spot typically has a field of about 500 gauss (a gauss is a unit of magnetic field strength), a large spot, 4,000 gauss. By comparison, Earth's natural magnetic field at the surface is somewhat less than one gauss, and the overall solar magnetic field is now believed to be about the same. In a simple spot the field is nearly axisymmetric; it has its maximum value at the centre of the umbra and decreases to a very small value at the outer edge of the penumbra. On the average, the lines of magnetic force stand vertical to the Sun's surface at the centre of the umbra but diverge outward so that at the edge of the penumbra the direction of the field has an inclination of about 70° to the vertical. An explanation of the relative coolness of sunspots must involve their associated magnetic fields. After the magnetic field is formed the region cools, presumably because convection, bringing energy from below—as exhibited by the surrounding granulation—is stopped. A balance is achieved when the gas pressure, the magnetic pressure, and the mass flow of material in and out of the spot are in equilibrium.

The sunspot cycle. That spots are not always equally numerous on the surface of the Sun is apparent to anyone who examines the Sun telescopically on every clear day. But it was not until 1851 that astronomers were able to find a regular periodicity of about ten years. At sunspot minimum the Sun may remain spotless for weeks; at maximum several large groups containing dozens of spots are generally visible on its surface. Complete records of the Wolf sunspot numbers (named after the Swiss solar astronomer Rudolf Wolf, whose work on the behaviour of sunspots was fundamental) are known for the past 120 years. Monthly means have been established from 1749 and the epochs of maximum and minimum are known from the year 1610.

Sunspot numbers

Because of the wide fluctuations of the daily number of spots, the level of solar activity is generally expressed as a running mean of a month or a year. The average interval between maximum is 11.2 years but is quite variable, ranging from eight to 16 years. The ascent to maximum, which averages 5.2 years, is steeper than the decline. In addition to the 11-year cycle, a low-amplitude 80-year cycle is recognized, though observations have not yet extended over an interval long enough to define the long wave completely. Sunspot numbers show an 11-year period, but the true length of the cycle is 22 years, when account is taken of the alternation of the sign of spot magnetic fields. The leading spot of a bipolar group in the northern hemisphere has a magnetic polarity opposite that of the leader in the southern hemisphere; and the polarity of the leaders reverse at the beginning of each 11-year cycle. Thus, the magnetic period of the Sun is twice the 11-year cycle.

The number of spots in each cycle is quite variable. Attempts have been made to fit the observed numbers by superposition of cycles, but, in general, these have failed to predict succeeding cycles. Others have tried to relate

the sunspot cycle to planetary influence, particularly that of Jupiter, which revolves around the Sun in a period of 11.8 years. Apparently the true cause of sunspots is deeply seated in the Sun itself.

Each new cycle begins with the appearance of a few spots at about ±30° north and south latitude. Four years later in the cycle, at sunspot maximum, the zone of eruption is at ±15° latitude, while the last few spots of the cycle appear at ±8°. When the latitude of spots, without regard to size, is plotted as a function of time, a so-called butterfly diagram results. This shows that new spots of a new cycle can appear at high latitudes before the disappearance of low-latitude, old-cycle spots, and suggests each new cycle is in some ways independent of the old.

Magnetic field theories

Many of the known features of the sunspot cycle have been explained by a theory of magnetic-field amplification by solar differential rotation. Very much simplified, the theory postulates that the Sun's differential rotation slowly winds up the lines of magnetic force like string on a ball. These lines of force, called flux ropes, which are formed beneath the surface, become buoyant and tend to rise to the surface; thus, bipolar magnetic regions (sunspots) are formed where the flux rope loops through the surface. One worker has separated the field into three components and adopted an interior rotation faster than the exterior with a strong concentration of radial shear toward the equator. With suitable parameters, he was able to reproduce with remarkable exactitude all the phenomena connected with the solar cycle.

OTHER PHENOMENA

Faculae, plages. Bright granulated structures, visible in white light, toward the Sun's limb, are called faculae. That they appear only near the limb and not at the centre of the disk means that the higher facular parts emit more radiation than the surrounding high photospheric layer and the lower parts emit less than the local photosphere. The temperature excess amounts to 250° C (480° F) at the most.

Spectroheliograms of the active Sun reveal large, bright areas closely associated with the photospheric faculae. These chromospheric faculae, or plages, are the extension of the photospheric faculae into the chromosphere. Ultraviolet plages and X-ray plages are observed coming from the high chromosphere and corona. Radio plages are also observed in the centimetre and decimetre radio wavelengths.

Prominences, filaments. Prominences are among the most beautiful of solar phenomena (see Figure 2). A typical fully developed quiescent prominence is a thin fish-tailed, or ribbon-like, gaseous structure standing above the Sun's surface; it may attain a size of 200,000 kilometres (125,000 miles) long, 40,000 kilometres high, and 6,000 kilometres thick. As seen on the disk, in hydrogen-alpha spectroheliograms, prominences appear as long, dark structures called filaments. At the Sun's limb projected against the dark sky they are seen as bright clouds composed of a myriad of fine threads. Generally, prominences are connected by appendages to the Sun's surface or rest upon it, but a few stand free. Like clouds, prominences take many fanciful forms. From their appearance and mode of formation they have been classified into the following types: active prominences, eruptive, sunspot, tornado, quiescent, and coronal prominences.

Time-lapse motion pictures of solar prominences

Time-lapse motion pictures reveal the intricate motions of prominences. Thus viewed, they show at least two active groups of the first type: prominences that appear as great funnels with downward-curving streamers feeding knots of material from the prominence into a nearby active area; and the spectacular arched prominences in which material condenses from the corona and moves down along each arm, often for several days, into a spot region. Quiescent hedgerow-type prominences are carried in, and are supported against the pull of gravity by, a trough of magnetic field lines; they separate regions of opposite magnetic polarity on the Sun's surface and may last for months. The large-scale structure of such quiescent prominences is fairly constant, but the fine structure shows much motion, generally downward. On occasion, a nearby solar flare may trigger great activity in the prominence such that it either flows downward into an active region and disappears, or that a portion of it may erupt from the surface of the Sun with velocities of 200–500 kilometres, or about 125–310 miles, per second. Strangely, the activity appears not to destroy the magnetic-field configuration, for there is a tendency for the prominence to reappear a few days later, phoenix-like, in its former place and form.

Spectroscopy reveals the physical conditions in prominences. There are both hot and cold prominences. Cold prominences with kinetic temperatures of 8,000° to 10,000° K (14,000°–18,000° F) show sharp spectral lines: the Balmer lines of hydrogen, H and K of calcium, and a number of lines of ionized and neutral metals. Active prominences, loops and surges, with temperatures up to 100,000° K (180,000° F) show lines broadened by thermal motions of the atoms; the metal lines are weak and ionized helium lines appear.

Flares. In plage areas, sudden and distinct brightenings occur from time to time. These brightenings, called flares, last from a few minutes for small ones, up to several hours for very large ones. Solar flares are among the most violent and dramatic events that occur on the Sun, yet, with few exceptions, they cannot be seen in white light and may well go unnoticed. They are easily seen, however, in the red light of hydrogen and they produce a number of effects observable at the Earth that make them of considerable importance.

Flares are classified by area and brightness on a scale of increasing importance, from — 1 to + 3, the latter group comprising only 1 percent of the total. As seen in hydrogen light, a typical large flare starts as a brightening of a pre-existing filamentary structure in the neighbourhood of a sunspot. The rise to maximum brightness is sudden, taking place in two to five minutes, with a rapid spreading of the flare area, the brightest parts of which are often elongated. The flare lasts about one hour longer during which a bifurcation and separation of the luminous ribbon and subsequent fading of the flare take place. Frequently, after the onset of a flare, a surge of gas is ejected with a velocity of perhaps 100 kilometres (about 60 miles) per second. A great flare can create an expanding magnetohydrodynamic wave, a complex effect involving both magnetic and hydrodynamic forces, propagating through the corona. Its passage is seen by the successive activation of disk filaments as it spreads over the surface with velocities of 1,000 kilometres (600 miles) per second and for distances of over 500,000 kilometres.

Geophysical phenomena

Many geophysical phenomena are caused by radiation and particle emission from flares. For example, simultaneously with the observed flare there is often a fade-out of shortwave-radio communication caused by a flash X-ray penetrating to the ionosphere of the Earth and greatly increasing its ionization and absorption. Plasma clouds from the flare region moving out at 2,000 kilometres (1,250 miles) per second produce auroras and geomagnetic storms a day or so later. Cosmic-ray particles from very large flares have been observed to penetrate to ground level. In the polar regions, measurements of cosmic-radio noise show a marked decrease, caused by bombardment of the upper atmosphere by high-energy protons from the flare.

Observations of the X-ray spectrum during a flare detect a great increase in the emission at the shorter X-ray wavelengths down to below ten angstroms. This reveals the existence of localized regions with temperatures of the order of 25,000,000° K and more. The formation of such regions is explained as being due to thermal electrons heated by accelerated electrons, which, in turn, are responsible for the microwave emission connected with flares. Usually the enhanced X-ray emission caused by the flare phenomenon appears after the emission at radio frequencies, but the X-ray emission lasts much longer and decays more slowly than the microwave emissions.

In addition to the above-mentioned flare-associated X-ray emission (so-called quasi-thermal due to the heating of localized regions), bursts of very short wavelength

X-ray and radio phenomena

(hard) nonthermal X-rays may occur which are observed in the range between about one and ten angstroms. They may last from a couple of seconds up to two minutes. Since the energies of these bursts are of the order of 10^4–10^6 electron volts, some of their electrons must be moving with velocities of between 60,000 to about 300,000 kilometres or, about 37,000 to 186,000 miles, per second. Consequently, the bursts are felt to be due to bremsstrahlung radiation (that emitted by decelerating electrons) or to synchrotron radiation, produced by electrons accelerating in a curved path. The hard X-ray bursts are correlated with various types of radio bursts which also are produced by the acceleration of particles to high speeds. There is, however, not always a one-to-one correspondence between X-ray and radio emission in a flare or any other flare phenomenon.

The interaction of high-energy electrons, protons, and nuclei in the solar atmosphere may also produce gamma rays in a flare region. A few major flares have been observed to eject gamma-ray bursts with energies up to one million electron volts (*i.e.*, the energy that would be gained by an electron in moving through a potential energy difference of 1,000,000 volts). Like X-ray bursts, gamma-ray bursts are short-lived phenomena; their duration, being of the order of a few seconds to a minute, depends on the time the high-energy particles are trapped in the solar plasma.

The radio phenomena associated with a large flare are numerous and complex; different events appear at different frequencies and different times. During the flash phase of a large flare, a stream of high-energy electrons moving with one-third of the speed of light are ejected from the flare into the corona to generate waves moving at speeds faster than the speed of light in the plasma called Cherenkov plasma waves and fast-drift radio bursts, that is, bursts of radiation at rapidly changing radio wavelengths. These are observed in the frequency range 600 to 20 megacycles per second. In five to ten minutes other fast electrons are trapped in the magnetic field and their oscillation, possibly between magnetic-mirror points, produces a broadband burst over areas of about one-quarter of the Sun's disk. Electrons accelerated downward from the ejection area of the flare plunge into the chromosphere creating X-rays and microwave bursts. In addition to the high-speed electrons, the flare explosively ejects a dense plasma cloud at a velocity of 1,000 kilometres (600 miles) per second. This cloud may interact with the corona at supersonic velocities to form shock waves and radio bursts. The magnetic field embedded in this plasma moving out from the Sun holds electrons moving at a high fraction of the speed of light, which emit synchrotron radiation that can be followed with radio telescopes as a moving burst. From deep in the flare site microwave emissions continue, generated by synchrotron and bremsstrahlung processes (see above). Continuous radio emission continues for hours at metre wavelengths.

Energy release

The physical causes of flares are still unknown, though it is believed that the energy released by a flare (about 10^{32} ergs in a large flare) must come from the intense magnetic or electric fields associated with the solar active regions. It has been estimated that the energy released by a single flare would be enough to provide electric power for the whole world for 1,000,000 to 100,000,000 years at the current rate of use, depending upon the size of the flare.

Coronal condensations. The corona is far from uniform. Coronal condensations occur over centres of activity. They dominate the scene in X-ray photographs of the Sun and can be observed with radio telescopes. The condensations are shaped in the forms of rays, brushes, loops, or arches, and are often associated with hot prominences. They vary rapidly in brightness (a factor of 2 in two hours is not uncommon) but may remain stable in form for quite some time; in fact, over a large, highly developed spot group or centre of activity, a condensation may remain visible for several months. At the time of a flare the structure of a condensation may suddenly undergo change and show great activity; arch streamers, for example, may expand outward with velocities of 500 kilometres (about 310 miles) per second. In a coronal condensation, the electron temperature approaches 5,000,000° K and the particle density is five to ten times greater than that of the surrounding quiet corona. At radio frequencies coronal condensations manifest themselves by enhancement of the radio emission in centimetre and decimetre wavelengths. By far, the greatest enhancement, about 80 times greater than in the quiet corona, can be observed in the X-ray region, in which the emission is proportional to the square of the electron density.

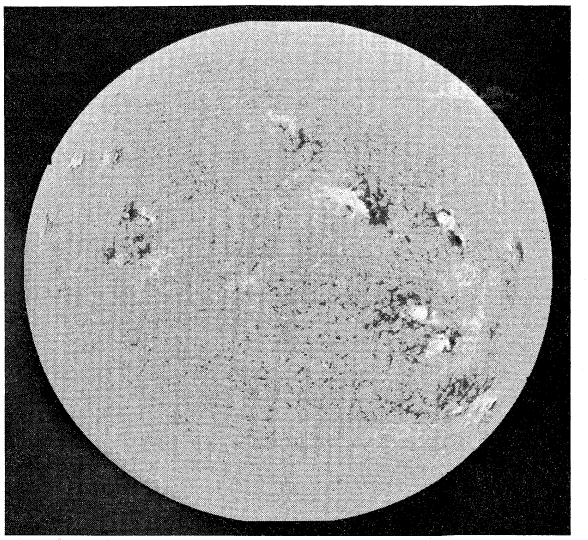


Figure 7: Photospheric magnetic fields on the Sun. The dark and light areas correspond to positive and negative (north and south) polarities. The bipolar nature of most regions is indicated. The figure shows a computer-generated photograph or map built up of direct photoelectric scans covering the entire visible surface.
By courtesy of Kitt Peak National Observatory, Tucson, Arizona

The Sun's general magnetic field. The existence of a weak polar magnetic field for the Sun has long been assumed. The measurement of small magnetic fields by the Zeeman effect is exceedingly difficult; hence, it need not be surprising that, although early measurements of polar photospheric fields gave values of 20–50 gauss, modern measurements have indicated much smaller values, of the order of one gauss, roughly the same as the natural magnetic field at the surface of the Earth. How the observed photospheric field is related to the internal field is not known. The field observed in the photosphere in the polar regions is not the dipole field but results from the slow migration and expansion of magnetic remanents from centres of activity in the sunspot zones to the polar regions. Furthermore, the irregular or random eruption of magnetic flux in the spot zones is eventually reflected in the polar fields and explains why the Sun has, on occasion, had a magnetic south pole simultaneously at both north and south poles and why there is often a lack of perfect symmetry between the two hemispheres. Figure 7 shows a continuous plot of the photospheric magnetic field.

The solar interior: energy generation and evolution of the Sun

The internal structure of the Sun is determined by its chemical composition, by its mass, and by its previous history. Today, with the development of large, fast computers, theoretical solutions to equations governing the structure of the Sun can be found in a few minutes and the evolutionary track showing the changes in the structure and compositions followed for billions of years.

The development of a mathematical model of the Sun would require that the chemical composition be known throughout its mass, and this is not known. Hence, a sequence of models must be computed and compared to the true Sun. It has been suggested that the flux of neutrinos from the hydrogen-burning fusion reactions in the

Sun's interior (see below) may allow direct observations of conditions in the Sun's core, and investigators have set up detectors to capture the neutrinos. The results so far indicate that the central temperature of the Sun is less than 20,000,000° K, but results have not yet fixed the chemical composition.

The Sun radiates 3.86×10^{33} ergs of energy every second (from energy = mass times velocity of light squared, $e = mc^2$; this corresponds to a mass loss of 4,700,000 tons per second or 1×10^{-11} solar mass per year). Nuclear reactions are the only possible means that can produce such a great amount of energy. In the hot, dense core of the Sun, within about one-tenth of a solar radius of the centre, nuclear fusions liberate the energy that is emitted millions of years later at the surface of the Sun. The most important fusion process in the Sun is the so-called proton–proton reaction or hydrogen-burning. The fusion of two hydrogen atoms, the nuclei of which are called protons, produces a deuterium atom and releases a positron and a neutrino.

Nuclear reactions

The neutrino escapes with the velocity of light and the positron quickly collides with an electron, whereby they annihilate each other, producing a very high-energy photon of radiation, called a gamma ray. The collision of the deuterium atom with another proton produces an atom of the light helium isotope of atomic mass 3, called helium-3, and the emission of another gamma ray. Eventually, two helium-3 atoms collide and in doing so, produce an ordinary stable helium atom of atomic mass 4 and release two protons and a gamma ray. In total, two electrons and six protons are used up in the production of one stable helium atom; two new protons are rebuilt and two neutrinos and five gamma rays are released in the process, which provide the energy radiated away at the solar surface. About 5 percent of the energy is carried away by the neutrinos (see NUCLEAR FUSION).

The Sun is presently in a state of equilibrium with a balance between its total energy production and its luminosity. The evolution of the Sun has been considered by many theoreticians. The U.S. astrophysicist Icko Iben sees a million years as having been taken for the mass of gas that is now the Sun to evolve from a wholly convective state to a body with a radiative core. The radiative core then grows to encompass nearly all the Sun's mass. After 10^7 years the brightness reaches a minimum and then increases. At 2.3×10^7 years nuclear burning begins. At 2.6×10^7 years a convective core appears and the Sun begins to take on a form much like its present one at 3.7×10^7 years. As the conversion of carbon to nitrogen proceeds to burn carbon rapidly, the convective core is replaced by the condition of radiative equilibrium, in which the amount of energy received is everywhere exactly passed on throughout the mass. In 4.5×10^9 years the brightness increased to the present solar value. At its present stage the Sun consists of about 60 percent hydrogen by weight. After about 6×10^9 years a slight change in mass and luminosity may be recognizable. During the past 4.5×10^9 years, that is, since the formation of the Earth's crust, the Sun has virtually remained unchanged.

Future evolution

The future evolution of the Sun is expected to be similar to that of other normal stars. Eventually all hydrogen will be burned up and nuclear reactions involving helium and heavier atoms will take over. This will change the chemical composition of the Sun; as a result, the Sun will increase in size and luminosity and thus turn into a red giant star. Computations on evolutionary models predict that in a few times 10^9 years the Sun will reach the red giant stage. Finally, when all nuclear energy sources are used up, the Sun will reach its last evolutionary stage, that is, it will become a white dwarf, a star of small radius. Its radius will be about 100 times smaller than it is now. Its internal temperature and its luminosity will gradually decrease on a time scale of 10^9 years and, eventually, the Sun will become a black dwarf, a very dense, nonluminous object of degenerate matter. The total lifetime of the Sun is estimated to be some 10^{10} years.

BIBLIOGRAPHY. A large fraction of modern solar research is published in *Solar Physics* (1967–), a monthly journal for solar research and the study of solar–terrestrial relations.

Descriptive material and physics of the Sun's atmosphere: Popular books on this subject include: D.H. MENZEL, *Our Sun*, rev. ed. (1959); G. ABETTI, *The Sun*, trans. by J.B. SIDGWICK (1957); G. GAMOW, *The Birth and Death of the Sun* (1945); K.O. KIEPENHEUER, *The Sun*, trans. by A.J. POMERANS (1959); and S.A. MITCHELL, *Eclipses of the Sun*, 5th ed. (1951, reprinted 1969). Works of a more technical nature include: H. ZIRIN, *The Solar Atmosphere* (1966); G.P. KUIPER (ed.), *The Sun* (1953); S. FLUGGE (ed.), *Encyclopedia of Physics*, vol. 52, *Astrophysics III: The Solar System* (1959); A. UNSOLD, *Physik der Sternatmosphären*, 2nd ed. (1955); and E. TANDBERG-HANSSEN, *Solar Activity* (1967).

Solar subjects: Specific subjects are treated in the following technical books: R.J. BRAY and R.E. LOUGHHEAD, *Sunspots* (1964), *The Solar Granulation* (1967); W. GLEISSBERG, *Die Häufigkeit der Sonnenflecken* (1952); M. WALDMEIER, *The Sunspot-Activity in the Years 1610–1960* (1961); H.J. and E.V.P. SMITH, *Solar Flares* (1963); Y. OHMAN (ed.), *Mass Motions in Solar Flares and Related Phenomena* (1968); N. ROBINSON (ed.), *Solar Radiation* (1966); D.E. BILLINGS, *A Guide to the Solar Corona* (1966); J.W. EVANS (ed.), *The Solar Corona* (1963); R.N. THOMAS and R.G. ATHAY, *Physics of the Solar Chromosphere* (1961); A.C. STRICKLAND (ed.), *Annals of the IQSY*, vol. 4, *Solar-Terrestrial Physics: Solar Aspects* (1969); T.R. OPPOLZER, *Canon of Eclipses*, trans. by O. GINGERICH (1962); J. MEEUS, C.C. GROSJEAN, and W. VANDERLEEN, *Canon of Solar Eclipses* (1966); K.O. KIEPENHEUER (ed.), *Structure and Development of Solar Active Regions* (1968); M.R. KUNDU, *Solar Radio Astronomy* (1965); A.G. SMITH, *Radio Exploration of the Sun* (1967); S. CHAPMAN, *Solar Plasma Geomagnetism and Aurora* (1964); R.v.d.R. WOOLLEY and D.W.N. STIBBS, *The Outer Layers of a Star* (1953); V.V. SOBOLEV, *Treatise on Radiative Transfer*, trans. by S.I. GAPOSCHKIN (1963); M. SCHWARZSCHILD, *Structure and Evolution of the Stars* (1958); R.F. STEIN and A.G.W. CAMERON (eds.), *Stellar Evolution* (1966); and D.H. MENZEL, P.L. BHATNAGAR, and H.K. SEN, *Stellar Interiors* (1963).

Atlases of Fraunhofer lines: G. BRUCKNER, *Photometric Atlas of the Near Ultraviolet Solar Spectrum 2988 Å–3629 Å* (1960); M. MINNAERT, G.F.W. MULDERS, and J. HOUTGAST, *Photometric Atlas of the Solar Spectrum—Utrecht* (1940); O.C. MOHLER *et al.*, *Photometric Atlas of the Near Infra-Red Solar Spectrum λ 8465 to λ 25242* (1950); L. DELBOUILLE and G. ROLAND, *Photometric Atlas of the Solar Spectrum from λ 7498 to λ 12016* (1963); and M. MIGEOTTE, L. NEVEN, and J. SWENSSON, *The Solar Spectrum from 2.8 to 23.7 Microns* (1956).

Lists of Fraunhofer lines: C.E. MOORE *et al.*, *The Solar Spectrum 2935 Å to 8770 Å* (1966); and O.C. MOHLER, *A Table of Solar Spectrum Wavelengths, 11984Å to 25578Å* (1955).

(A.K.P./E.A.M.)

Sung T'ai Tsu

The Sung dynasty, which ruled China from 960 to 1279, was founded by Sung T'ai Tsu (in Pin-yin romanization, Song Tai Cu), an outstanding military leader, statesman, and emperor, who laid the foundations for the essential political institutions of a remarkable epoch. The political order of his dynasty combined to a surpassing degree freedom of discussion, innovation in bureaucratic methods, internal reform, peace, and stability. This atmosphere undoubtedly facilitated the pioneering in economic techniques, scientific advances, and achievements in philosophy, art, and literature that distinguished the Sung period.

Chao K'uang-yin (who posthumously received the dynastic temple name of T'ai Tsu, or Grand Ancestor) was born in 927 in a military base at the former capital, Lo-yang. He was the second son of a military officer, Chao Hung-yin. China was then in chaos. The once great T'ang dynasty, fragmented by rebellions, had been extinguished in 907. A succession of warlords—Chinese, part Chinese, or foreign—claimed the Imperial throne briefly in devastated North China, while the more prosperous south was divided among satraps who were independent in fact and sometimes in name. The boy's forbears had in three previous generations won a certain standing as military leaders under one or another Imperial claimant, and his father reached a post of high command before his death in 956. The wisdom and foresight of Chao K'uang-yin's remarkable mother influenced his decisions even after her death in 961.

Rise to power

At about age 20 Chao joined a leader whose adoptive father soon afterward established the Later Chou dynasty at K'ai-feng; Chao's patron succeeded to the throne in 954 and fought to extend his sway into South China and to eliminate a rival who, established to the north in Shansi and supported by the Khitan empire, laid claim to the rule of China. Through a series of daring and successful actions Chao quickly rose to the chief command of the Later Chou forces.

Early in 960 Chao's patron died and was followed on the throne by his son, a child. At this time the armies of the Khitan and their Chinese allies prepared a concerted invasion. Chao marched northward to meet them. Discontent arose among Chao's troops, who in the crisis did not wish a child as ruler. When the army was encamped for the night at a bridge outside the capital, the officers awakened Chao (who had drunk well before retiring), hailed him as emperor, robed him in Imperial yellow, mounted him on horseback, and urged him to return and take over the government; the records, which are unverifiable on this point, imply that Chao lacked forewarning of the coup. On the officers' pledge of obedience and promise not to molest the existing Imperial family and its councillors, the public buildings, and the homes of the people, Chao complied. He named his dynasty the Sung.

By courtesy of the National Palace Museum, Taipei, Taiwan



Sung T'ai Tsu, portrait by an unknown artist. In the National Palace Museum, Taipei.

In rapid succession the dissident Chinese states now came under the new emperor's control. In 963 two submitted voluntarily and two more were conquered. By 976 all but the northern rival house were mastered, and its Khitan backers had seen more than one defeat. In that year the Sung armies were mobilized against this last rival when the emperor T'ai Tsu died at the age of 49. Within three years his younger brother T'ai Tsung, who succeeded him, completed the unification of China (except for a small area near Peking that remained in Khitan hands).

The task of unification had not been easy, and in parts of China revolts of local autonomists further complicated it. Yet the Sung founder had also turned his mind to ways of avoiding the dangers that had been fatal to the T'ang and to encourage the success of his dynasty. T'ai Tsu's policies no doubt owed much to his personality, a striking combination of qualities that inspired in his generation and later a multitude of anecdotes about him. Though some of these may include fictional elements, they convey the impression he made on his countrymen. A highly skilled archer and horseman in his youth, Chao survived daredevil equestrian exploits unscathed. As emperor he said that destiny had given him the throne and would determine his life or death; man could not deflect it. Despite remonstrances of his advisers, he persisted in going about incognito to observe conditions among the people. He rejected indignantly the gift of a sword stick for protection in emergency. His tastes were simple; when shown the inlaid urinals captured from the former Szechwan princeling, he had them destroyed. He visited his ministers informally and frankly admitted to them his chagrin over his own errors. In his last year he declined the title of unifier and pacifier that was offered him.

T'ai Tsu's policies and personality

T'ai Tsu was strict in holding his officials to account in important matters; his councillors held him in awe. On the other hand, he accepted minor faults or impertinences with a laugh. He was slow to entertain suspicion. He sometimes acted impetuously, and some have suggested, on rather limited evidence, that he indulged in wine to excess. On occasion, when severely provoked by a presumptuous official or subject, he was given to outbursts of violent rage. At such times, however, his temper cooled quickly, and he then softened penalties he had given in anger and even compensated the unfortunate culprit for abuses suffered.

T'ai Tsu was active by nature. Even as emperor he conducted military campaigns personally from time to time. Rather than simply approving governmental papers in finished form, as T'ang emperors had done, he let his ministers submit rough drafts to him for preliminary criticism.

The functional arrangements of T'ai Tsu's government reflected both his active disposition and his refusal to pretend infallibility. He continued the existing system by which three ministers were directly responsible to him for different aspects of administration (fiscal, military, and general), thus limiting the power of each. By generously granting them consideration and responsibility, however, he encouraged a certain balance between the functions of ruler and minister. The control of the central government over the local was also strengthened. Beginning in 963, the administration of the prefectures was cautiously but steadily transferred from the unruly military to civil officials. Court officials were sent to govern subprefectures. From 965 taxes were remitted directly to the national treasury. The first fiscal intendants—forerunners of the Sung "circuit" system—were established to supervise local functionaries. To counter the military threat to the state's integrity, T'ai Tsu transferred the best troops to the capital and on suitable opportunities induced the most powerful commanders to accept retirement.

Reform of the examination

T'ai Tsu's policies were clearly directed toward the creation of a bureaucracy based on demonstrated abilities rather than birth or favour. This is evident in his steps to strengthen the examination system. By 963 he had forbidden court officials to recommend candidates and had forbidden graduates to consider examiners their patrons. He ordered re-examinations on the petition of a rejected candidate or on even a hint of favour in the selection of graduates. By 973 he had established the final examination in the Imperial palace to verify the rankings and had ordered the list of successful competitors to be announced publicly. He began to award larger numbers of degrees.

A mildness and humanitarian tone pervade T'ai Tsu's policies, which on the whole conform to the Confucian ethos. He extended clemency toward defeated opponents rather consistently. He showed concern for the adherents of the dynasty he displaced. His generals were repeatedly admonished to shun avoidable harm to the citizens of places they occupied and even to spare captured soldiers and leaders; among the latter was the poet-prince Li Yü. His own ministers who lost his favour were treated well. In his legal reforms, though he dealt more severely with corruption and irresponsibility of officials in several measures, he lightened the punishments for violations of the state controls over salt and wine and required a review of all capital sentences by the high court at the capital. His early measures also show a special concern with improving the economic lot of the poorer citizens and easing their burdens of taxation.

When T'ai Tsu died, the construction of the new state was far from finished. The ensuing peace and prosperity would also bring new problems calling for new solutions.

Of the succeeding emperors, none quite matched him in stature or in character. But Confucian ancestral piety reinforced the attraction of his proven policies. The traditions of his active concern in administration and of close association with the bureaucracy's leaders persisted in greater or lesser degree among later Sung rulers. Subsequent developments on the whole moved in directions indicated by T'ai Tsu. Their benefits were scarcely unadulterated; safeguards against the ambitions of military commanders, for example, perhaps hampered Sung armies in meeting powerful foreign invaders. Still, the efforts of his successors to further popular welfare, to find and train the best talent for the civil service, and to defend the state's stability and the unbroken rule of the dynasty for three centuries no doubt owe much to T'ai Tsu's concepts of statecraft. (E.A.K.)

Sun Yat-sen

Sun Yat-sen (in Pin-yin romanization, Sun I-xian, also known as Sun Wen or Sun Chung-shan), early leader of China's Kuomintang (Nationalist Party), played a leading role in the overthrow of the Ch'ing (Manchu) dynasty in 1911 and was the first provisional president of the Republic of China. He is revered by the Communists and is called by the Nationalists the "Father of China."

Early life and influences. Sun was born on November 12, 1866, to a family of poor farmers in Hsiang-shan, in the South Chinese province of Kwangtung. In 1879 his brother Sun Mei, who had earlier emigrated to Hawaii as a labourer, brought him to Honolulu, where, as a student at a British missionary school for three years and at an American school, Oahu College, for another year, he first came into contact with Western influences. Because his brother objected to his penchant for Christianity, Sun returned to his native village in 1883 and went to study at the Diocesan Home in Hong Kong in the fall of the next year. In 1885 he transferred to the Government Central School (later known as Queen's College) and married Lu Mu-chen (1867–1952), the girl chosen for him by his parents. Out of this marriage a son and two daughters were born. In either 1884 or 1885 Sun was baptized by an American missionary. After another trip to Hawaii, he enrolled in the Canton Hospital Medical School in 1886. He transferred later to the College of Medicine for Chinese in Hong Kong and was graduated in 1892.

Education and marriage

Although not trained for a political career in the traditional style, Sun was nevertheless ambitious and was troubled by the way China, which had clung to its traditional ways under the conservative Ch'ing dynasty, suffered humiliation at the hands of more technologically advanced nations. Forsaking his medical practice in Canton, he went north in 1894 to seek political fortunes. In a long letter to Li Hung-chang, governor general of Chihli, he set forth his ideas of how China could gain strength; but all he received from Li was a perfunctory endorsement of his scheme for an agricultural–sericultural association. With this scant reference, Sun went to Hawaii in October 1894 and founded an organization called the Revive China Society (Hsing Chung hui), which became the forerunner of the secret revolutionary groups Sun later headed. As far as it can be determined, the membership was drawn entirely from natives of Kwang-tung and from lower social classes, such as clerks, peasants, and artisans.

Brown Brothers



Sun Yat-sen.

Years in exile. Taking advantage of China's defeat in the Sino-Japanese War (1894–95) and the ensuing crisis, Sun went to Hong Kong in 1895 and plotted for an uprising in Canton, the capital of his native province. When the scheme failed, he began a 16-year exile abroad. In 1896 he was detained for 13 days by the Chinese legation in London. How the incident began is subject to dispute, but it appears likely that Sun ran into a fellow Cantonese who worked for the legation and was found out and caught while visiting him under an alias. The legation planned to ship Sun back to China, but, before this could be done, Sun had converted a British employee at the legation to his side and got a word through to James Cantlie, former dean of Hong Kong College of Medicine. The British Foreign Office intervened, and Sun was released from his captivity. The incident engendered great publicity and gave Sun's career a powerful boost.

Detention in Chinese legation

After spending much of the ensuing eight months reading in the British Museum, Sun travelled to Japan by way of Canada. Arriving in July 1897, he was met by Miyazaki Torazō, an adventurer who had heard of the London incident and who was willing to help Sun in his political activities. Miyazaki introduced Sun to many influential Japanese, including Ōkuma Shigenobu, Soejima Taneomi, and Inukai Tsuyoshi, from some of whom Sun was to receive both political and financial assistance. During the turmoil of 1900, Sun participated in secret manoeuvres involving Sir Henry Blake, the British governor of Hong Kong, and Ho Kai, an influential Chinese in that colony. Their aim was to persuade Li Hung-chang to declare independence from the Ch'ing. Responding to an invitation by Li's staff, Sun journeyed to Hong Kong, but, fearing a trap, he did not go ashore. Instead, he was represented by Miyazaki and two other Japanese at the meeting, which proved fruitless.

Previously, Sun had made contact with bandits and secret societies in Kwangtung. These forces began a revolt in Hui-chou in October 1900. The campaign, the second of ten claimed by Sun between 1895 and 1911, lasted 12 days.

The year 1903 marked a significant turning point in Sun's career; from then on, his following came increasingly from the educated class, the most prestigious and influential group in China. For this decisive change Sun owed much to two factors, the steady decline of the Ch'ing dynasty and the powerful propaganda of Liang Ch'i-ch'ao, a reformist who fled to Japan in 1898, founded a Chinese press, and turned it into an instant success. Liang did not actually oppose the Ch'ing régime, but his attacks on the Dowager Empress, who effectively ruled the country, served to undermine the regime and make revolution the only logical choice. As a consequence, Sun's stock rose steadily among the Chinese students abroad. In 1904 he was able to establish several revolutionary cells in Europe, and in 1905 he became head of a revolutionary coalition, the Alliance Society (T'ung-meng hui), in Tokyo. For the next three years the society propagandized effectively through its mouthpiece, "People's Journal" (*Min pao*).

Founding of the Alliance Society

The rise in Sun's fortune increased many of his difficulties. The Alliance Society was very loosely organized, and Sun had no control over the individual members. Worse still, all the revolts Sun and the others organized ended in failure. The members fell into despair, and outside financial contributions declined. Furthermore, as a result of pressures exercised by the Ch'ing, foreign governments increasingly shunned Sun. In 1907 the Japanese government gave him a sum of money and asked him to

leave the country. A year later French Indochina, where Sun had hatched several plots, banned him completely. Hong Kong and several other territories were similarly out of his reach. In the circumstances, Sun spent a year in 1909–10 touring Europe and the United States. Returning to Asia in June 1910, he left for the West again in December after a meeting with other revolutionaries, in which they decided to make a massive effort to capture Canton. This time Sun raised more money in Canada and the United States, but the uprising of April 27 in Canton (known as the March 29 Revolution, because of its date in the Chinese calendar) fared no better than the earlier plots. The possibility of revolutionary success seemed more remote than ever.

But help was to come from the Ch'ing. If only for the purpose of self-preservation, the court had sponsored reform since 1901. In the next few years it reorganized the army, instituted a school system, abolished the civil-service examinations based on traditional Chinese scholarship, reconstructed many government organs, and convened provincial and national assemblies. The educated class nevertheless remained unsatisfied with the tempo of change, and the regime was rapidly losing its grip over the situation. In 1911 it decided to nationalize all the trunk railways, thus incurring the wrath of local vested interests. Armed rebellion broke out in the province of Szechwan and the court exposed itself to further attacks by failing to suppress it. In October of the same year a local revolutionary group in Wu-han, one of many in China by this time, began another rebellion, which, in spite of its lack of coordination, unexpectedly managed to overthrow the provincial government. Its success inspired other provincial secessions.

The revolution of 1911

Sun Yat-sen learned of the Wu-han revolution from the newspapers while he was in Denver, Colorado. He returned to Shanghai in December and was elected provisional president by delegates meeting in Nanking. Knowing that his regime was weak, Sun made a deal with Yüan Shih-k'ai, an Imperial minister who had been entrusted with full power by the court. On February 12, 1912, the Emperor abdicated; the next day Sun resigned, and on the 14th Yüan was elected his successor.

Later struggles. In September, Yüan appointed Sun director general of railway development. Their entente might have lasted if Sung Chiao-jen, who had reorganized the Alliance Society into the Kuomintang and was serving as its head, had not been assassinated in March 1913, reportedly at Yüan's instigation. This precipitated a second revolution, in which Sun opposed Yüan. When the campaign failed, Sun fled once again to Japan. While there, he unavailingly sought Japanese aid by promising vast concessions in China, and he also alienated many revolutionaries by requiring them to take an oath of personal allegiance to him. He was also criticized for marrying his secretary, Soong Ch'ing-ling (November 1914), without divorcing his first wife.

A combination of internal opposition and external pressures defeated Yüan in 1916. The next year Sun went from Shanghai to Kwangtung to launch a movement against the premier, Tuan Ch'i-jui. Elected generalissimo of a separatist regime in July, Sun had to resign and leave for Shanghai toward the middle of 1918, when he lost the support of Lu Jung-t'ing, the military overlord of Kwangtung.

Earlier, Lu had agreed to Sun's gaining control over 20 battalions of armed guards if the forces would remain outside Kwangtung. Accepting this condition, Sun appointed Ch'en Chiung-ming as the commander and dispatched his men to Fukien. By persuading Ch'en to fight Lu, Sun found his way back to office for another 16 months, at the end of which Ch'en turned against him, and Sun had to leave for Shanghai again. From that sanctuary, he wooed the troops from Kwangsi and Yunnan, and with their help he again returned to Canton. In February 1923 he installed himself as generalissimo of a new regime.

Meanwhile, a new factor had risen in Sun's political life. Unsuccessful at obtaining aid from the West and Japan, he had looked increasingly to the Soviet government, which came to power in Russia in 1917. A Soviet diplomat, Adolf Joffe, visited Sun in Shanghai in both 1922 and 1923. On the latter occasion the two issued a joint manifesto declaring that the Communist system was not suitable for China, that Russia intended to give up its privileges there, and that Russia had no intention of extending its influence over Outer Mongolia. At Soviet prodding, the Chinese Communist Party resolved to cooperate with the Kuomintang. In October 1923, Mikhail Borodin, a representative of the Comintern (Communist International), arrived at Canton and soon gained Sun's confidence. Early in 1924 Sun reorganized the Kuomintang as a tightly disciplined body with authority descending from the top to the lower levels on the model of the Soviet Communist Party. Under his directive a party congress elected three Communists to its central executive committee and approved the establishment of a military academy (of which Sun appointed Chiang Kai-shek president). Part of his party-building efforts were a series of lectures Sun delivered on his own doctrine, the Three Principles of the People.

Ties with the Soviets

Although these actions strengthened the Kuomintang, there was still considerable opposition to Sun's authority when he died of cancer in Peking on March 12, 1925. His coffin remained uninterred in a temple in Western Hills (Hsi-Shan) until 1929, when it was moved to a mausoleum in Nanking.

Assessment. Sun's political doctrines are summarized in his Three Principles of the People (nationalism, democracy, and people's livelihood—the last involving the regulation of private capital and "equalizing land rights") and his Plan for National Reconstruction (*Chien-kuo fang-lueh*), which explained basic parliamentary procedures, attacked the traditional Chinese saying that to know is easier than to do and set forth a grandiose plan for China's industrialization, concocted by Sun without much help from engineers or economists.

Although sanctified by his followers, Sun's doctrine was not his major strength. All contemporary sources attribute to him a magnetic personality, a great capacity for tolerating others' weaknesses, a singular dedication to the pursuit of power, and a knowledge of the West unequalled by any of his political rivals. Perhaps the last factor is the most important, for it is this that set Sun apart and made him the symbol of Chinese modernization. Quite fittingly, the Chinese Communists call him "a pioneer of the revolution."

BIBLIOGRAPHY. An old but still useful biography is LYON SHARMAN, *Sun Yat-sen: His Life and Its Meaning* (1934). A more recent work, *Sun Yat-sen and the Origins of the Chinese Revolution* by HAROLD Z. SCHIFFRIN (1968), is a solidly researched piece, although it covers only the years to 1912. For a sympathetic treatment by a lawyer who knew Sun well, see PAUL LINEBARGER, *Sun Yat-sen and the Chinese Republic* (1925). In *The Japanese and Sun Yat-sen* (1954), MARIUS B. JANSEN analyzes the support Sun solicited and received from the Japanese from the 1890s to 1917. A recent biography is HOWARD L. BOORMAN (ed.), *Biographical Dictionary of Republican China*, vol. 3, pp 170–189 (1970).

(Y.C.W.)

Superconductivity

Superconductivity is a state of matter into which many chemical elements, compounds, and alloys enter on being cooled to exceedingly low temperatures; and while in this state their thermal, electric, and magnetic properties are strikingly different from their properties in the normal state and from those of other materials that do not become superconducting at such temperatures. The unique properties of superconductors have stimulated intensive research and have inspired the development of many new devices (see CRYOGENICS, APPLICATIONS OF).

HISTORY

Superconductivity was discovered in 1911 by a Dutch physicist, Heike Kamerlingh Onnes, who was awarded the Nobel Prize in 1913 for his low-temperature research. Kamerlingh Onnes found that the electrical resistivity of a mercury wire disappears suddenly when it is cooled below a temperature of about 4° Kelvin (—269° C or

Discovery by Kamerlingh Onnes

—452° F; absolute zero is 0° K, a temperature at which all matter theoretically is considered to have lost disorder). Kamerlingh Onnes' discovery was a completely unexpected phenomenon. He soon discovered that a superconducting material can be returned to the normal state either by passing a sufficiently large current through it or by applying a sufficiently strong magnetic field to it. At that time, there was no theory to explain this.

For many years it was believed that, except for the fact that they had no electrical resistance (*i.e.*, they had infinite electrical conductivity), superconductors had the same properties as normal materials. This belief was shattered in 1933 by the discovery that a superconductor is highly diamagnetic; that is, it is repelled by a magnetic field. A ferromagnetic substance easily acquires and then retains strong magnetic properties. A paramagnetic substance is only weakly magnetizable, whereas diamagnetic substances may be said to expel magnetic field that would permeate a magnetizable substance. This phenomenon in superconductors is called the Meissner effect for one of the two men who discovered it. Because such behaviour relates to changes in energy, its discovery made it possible, in the same year, to consider the transition from normal to superconducting states as being thermodynamically reversible. The discovery of the Meissner effect also made it possible to formulate, in 1934, a theory of the electromagnetic properties of superconductors that predicted the existence of an electromagnetic penetration depth (for a definition, see below *The Meissner effect*), which was first confirmed experimentally in 1939. In 1950, it was clearly shown for the first time that a theory of superconductivity must take into account the fact that free electrons in a crystal are influenced by the vibrations of atoms that define the crystal structure, called the lattice vibrations. In 1953, in an analysis of the thermal conductivity of superconductors, it was recognized that the distribution of energies of the free electrons (those not attached to atoms) in a superconductor is not uniform from one extreme to the other but has a separation called the energy gap.

The theories referred to thus far served to show some of the interrelationships between observed phenomena but did not explain them as consequences of the fundamental laws of physics. For almost 50 years after Kamerlingh Onnes' discovery, theorists were unable to develop a fundamental theory of superconductivity. Finally, in 1957, such a theory was presented by physicists John Bardeen, Leon N. Cooper, and John Robert Schrieffer of the United States, which won for them (1972) the Nobel Prize in physics. Now called (for them) the BCS theory, all later theoretical work is based on it. The BCS theory also provided a foundation for an earlier model that had been introduced by the Soviet physicists Lev Davidovich Landau and Vitaly Lazarevich Ginzburg (1950); this model has been useful in understanding electromagnetic properties, including the fact that the internal magnetic flux in superconductors exists only in discrete amounts (instead of in a gradual and continuously changing spectrum of values), an effect called the quantization of magnetic flux. This flux quantization, which had been predicted from quantum-mechanical principles, was first observed experimentally in 1961.

The BCS theory

In 1962 Brian D. Josephson, a physicist in England, predicted that two superconducting objects placed in electrical contact would display certain remarkable electromagnetic properties. These properties have since been observed in a wide variety of experiments, demonstrating quantum-mechanical effects on a macroscopic scale.

The theory of superconductivity has been tested in a wide range of experiments, involving, for example, ultrasonic absorption studies, nuclear-spin phenomena, low-frequency infrared absorption, and electron-tunnelling experiments (see below *Energy gaps*). The results of these measurements have brought understanding to many of the detailed properties of various superconductors. The search for materials that remain superconducting at higher temperatures and in stronger magnetic fields has been pursued on an empirical basis, particularly by Bernd Teo Matthias, a German-born U.S. physicist.

THERMAL PROPERTIES OF SUPERCONDUCTORS

Superconductivity as a property of matter is best understood in terms of atomic structure, specifically of the arrangement of electrons in atoms. Every atom is identified by its nucleus, consisting of two kinds of subatomic particles, positively charged protons and uncharged neutrons. The nucleus is surrounded by electrons equal in number (the atomic number of the element) to the protons and arranged in what are called shells and subshells, or orbitals. Each shell can contain any number of electrons up to a maximum for each shell. In lighter elements the inner shells are filled and only the outermost shell is incomplete; in the heavier elements one or more inner shells, as well as the outermost, may be incomplete. Positions the electrons occupy in the structure are identified on the basis of energy relationships, and electrons can be excited from a normal, or ground state of energy, or energy level, to other energy levels (*i.e.*, to other positions in the structure, or to leave the atom completely). All chemical and most physical properties of the elements —and of the compounds, gases, liquids, and solids they form—are understood in terms of energy changes in the electron structures of atoms. The energy that an electron can absorb or emit in terms of specific units of energy that can be exchanged by electrons or between electrons and radiations such as light is defined by quantum mechanics. As a result of the complex arrangement of electron orbitals in a material when in the superconductive state, there may occur a large interval of unoccupied energy levels assigned to electrons, called an energy gap. Atoms bound into crystals hold their electrons tightly, but atoms of metals release a few electrons, and these free electrons conduct heat and electricity.

Hundreds of materials are known to become superconducting at low temperatures. Twenty-six of the chemical elements, all of them metals, are superconductors in their usual crystallographic forms at atmospheric pressure. Among these are commonly known metals such as aluminum, tin, lead, and mercury, and less common ones such as rhenium, lanthanum, and protactinium. In addition, 13 chemical elements that are metals, semimetals, and semiconductors (nonmetals that are electrically conductive under certain conditions) are superconductors at low temperatures and high pressures. Among these are uranium, cerium, silicon, and selenium. Bismuth, though not superconducting in its usual crystallographic form, can be made superconducting by preparing it in the amorphous (noncrystalline) form, which is stable at extremely low temperatures. Superconductivity is not exhibited by any of the magnetic elements chromium, manganese, iron, cobalt, and nickel.

Superconducting materials

Most of the known superconductors are alloys (homogeneous mixtures of metals) or compounds. It is possible for a compound to be superconducting even if the chemical elements constituting it are not; examples are disilver fluoride (Ag_2F) and a compound of carbon and potassium (C_8K). Some semiconducting compounds, such as tin(II) telluride (SnTe), can become superconducting if they are properly infused with impurities.

Transition temperatures. The temperature below which a substance is superconducting is called the transition temperature (T_c). The vast majority of the known superconductors have transition temperatures that lie somewhere between 1° Kelvin and 10° K. Of the chemical elements, tungsten has the lowest transition temperature, 0.015° K, and niobium the highest, 9.5° K. The transition temperatures of some of the commonly known superconducting elements are indicated in the Table. An alloy of niobium, aluminum, and germanium ($Nb_{12}Al_3Ge_1$) has the highest known transition temperature, 21° K.

Among the transition elements there is a strong correlation between transition temperature and the number of electrons in the outer shell of the atom. When this number is five (vanadium, niobium, and tantalum) or seven (technetium and rhenium) the transition temperature is markedly higher than for other transition elements.

The transition temperature is sometimes unusually sensitive to chemical composition, as in the case of magnetic impurities. A few parts per million of manganese in zinc,

Transition Temperatures and Low-Temperature Values of the Critical Magnetic Field of Some Superconducting Elements

	T_c (° Kelvin)	H_0 (oersted)
Zinc	0.88	53
Aluminum	1.20	99
Indium	3.41	282
Tin	3.72	306
Mercury	4.15	411
Lead	7.19	803

for example, lowers the transition temperature considerably. It has already been mentioned that the transition temperature can depend on pressure and crystal structure. The latter is more effective in influencing the transition temperatures of compounds than of alloys, especially of compounds containing at least one transition element. Some types of crystal structure tend to have higher transition temperatures than others, and some types are seldom superconductors, but some types are found in both groups.

Specific heat and thermal conductivity. The thermal properties of a superconductor can be compared with those of the same material at the same temperature in a normal state. (The material can be forced into a normal state at the temperature by a large enough magnetic field.)

When a small amount of heat is put into a system, some of the energy is used to increase the lattice vibration (an amount that is the same for a system in the normal and in the superconducting state) and the remainder is used to increase the kinetic energy of the free electrons. The specific heat (C_e) of the electrons is defined as the ratio of that portion of the heat used by the electrons to the rise in temperature of the system. Figure 1 shows how the

Adapted from *Zeitschrift fur Physik* (1959)

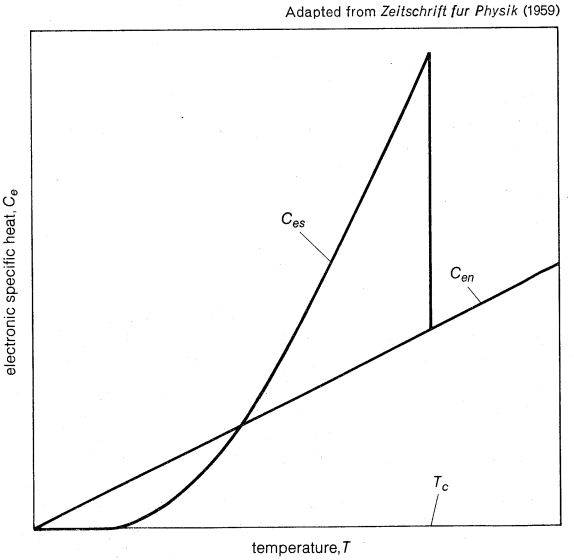


Figure 1: Specific heat in the normal (C_{en}) and superconducting (C_{es}) states as a function of absolute temperature. The two functions are identical at the transition temperature (T_c) and higher.

specific heat of the electrons in a superconductor varies with the absolute temperature (T) in either the normal or the superconducting state. It is evident from the figure that the electronic specific heat in the superconducting state, designated C_{es}, is smaller than in the normal state, designated C_{en}, at low enough temperatures, but that C_{es} becomes larger than C_{en} as the transition temperature T_c is approached, at which point it drops abruptly to C_{en}.

Temperature dependence of electronic specific heat

To a good approximation C_{es} is found to be proportional to the cube of the absolute temperature. Precise measurements, however, have shown deviations from this temperature dependence and have indicated that, at temperatures considerably below the transition temperature, the logarithm of the electronic specific heat is inversely proportional to the temperature.

This temperature dependence, together with certain principles of statistical mechanics (see PHYSICAL THEORIES, MATHEMATICAL ASPECTS OF), strongly suggests that there is a gap in the distribution of energy levels available to the electrons in a superconductor, so that a minimum energy is required for the excitation of each electron from a state below the gap to a state above it.

The heat flow per unit area of a sample equals the product of the thermal conductivity (K) and the temperature gradient ∇T. This may be expressed as $J_Q = -K \nabla T$, the minus sign indicating that heat always flows from a warmer to a colder region of a substance.

It is a fact that thermal conductivity in the normal state (K_n) approaches thermal conductivity in the superconducting state (K_s) as the temperature (T) approaches the transition temperature (T_c) for all materials, whether they are pure or impure. This suggests that the energy gap (Δ) for each electron approaches zero as the temperature (T) approaches the transition temperature (T_c). This would also account for the fact that the electronic specific heat in the superconducting state (C_{es}) is higher than in the normal state (C_{en}) near the critical temperature: as the temperature is raised toward the transition temperature (T_c), the energy gap in the superconducting state decreases, the number of thermally excited electrons increases, and this requires the absorption of heat.

Energy gaps. As stated above, the thermal properties of superconductors indicate that there is a gap in the distribution of energy levels available to the electrons, and so a finite amount of energy, designated as delta (Δ), must be supplied to an electron to excite it. This energy is maximum at absolute zero (designated Δ_0) and changes little with increase of temperature until the transition temperature is approached, where Δ decreases to zero, its value in the normal state. The BCS theory predicts an energy gap with just this type of temperature dependence.

According to the BCS theory, there is a type of electron pairing (electrons of opposite spin acting in unison) in the superconductor that is important in interpreting many superconducting phenomena. The electron pairs, called Cooper pairs, are broken up as the superconductor is heated. Each time a pair is broken, an amount of energy that is at least as much as the energy gap (Δ) must be supplied to each of the two electrons in the pair, so an energy at least twice as much (2Δ) must be supplied to the superconductor.

Cooper electron pairs

The value of twice the energy gap at 0° K (which is $2\Delta_0$) might be assumed to be higher when the transition temperature of the superconductor is higher. In fact, the BCS theory predicts a relation of this type, namely, that the energy supplied to the superconductor at absolute zero would be equal to 3.53 times the product of the Boltzmann's constant, k, and the transition temperature, T_c, in degrees Kelvin; *i.e.*, $2\Delta_0 = 3.53\ kT_c$. Boltzmann's constant is derived from theoretical considerations of the energy distribution among aggregates of particles (it equals 1.38×10^{-23} joule per degree Kelvin, one joule being equal to 0.239 calorie).

The energy gap (Δ) can be measured most precisely in a tunnelling experiment (a process in quantum mechanics that allows an electron to escape from a metal without acquiring the energy required along the way according to the laws of classical physics). In this experiment, a thin insulating junction is prepared between a superconductor and another metal, assumed here to be in the normal state. In this situation, electrons can quantum-mechanically tunnel from the normal metal to the superconductor if they have sufficient energy. This energy can be supplied by applying a negative voltage (V) to the normal metal, with respect to the voltage of the superconductor.

Tunnelling will occur if eV, the product of the electrons' charge, e (-1.60×10^{-19} coulomb) and voltage, is at least as large as the energy gap Δ. The current flowing between the two sides of the junction is small up to a voltage equal to the energy gap divided by the electron's charge ($V = \Delta/e$), and then rises sharply. This provides an experimental determination of the energy

gap (Δ). In describing this experiment it is assumed here that the tunnelling electrons must get their energy from the applied voltage, rather than from thermal excitation.

MAGNETIC AND ELECTROMAGNETIC PROPERTIES OF SUPERCONDUCTORS

Critical field. One of the ways in which a superconductor can be forced into the normal state is by applying a magnetic field. The minimum magnetic field required to cause this transition is called the critical field (H_c) if the sample is in the form of a long, thin cylinder or ellipsoid and the field is oriented parallel to the long axis of the sample. (In other geometries the sample goes from the superconducting state into an intermediate state in which some regions are normal and others are superconducting, and finally into the normal state.) The critical field increases with decreasing temperature toward a value at absolute zero (H_0). For the superconducting elements, values of H_0 range from 1.2 oersted (oersted is a unit of magnetic field intensity) for tungsten to 830 oersteds for tantalum. Values of H_0 are listed in the Table for some common superconductors.

These remarks about the critical field apply to ordinary (so-called type I) superconductors. In the following section the behaviour of other (type II) superconductors is examined.

The Meissner effect. As was stated above, a type I superconductor in the form of a long, thin cylinder or ellipsoid remains superconducting at a fixed temperature as an axially oriented magnetic field is applied, providing the applied field does not exceed a critical value (H_c). Under these conditions, superconductors exclude the magnetic field from their interior, as could be predicted from the laws of electromagnetism and the fact that the superconductor has no electrical resistance. A more astonishing effect occurs if the magnetic field is applied in the same way to the same type of sample at a temperature above the transition temperature and is then held at a fixed value while the sample is cooled. It is found that the sample expels the magnetic field or flux as it becomes superconducting. This is called the Meissner effect. Complete expulsion of the magnetic flux (a complete Meissner effect) occurs in this way only for certain superconductors, called type I superconductors, and only for samples that have the described geometry. For samples of other shapes, including hollow structures, some of the magnetic flux can be trapped, and one has an incomplete or partial Meissner effect. Figure 2 shows a graph of the magnetization of a type I sample with a complete Meissner effect.

Reprinted with permission from K. Maki, *Physics* (1964); Pergamon Press

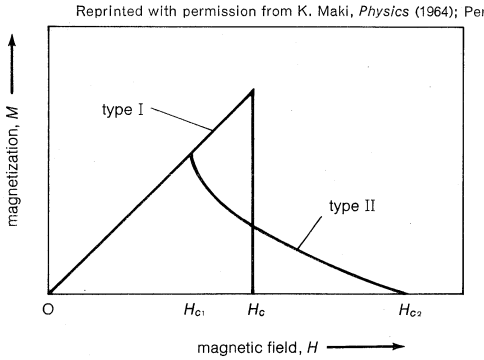


Figure 2: Magnetization as a function of magnetic field for a type I superconductor and a type II superconductor. For type I, magnetic flux is expelled, producing a magnetization (M) that increases with magnetic field (H) until a critical field (H_c) is reached, at which it falls to zero as with a normal conductor. A type II superconductor has two critical magnetic fields (H_{c1} and H_{c2}); below H_{c1}, type II behaves as type I and above H_{c2} it becomes normal.

Behaviour of type II superconductors

Type II superconductors have a different magnetic behaviour. Examples of materials of this type are niobium and vanadium (the only type II superconductors among the chemical elements), concentrated alloys, and superconducting semiconductors. As a sample of this type, in the form of a long thin cylinder or ellipsoid, is exposed to a decreasing magnetic field that is axially oriented with the sample, the increase of magnetization instead of occurring suddenly at the critical field (H_c) sets in gradually. Beginning at the upper critical field (H_{c2}), it is ideally completed at a lower critical field (H_{c1}; see Figure 2). If the sample is of some other shape or is inhomogeneous or strained, its magnetic behaviour may deviate from the ideal behaviour just described; in particular, some magnetic flux frequently remains trapped and some magnetization of the sample remains after the applied field is completely removed. Known values of the upper critical field extend all the way up to 3.5×10^5 oersteds, the value for the compound of vanadium and gallium with formula V_3Ga.

The expulsion of magnetic flux by type I superconductors in fields below the critical field (H_c) or by type II superconductors in fields below H_{c1} is never quite as complete as has been stated in this simplified presentation, because the field always penetrates into a sample for a small distance, known as the electromagnetic penetration depth. Values of the penetration depth for the superconducting elements at low temperature lie in the range from about 390 to 1300 angstroms ($1\text{Å} = 10^{-8}$ centimetre). As the temperature approaches the critical temperature, the pentration depth becomes extremely large.

Electromagnetic penetration depth

High-frequency electromagnetic properties. Thus far the discussion has pertained to the behaviour of superconductors in the absence of electromagnetic fields (represented by infrared waves, for example) or in the presence of steady or slowly varying fields, but high-frequency properties of superconductors have also been studied.

The energy gap in a superconductor has a direct effect on the absorption of electromagnetic radiation. At low temperatures, at which a negligible fraction of the electrons are thermally excited to states above the gap, the superconductor can absorb a quantum of energy only if it is at least twice the gap energy (at absolute zero, $2\Delta_0$). In the absorption process, a photon (a quantum of electromagnetic energy) is absorbed and a Cooper pair is broken, exciting both electrons in the pair. The photon's energy (E) is related to its frequency (f) by the famous Planck relation, energy is equal to Planck's constant times the frequency, $E = hf$, in which h is Planck's constant (6.63×10^{-34} joule second). Rearranging the equation gives $f = E/h$. Hence the superconductor can absorb electromagnetic energy only for frequencies at least as large as twice the energy gap divided by Planck's constant, or $2\Delta_0/h$. Data taken in the infrared and microwave regions of the spectrum therefore determine the size of the energy gap for a particular material.

Magnetic-flux quantization. In quantum mechanics, electrons have the properties of waves, and the properties of an electron can be summed up in what is called a wave function. If the wave functions are in phase (*i.e.*, act in unison), they are said to be coherent. The theory of superconductivity indicates that there is a single, coherent, quantum-mechanical wave function that determines the behaviour of all of the superconducting electrons. As a consequence, a direct relationship can be shown to exist between the velocity of these electrons and the magnetic flux (Φ) enclosed within any closed path inside the superconductor. Indeed, inasmuch as the magnetic flux arises because of the motion of the electrons, the magnetic flux can be shown to be quantized; *i.e.*, the intensity of this trapped flux can change only by units of Planck's constant divided by twice an electron's charge.

When a magnetic field enters a type II superconductor (in an applied field between the lower and upper critical fields, H_{c1} and H_{c2}) it does so in the form of quantized fluxoids, each carrying one quantum of flux. These fluxoids tend to arrange themselves in regular patterns that have been seen by electron microscopy and by neutron diffraction. If a current is passed through the superconductor, the fluxoids can be made to move. This motion leads to energy dissipation that can heat the superconductor and drive it into the normal state. In making wire for superconducting high-field magnets, manufacturers thus try to fix the positions of the fluxoids by deliberately making the wire inhomogeneous in composition.

Fluxoids

Josephson currents. If two superconductors are separated by an insulating film that forms a low-resistance junction between them, it is found that Cooper pairs can tunnel from one side of the junction to the other. (This process occurs in addition to the single-particle tunnelling already described.) Thus a flow of electrons, called the Josephson current, is generated and is intimately related to the phases of the coherent quantum-mechanical wave function for all of the superconducting electrons on the two sides of the junction. As a result, several novel phenomena could be predicted, and experiments have demonstrated them. These are collectively called the Josephson effect or effects.

The first of these phenomena is the passage of current through the junction in the absence of a voltage across the junction. The maximum current that can flow at zero voltage depends on the magnetic flux (Φ) passing through the junction as a result of the magnetic field generated by currents in the junction and elsewhere. The dependence of the maximum zero-voltage current on the magnetic field applied to a junction between two superconductors is shown in Figure 3.

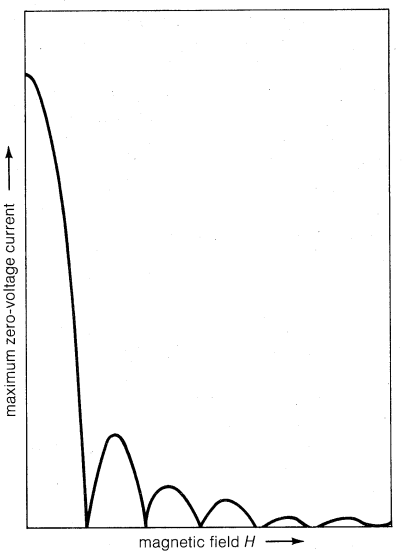


Figure 3: Maximum zero-voltage (Josephson) current passing through a junction by Cooper-pair tunnelling as a function of magnetic field.

Oscillating currents

A second class of Josephson effects is an oscillating current resulting from a relation between the voltage across the junction and the frequency (f) of the currents associated with Cooper pairs passing through the junction. The frequency (f) of this Josephson current is equal to two times the product of the electron charge (e) and the impressed voltage (V) divided by Planck's constant (h); that is, $f = 2eV/h$. Thus the frequency increases by 4.84×10^{14} hertz (cycles per second) for each additional volt applied to the junction. This can be demonstrated in various ways. The voltage can be established with a source of direct current (dc) power, for instance, and the current be detected by the electromagnetic radiation of frequency (f) that it generates. Another method is to expose the junction to radiation of another frequency (f') generated externally. It is found that a graph of the dc current versus voltage has current steps at values of the voltage for Josephson frequencies that are integral multiples (n) of the external frequency ($f = nf'$); that is, the voltage is equal to an integer times Planck's constant times the external frequency divided by twice the electron charge, or in equation form, $V = nhf'/2e$. The observation of current steps of this type has made it possible to measure the ratio of Planck's constant to the charge on the electron (h/e) with far greater precision than by any other method and has therefore contributed to a knowledge of the fundamental constants of nature.

The Josephson effect has been utilized in the invention of novel devices for high-sensitivity measurements of currents, voltages, and magnetic fields.

CONTINUING INVESTIGATIONS OF SUPERCONDUCTIVITY

In addition to the types of tunnelling measurements that have already been described, experiments have been performed in which the single-particle tunnelling process is assisted by energy supplied by phonons (lattice vibrations) or photons (electromagnetic oscillations). These have yielded valuable information on the interactions between electrons and these forms of energy. Experiments in which magnetic impurities are present, either in the insulating layer or in one of the metals at the tunnelling junction, have been used to explore the electronic states of these impurities.

Measurements of the attenuation of ultrasonic waves in an elastic medium have yielded values of the energy gap for electrons moving in particular directions in a crystal. The attenuation of ultrasonic waves in the superconducting state is less than that in the normal state, for frequencies that are less than the gap energy (associated with a Cooper electron pair) divided by Planck's constant ($2\Delta/h$), because a phonon in that frequency range has insufficient energy to excite the electrons that are in unexcited states below the gap.

Extensive measurements have been made of the effect of magnetic impurities on the transition temperature and other properties of superconductors. The effects are sensitive to any spin ordering (orientation of electron spins in a crystal), such as the ordering in nickel and iron, and therefore yield information about this ordering.

The properties of a superconductor never change discontinuously at the transition temperature. The direct-current electrical resistance, for example, decreases continuously as the sample is cooled through a finite range of temperatures. The width of this range is usually small and has been observed to be as small as 2×10^{-5}° K in a pure, unstrained superconductor. (The transition range can be as wide as several degrees, as a result of large inhomogeneities or strains in a sample.) Near the transition temperature, thermodynamic fluctuations between the superconducting and the normal states occur at localized regions of the material. When the width of the transition is mainly a result of these fluctuations, information about them can be obtained by measurements of the properties of the material in the transition region.

Transition range

BIBLIOGRAPHY. E.A. LYNTON, *Superconductivity,* 3rd ed. (1969), a short, good introduction; P.G. DE GENNES, *Superconductivity of Metals and Alloys* (1966), an excellent presentation of the foundations of the subject showing how the theory is applied to many phenomena; R.D. PARKS (ed.), *Superconductivity,* 2 vol. (1969), an excellent and exhaustive compendium; D.M. GINSBERG, "Resource Letter Scy-1 on Superconductivity," *Am. J. Phys.,* 32:85–89 (1964), and ". . . Scy–2 . . . ," *ibid.,* 38:949–955 (1970), introductory review articles with descriptive bibliographies.

Classic articles: H. KAMERLINGH ONNES, "Superconductors," *Communs Kammerlingh Onnes Lab.,* suppl. 34b:55–70 (1913), a summary of the first two years of Kamerlingh Onnes' brilliant work on superconductivity; C.J. GORTER and H. CASIMIR, "On Superconductivity I," *Physica,* 1:306–30 (1934), the classic thermodynamic treatment of the subject; F. and H. LONDON, "The Electromagnetic Equations of the Superconductors," *Proc. R. Soc.,* A149:71–88 (1935), presents the London theory of the electromagnetic properties of superconductors, the first theory to emphasize the importance of magnetic properties (the Meissner effect); A.B. PIPPARD, "An Experimental and Theoretical Study of the Relation Between Magnetic Field and Current in a Superconductor," *ibid.,* A216:547–568 (1953), a classic experimental and theoretical investigation into the high-frequency electromagnetic properties of superconductors; J. BARDEEN, L.N. COOPER, and J.R. SCHRIEFFER, "Theory of Superconductivity," *Phys. Rev.,* 108:1175–1204 (1957), the fundamental microscopic theory of superconductivity; I. GIAEVER, "Energy Gap in Superconductors Measured by Electron Tunneling," *Phys. Rev. Lett.,* 5:147–148 (1960), the first tunnelling experiment on a superconductor; B.D. JOSEPHSON, "Supercurrents Through Barriers," *Adv. Phys.,* 14:419–451 (1965), the theory of the Josephson effect; J.P. GOLLUB *et al.,* "Observation of Enhanced Diamagnetism Above T_c in Indium Due to Thermodynamic Fluctuations," *Phys. Rev. Lett.,* 22:1288–1291 (1969), the first experiment on the influence of thermodynamic fluctuations on the diamagnetism of a superconductor near its transition temperature.

(D.M.Gi.)

Surgery

Surgery is identified by the general public as the surgical operation and is defined by some dictionaries as the art and practice of treating injuries, deformities, and other disorders by manual and instrumental means. Surgery actually includes far more than this. The modern surgeon conducts the patient through the diagnostic phase, the preoperative preparation, the operation itself, and the postoperative and rehabilitative period. Thus, it is obvious that the operation takes relatively little of the surgeon's time and thought except in minor injuries. It also should be evident that the surgeon must have an extensive medical knowledge of the working of the body as well as a knowledge of the diseases that can be treated by operative means.

Surgery is fundamentally involved in the management of acute injury as differentiated from chronic, slowly progressing diseases, except when patients with the latter type of disease must be operated upon. Not only does the surgeon take care of the accidentally injured man and the battle casualty but his deliberate violation of the body, as he cuts into it in the elective operation, creates the same sort of damage as is received in an automobile accident. Secondly, there are but three basic acute bodily injuries in the surgeon's province of treatment. These are: loss and disruption of body tissue, as represented by the wound; loss of body fluids, such as blood and gastrointestinal juices, resulting in shock and dehydration; and the invasion of the body by disease-causing pathogenic organisms that produce localized or generalized infection. These three injuries may exist singly or in combination and may result in death if not combatted.

Since the wound is the central happening in surgery, a knowledge of the significance of the wound to the body as a whole and knowledge of wound care make surgery a unique medical specialty. The surgeon's education and knowledge are among the most comprehensive in medicine.

DEVELOPMENT OF SURGERY

Surgery is as old as mankind, and anyone who has ever staunched a wound has acted as a surgeon. In some ancient civilizations surgery reached a rather high level of development, particularly in India, China, Egypt, and, later, in the Hellenistic period of Greek civilization (from the 4th century BC to the 2nd century AD). Medical knowledge and, thus, also surgical knowledge was usually invested in the hands of the priests or the religious orders, particularly in the Western world. It was not until a 13th-century pope interdicted the practice of law and medicine by the clergy that medicine came into the hands of the laymen. The practice of surgery was not taught in most universities, and ignorant barbers wielded the knife. Often they acted on their own responsibility, but frequently they were called in by physicians, who directed their operating. Some barber-surgeons became eminent, wrote treatises on surgery, and were good observers. The Huguenot Ambroise Paré, who lived in 16th-century France, was a barber-surgeon of this type. He used the ligature for blood vessels, for example, instead of cauterizing wounds with hot oil, an Arabic practice. He also used the "kind" treatment of cleansing wounds by washing them with wine and water. The principles of his wound treatments resembled those of the Hippocratic school of Greek medicine (5th–4th centuries BC). Hippocrates stressed that the surgeon should foster and use the healing power of nature rather than interfere unnecessarily. In 1540, in the reign of Henry VIII, organization of the Company of Barber Surgeons of London marked the beginning of some control in qualification for performing operations. This guild, which lasted two centuries, was the precursor of the Royal College of Surgeons of England.

Surgery by barbers

At the end of the 18th and the beginning of the 19th centuries, surgery was firmly established in the medical curriculum. In the period that closed with the public demonstration of anesthesia by the use of ether as an anesthetic, in 1846, the development of surgery was remarkable, despite the lack of anesthesia. Progress at this time was achieved by the application of accurate anatomic knowledge and by the development of skill and rapidity of operating. It was the time of great names in surgery, the cult of the personality. Patients were sedated with opium, made drunk with alcohol, and tied down, so that a leg amputation, for example, could be done in three to five minutes.

Almost simultaneously with the beginnings of modern surgery, which dates from the introduction of ether anesthesia (1846), came the identification of micro-organisms as the cause of infection in wounds. Promulgation of the germ theory of disease by the French microbiologist Louis Pasteur (mid-19th century) and the application of this theory to wound sepsis by the British surgeon Joseph Lister resulted in the technique of antisepsis, which brought about a remarkable reduction in the mortality rate from wounds and elective operations. Before, micro-organisms had been observed in pus from infected wounds, but were thought to be the product of sepsis (putrefaction) rather than its cause. Infection hitherto was considered of biochemical origin. Hospital gangrene, so-called, was so common as to interdict elective surgery at various times in the 19th century. The spread of infection from one patient to another by attendants, including the surgeon, was the cause of the high rate of sepsis. Lister's observation that it was only compound fractures (those in which the skin had been broken and the deep tissue exposed to the environment) that became infected made him sure that the micro-organisms introduced from the environment were the cause of hospital gangrene. Lister's technique of employing antibacterial carbolic-acid spray was instituted gradually in the latter part of the 19th century. It was supplanted by the aseptic technique of sterilizing instruments and materials by moist heat under pressure. This technique is still in use.

20th-century developments

At the turn of the century, the discovery of X-rays by the German physicist Wilhelm Conrad Röntgen added an important diagnostic tool without which modern surgery would not exist. Next, the discovery of blood types in 1900 by the Austrian biologist Karl Landsteiner made transfusions safe. New techniques of anesthesia involving not only new agents for inhalation but also regional anesthesia accomplished by nerve blocking (spinal and local anesthesia) also were introduced. Positive pressure and controlled respiration techniques made chest surgery safe. Ultimately, the intravenous administration (injection into the veins) of anesthetic agents made its appearance. Finally, in the present period, the replenishment of body fluids by intravenous infusion, the introduction of chemicals and antibiotics (in the 1930s and 1940s, respectively) for antimicrobial use and the treatment of the metabolically disturbed body as well as the local wound were additional factors in bringing surgery to its present state, in which every body cavity, system, and area can safely be operated upon.

Today, experimental surgery in animals prepares the way for new and better human application. New operative techniques, such as the application of the laser beam and ultrasonics, have emerged in the experimental and in some cases the practical phases of surgery. These will be discussed in more detail later.

SURGICAL THERAPY OF THE BODY AS A WHOLE

That an injury has a general effect on the body is a most important concept of surgical thought, for an understanding of it as a guide to therapy affects the outcome of serious operations or injuries. The celebrated British surgeon John Hunter appreciated the magnitude of the general effect of acute injury when, in the early 19th century, he wrote: "What does the wound signify to the organism? By what mechanism does the organism safeguard itself against the effect of the wound immediate and remote?" It was not until the late 19th century, however, and more recently that the physiology of bodily response to injury became understood. The dread derangement of metabolism in surgical disease called surgical shock usually results from blood loss but may result from any of the three acute injuries.

The word injury, which formerly was used in surgery to signify trauma, now has an extended meaning that is suggested by the adjective injurious. In other words, it can refer to any action that does harm to the human body. And this extended definition is correlated to a similar development of surgical therapy extending beyond the operation itself.

The bodily response to acute injury protects the organism and preserves its resources, essentially under the control of the entire endocrine system and particularly the adrenal gland, this gland being recognized as the principal organ involved in the immediate response to stress; *i.e.*, triggering the stress reaction (see also the article HORMONE). The physician and physiologist Hans Selye spoke of "the alarm reaction" in describing the events that follow acute injury. Studies of the phenomenon have reached a stage at which the severity of a given injury can be measured in certain respects by the changes in metabolism and particularly by studies of the body fluids—blood, body water, and electrolytes (the salts of body water)—and the excretory function of the kidney.

The relatively new concept that the reaction to injury constitutes a complex homeostatic mechanism has developed from the work of the physiologists Claude Bernard, in the late 19th century, and Walter B. Cannon, in the first part of the 20th century. Cannon coined the term homeostasis for the ordered and coordinated progression of physiologic processes that keep most of the steady states in the body within a small range of variations. Homeostasis holds the body temperature relatively constant, for example, and serves to regulate the quantity and composition of body fluids, including the level of sugar in the blood. The endocrine glands have an active role in homeostatic mechanisms, and there is simultaneous involvement of the heart, brain, lungs, kidneys, and other organs.

Protective mechanisms

Not only complex metabolic reactions but also simple protective mechanisms are brought into play by injury. Present understanding of the way that the body responds to injury sharpens the realization that natural processes can point the way toward proper treatment. Thus an effort is now made to preserve the normal patterns within the body by introducing substances that are in short supply and to correct other abnormalities in the body's metabolism, rather than to employ more drastic surgical measures. One of the greatest blunders during the Middle Ages and the Renaissance was the practice of bloodletting, which killed many people who had already lost much blood and were in shock. The great success that the followers of Hippocrates had in their day rested on this understanding of the ability of the body to restore itself to health. Simple responses to injury can be seen, for example, in the man whose peptic ulcer begins bleeding while he is on a hunting trip and has no help at hand. He does not set out for camp as fast as he can run, as he might if the decision rested with him, but instead he succumbs to involuntary mechanisms that save his life. He goes into the state of surgical shock, usually remaining conscious after a preliminary fainting spell. This manoeuvre is helpful in preserving and maintaining every vital function in the presence of a diminished blood volume. Thus, his vital organs, particularly in the brain, can get enough nutriment (sugar) and oxygen to sustain life until the bleeding stops spontaneously. While his blood pressure remains below normal, his arterioles (small arteries) contract, which makes the circulating vascular system smaller and thus favours the best use of what blood is left to supply vital areas. The deficient oxygen supply to the brain is relieved as the pulse pressure increases. This is nature alone at work, for if the man's brain lacks oxygen for ten minutes he may die from that alone. Other factors favourable to spontaneous stopping of bleeding then begin to operate. If the vessel from which the blood is escaping has sufficient elasticity, it will close its channel spontaneously, and clotting will occur, which eliminates the hazard of continued blood loss. The man's body will adjust to the stress, and, after a period of continued rest, his blood volume may be partially restored from available fluid stored in his tissues outside the cells. Other mechanisms for conserving the body fluids come into play, the most obvious of which is the diminution of the output of urine. But, first, total body rest and rest of a particular part of the body are enforced by nature and must be recognized as homeostatic mechanisms of a high order. These physical homeostatic mechanisms are employed daily in surgical practice; they are called bed rest and immobilization of the injured parts. That is the reason fractures are immediately splinted at the scene of accidents.

Bed rest and immobilization

There is a sameness to all bodily injuries (wounding, blood and fluid loss, and infection) that is expressed by similar endocrine and metabolic responses. Overwhelming infection with septicemia (invasion of the bloodstream by the disease organisms), for example, produces shock similar to that produced by loss of blood. The normal metabolic response to injury is an overlapping interplay among its three basic phases: conservation, degradation, and repair. Study of the conservation phase has given the surgeon insight into how to help the organism that is under the stress of injury. The key to understanding the human body is that it is largely composed of water (approximately 60 percent by weight) and that this water is distributed in three compartments containing salts suitable for cellular function. The first compartment is the circulating blood, which, in addition to its water and salt content, has red blood cells, essential to the oxygenation of vital organs, and white blood cells, concerned in a protective antibacterial function. The volume of blood varies from 4,000 to 5,000 cubic centimetres (four to five quarts), and its loss by extensive bleeding results in surgical shock, the most critical condition engendered by injury. Blood replacement not only is essential to the treatment of shock when it occurs but it is the key to prevention of shock during elective operation. The other two fluid compartments are within the body cells and outside the cells in the tissues. A shift of fluid and salts (electrolytes) among the three compartments occurs in severe injury and changes the electrolytic constituency of blood serum. Such changes can be detected by chemical examination and thereby serve as a guide to fluid and salt replacement. Decreased elimination of water and salts by the kidneys invariably occurs in shock. This reaction, largely under the control of the adrenal glands, tends to preserve the organism, as has been said before. When an injured person cannot eat and drink and may be losing blood or extracellular fluids and salts in excessive amounts, it is the surgeon's job to replace loss quantitatively and qualitatively. This is usually done by intravenous infusions—introduction of fluids into a vein. Cortisone, a product of the cortex (outer substance) of the adrenal gland, is of the greatest importance in regulating water and salt retention and excretion. Measurements of obvious body-water losses and estimations of average known losses of water in exhaled air and sweat are recorded, and the differences between losses and intake of fluid tell whether the person has a subnormal, normal or greater than normal supply of fluid. Examination of blood serum (the blood minus its cells and clotting factors) must be made periodically and often several times daily. The aim of treatment is to keep the internal environment as near normal as possible.

Resuscitation

Resuscitation of the severely injured requires prompt measures to assure a good airway, essential to oxygenation of the blood. It may even be necessary to make an opening through the wall of the windpipe, especially if there are injuries to the chest. Air must be removed that has entered the pleural cavity—the space between the membrane covering the lungs and that which lines the chest—since air in this cavity makes it impossible for the lung on that side to functon and displaces the heart. Other measures that need rapid application include starting an arrested heart and acting to maintain body heat.

MONITORING

At present, surgical therapy is greatly helped by modern technology, particularly by monitoring devices that are used during surgery for the person who is severely in-

jured, during shock, and also during the postoperative period. The surgeon and his assistants, however, are still the best monitors.

Monitoring during the operation consists in obtaining all or part of the following data, depending upon the case. Arterial and venous blood pressure and pulse rate must be monitored. A fall in blood pressure accompanied by a rise in pulse rate is an indication of impending shock, since it gives evidence of critical loss of the volume of circulating blood. Other items monitored are the heart contractions as indicated by electrocardiograms; electroencephalographic records (tracings of brain waves), which reflect changes in brain function and may reveal insufficient oxygenation of this vital organ; the oxygen level in the arteries and veins; carbon dioxide partial pressure in the circulating blood; and respiratory volume and exchange. More extensive monitoring is useful in patients in shock and in the critical postoperative period.

Automatic monitoring

There are, at present, three general areas in which engineering technology can provide monitoring for critically injured patients. The simplest but most useful of such monitoring is that in which machines record heart and respiratory rate, continuous electrocardiographic tracings, urinary-output volume, and similar values. These machines are essentially replacements of human monitors (nurses and technicians) and thus have not gained universal favour, since too great a separation of the patient from human contact might occur from their use. The machines also are expensive. This type of monitoring is of considerable value, however, in patients undergoing open-heart surgery. It allows the simultaneous observation of variation in arterial and venous blood pressure, the filling pressure of the heart, and the resistance to circulation. From these data, plus simpler ones, the output of the heart can be calculated.

The second monitoring installation makes use of automatically obtained data and simultaneously carries out measures to correct observed abnormalities. Intravenous fluid administration for example, can be adjusted to the output of urine. Blood transfusions and drugs can be given as needed in the same manner. This represents a substitution for human performance as well as observation.

The third realm of the use of advanced engineering techniques involves the application of complex computer programming and analyses to clinical investigation of disease. The types of instruments in general use at present are those made for other purposes, although special designs for medical use are being developed. The great obstacle in computer usage is that a considerable amount of modern medicine is still empirical and subjective in nature and thus difficult to translate into objective language for programming.

OPERATIVE SURGERY

Common materials and instruments. The first and essential knowledge for operative surgery is of anatomy and pathology, the structure of the body and the changes wrought by disease; manual dexterity comes second. Next in importance are prevention of bleeding and adequate exposure of the part to be operated upon.

Asepsis, the avoidance of contamination by disease organisms, requires that all instruments and dry goods coming in contact with the surgical field be sterilized. This is done with the autoclave, a device that subjects them to 30 minutes of moist heat at 260°F (127°C) with 20 pounds of pressure. Chemical sterilization of some instruments is also used. The patient's skin is sterilized by chemicals. The surgical teams scrub their hands and arms for at least five minutes, after which hands and arms are bathed in alcohol. Sterilized gowns, caps, and masks that filter the team's exhaled air and sterilized gloves of rubber or disposable plastic complete the picture. Thereafter, attention in avoiding contact with nonsterilized objects is the basis of maintaining the technique. It must be said, however, that disease organisms are ubiquitous in the operating-room air, on the patient's skin, in his gastrointestinal tract, and in the exhalations of all persons in the room. There is an irreducible incidence of wound infection resulting from operations, and more so when the wounds are accidental. The greatest source of wound infection in elective surgery is from the patient himself, since not only his skin but also his gastrointestinal tract harbours a large number and variety of disease organisms. The gastrointestinal tract must be opened in many operations, and this contaminates the operative field. Since the advent of antibiotics, preoperative "sterilization" of the gastrointestinal tract is often done by the use of nonabsorbed agents, particularly in surgery of the large bowel (colon). Before certain operations antibiotics are administered as prophylactics against wound infection, particularly when foreign bodies such as plastic arterial grafts are to be implanted permanently. These measures have greatly reduced septic complications of intestinal surgery.

Aseptic techniques

Hemostasis (the arresting of bleeding) is achieved by use of the hemostat, a clamp with ratchets that grasps blood vessels or tissue; after application of hemostats, suture materials are tied around the bleeding vessels. Electrosurgery (cauterization) can be used to stop bleeding but is not as popular or as secure as the ligature. In a long, bloody operation hundreds of hemostats are used. Absorbent cotton napkins, called sponges, are used for drying the field (at one time marine sponges were used; thus the use of the word sponge). Estimation of blood loss is made by weighing discarded bloody sponges. Deficiencies in the patient's blood-clotting mechanism are sometimes encountered; these can be controlled in some instances. Thus, it is now possible to perform elective operations on persons with hemophilia, an inherited disease involving a defect in the coagulating power of the blood.

Although there is a constant increase in sophisticated surgical instruments, the basic ones have remained little changed since the 19th century. Those most commonly used are the scalpel (knife), scissors, hemostatic forceps, flexible tissue-holding forceps, wound retractors for exposure, crushing and noncrushing clamps for intestinal and vascular surgery, and the curved needle for working in depth.

Suture materials

There are two basic types of suture materials: absorbable catgut, which comes from sheep intestine and which is of varying sizes and strength depending on its chemical treatment, and nonabsorbable materials, made of silk, cotton, steel wire, and synthetic materials.

Catgut is used extensively to tie off small blood vessels that are bleeding, and, since the body absorbs it in time, no foreign materials are left in the wound to become a focus for growth of disease organisms. The nonabsorbable sutures, on the other hand, when buried remain forever in the wound and can harbour micro-organisms. Nevertheless, these materials are needed in wound closure in many instances because of their sustained strength. Personal preference for various suture materials also enters into the selection of one or another type. Suture materials used to close the skin are removed when sufficient healing has taken place. The only important rules that guide the surgeon in selection of suture material are that it is preferable not to use nonabsorbable suture, especially silk, in infected wounds and that monofilament stainless-steel wire of a single strand is the best suture material in situations in which sustained strength is needed over a long period.

Unusual and specialized techniques and instruments. *Stereotactic surgery.* Stereotactic surgery, used in neurosurgery, is the accurate placement of objects, hollow probes or electrodes, by three-dimensional vision using X-ray photography. The apparatus is extremely sophisticated. For brain stereotactic placement the head must be held in a vise, known landmarks visualized by X-ray, and very fine wire electrodes or probes advanced to the desired site under mathematical guidance. The desired site can then be verified by observing that the response to electrostimulation is in accord with the expected response. Diseased tissue can then be destroyed by a variety of agents (cold, chemicals, electrocoagulation). In the case of the brain, only a small burr hole in the skull is needed for the operation. The obvious value of this technique is its avoidance of extensive and difficult dissec-

tion. Also, areas of the brain can be reached by stereotaxis that defy conventional exposure. This field is extremely important in neurosurgery and especially in developing new physiologic surgery (beneficially altering bodily function). The pituitary gland, for example, situated at the bottom of the brain and difficult to expose conventionally, can be reached stereotactically through the nasopharynx (that part of the pharynx, or throat, that is above the soft palate) by electrode or probe and thus destroyed when the procedure is indicated.

Laser. The word laser is an acronym for "light amplification by stimulated emission of radiation." The apparatus, first constructed in 1960, produces high-intensity light or radiation that is monochromatic (of one colour) and emitted in almost perfectly parallel waves. Therefore, much power can be generated in a highly concentrated area. The laser is essentially a cauterizing agent and is absorbed by pigmented lesions. The use of the laser in surgery is still experimental and confined to the hands of experts. Pigmented tumours, freckles, and tattoos have been treated by this method. It has been used palliatively on cancer of the tongue and other superficial tumours. In depth, the laser has been used to destroy brain tumours by special techniques. A number of other experimental uses of the laser are found in the treatment of dental decay and in surgery on the blood vessels. With the development of sophisticated instruments for its application, the laser is expected to be useful in treatment of areas inaccessible by conventional surgery.

Electroencephalography. Electroencephalography, the recording of electrical activity of the brain by tracings, has been in use for a long time. Normal patterns of brain waves are known. Abnormalities in the electroencephalogram, or EEG, as it is commonly called, reflect alterations in the excitability of the brain cortex (the outer substance of the brain). This apparatus is useful in the diagnosis of epilepsy, brain tumours, and hematomas and infections in the brain. Hematomas are masses of blood that have escaped from vessels and press on adjacent tissues.

Ultrasonics. Ultrasonics is the science concerned with sounds too high to be heard by the human ear. These sound waves usually exceed 20,000 vibrations per second. Ultrasonic devices were used extensively during World War II for the underwater detection of submarines (sonar), and in medicine and surgery they have been found to be particularly useful in diagnosis. Ultrasonic waves produce certain characteristic echoes when they strike the boundaries between tissues of different densities. These echoes, or reflections, are then recorded. Since 1956 ultrasonic devices have been used to sterilize surgical instruments. In the treatment of the eye, high-frequency sound waves are being used experimentally to dissolve cataracts (opaque lenses) and practically to diagnose detachments of the retina, the light-sensitive layer in the back of the eye, and in other intraocular conditions. Ultrasonic nebulizers provide humidity to the windpipe and to the bronchioles—small air passages in the lungs—in persons with respiratory diseases. In the field of neurosurgery the ECHO encephalogram uses the reflection of ultrasonic waves to localize structures within the skull. Also in the field of neurosurgery, ultrasonic energy is being used to destroy certain nerve tissue and thereby alleviate disorders involving overactivity or intractable pain. Devices employing ultrasound to produce heat locally are used for treating inflammation in muscles and joints. Ultrasound has been used in studies of fetal growth and of the functioning of the heart valves. Its greatest use in the future will probably be in gathering information of this type and in the general area of diagnosis.

Cryosurgery. Cryosurgery is the destruction of tissue by means of extremely low temperatures rapidly achieved. Liquid nitrogen is the usual agent used. Temperatures of −125° to −130° C (−195° to −200° F) are usually obtained in clinical use. Cryosurgery has several desirable features. It can usually be employed without general anesthesia, and, by the use of special instruments —essentially hollow probes through which the destructive agent is delivered—brain tumours and other diseased brain tissue, difficult or impossible to reach by conventional surgical techniques, can be destroyed. Probes are placed with the aid of stereotactic techniques. Inoperable cancers of the tongue and mouth can be treated palliatively by this method, since it usually relieves pain and reduces the size of the tumour. It is safer than the laser in its application.

Advantages of cryosurgery

Microsurgery. In the mid-1950s the surgical microscope was developed for operations on the eye and ear. The surgical binocular microscope is now extremely complex. Greater visual facility has been gained with the use of colour filters to increase contrast. Multiple eyepieces to fit the eyes of several surgeons and magnification powers of small intervals such as six times, ten times, 16 times, 25 times, and 40 times (the limit of capability of the hand) have been additional valuable features of microsurgery.

Although surgical techniques had to be modified for microsurgery, the needed dexterity can be obtained by most surgeons. Special small instruments have been developed.

Most procedures are performed by the use of finger rather than wrist motions, especially in suturing. The fineness of suturing is limited by the needle diameter, but needles of $\frac{1}{5,000}$ of an inch in diameter have been used.

The microscope has been a particular boon to the surgery of small vessels, structures difficult to see, to dissect, or to suture under ordinary vision; these include the eye, nerves, the ureters, the bile ducts, and structures in the male and female reproductive organs. The surgical microscope has also proved useful in surgery on the coronary arteries.

Closed and open surgery

Basic techniques. There are two general techniques of wound treatment; the closed technique, in which all tissues, including the skin, are closed with suture material after completion of the operation; and the open technique, in which the wound is left open, to be closed later. The open technique is used in badly contaminated wounds in which infection might be anticipated if immediate closure were done. Open treatment allows wound drainage and thus avoids entrapment of micro-organisms, while also permitting observation of the wound. In military surgery, wounds are usually badly contaminated. Such wounds are subjected to debridement. Debridement consists of removing foreign bodies, such as shell fragments, fragments of clothing, and dirt, as well as all dead tissue upon which micro-organisms thrive. The wound is then extensively washed. It is left open for seven to ten days; after this period the wound is closed if it is clean. Serious infection is avoided by this method of treating contaminated wounds. The technique has a place in similarly contaminated civilian wounds; notably, compound fractures—fractures in which bone fragments have penetrated the skin. Infected wounds or tissues are cut into and drained, but drainage is sometimes delayed. The body's antimicrobial forces must be given time to work; during this process, an abscess is formed, which then can be drained. The principles for the treatment of all infection are the same. Drainage is established at the proper time. Resting of the body as a whole and of the part involved, general restoration of the bodily fluids, and the use of antimicrobials—substances to combat infection—complete the picture.

Major categories. There are five major categories of surgery. These are (1) wound treatment, (2) extirpative surgery, (3) reconstructive surgery, (4) physiologic surgery, and (5) transplantation of tissues and organs. The five categories are used in all surgical specialties.

Wound treatment. The technical aspects of wound treatment, already discussed to some extent, revolve around procuring good healing and the avoidance of infection. The elective wound, the wound made in the process of performing an operation, must be large enough so that the area to be operated upon can be seen accurately and injury to other structures can be avoided. In surgical jargon this is called exposure. Instruments called retractors, held by assistants, draw back the sides of the wound and further enhance exposure. Hemostasis not only prevents blood loss and improves visibility but

prevents other complications. In operations on the brain, for example, small, slow leaks of blood can cause serious pressure on the brain and result in unconsciousness. Inadequate hemostasis leaves pools of blood, which promote growth of micro-organisms and encourage wound infections.

Extirpative surgery. Extirpative surgery, one of the earliest types of surgery after wound treatment, involves removal of diseased tissue or organs.

Cancer surgery

Cancer surgery and "radical" operations are of this category. In cancer surgery, not only is the organ or part of the organ containing the cancer removed but so also is a wide margin of adjacent tissue, which might even include adjacent organs. Most importantly, the lymphatic drainage of the organ and its lymph nodes are cut out in one continuous mass (*en bloc*), since it is known that one route of the spread of cancer is by lymphatic channels. Until the present time, surgical excision has had no peer in the cure of cancer, and indeed most of the progress obtained in cancer survival has been made in the extension of the use of operative surgery and in the reduction of operative mortality. Radiation therapy plays some part in "cures" but is most used in palliation. Surgical excision is also palliative when it does not cure. The word radical, applied to surgery, refers to extensive surgical excision and not to heedless abandon in the use of the knife. In radical mastectomy, for example, the operation for cancer of the breast, the breast and the contents of the armpit, including the lymph nodes, are removed.

Reconstruction and replacement surgery. Reconstructive and replacement surgery deals with the consequences of injury—fractures, burns, congenital deformities, and the end results of degenerative-disease processes. To replace lost tissues, grafts, from the patient himself or from a donor, and plastic materials are used. Common examples of this category of surgery are the use of metal in reconstructing hip joints, replacement of diseased heart valves with valves of plastic, and the use of synthetic materials for replacing or bypassing diseased arteries.

Physiologic surgery. Physiologic surgery is a new field. It does not attack the disease but rather alters the body in such a way as to change its function for the betterment of the patient by eliminating the symptoms of the disease or by lessening their impact. An example of this kind of surgery is cutting the vagus nerve in persons who have peptic ulcer (an ulcer in the wall of the stomach or the duodenum). It is known that acid is the principal factor in the production of such ulcers and that the vagus nerve controls one phase of gastric-acid secretion. Severing the vagus nerve, therefore, reduces the secretion of gastric acid and improves the condition of the ulcer, although the fundamental cause of the disease cannot be attacked because its cause is not known.

Transplantation surgery. Grafting of skin from one point on a person's body to another has long been practiced. Grafts of skin from other persons are ultimately rejected but are of use temporarily in the coverage of burn surfaces. Progress has been made in transplantation of organs, including the heart, but the problem of rejection has not been solved. Various methods have been used to reduce the immunologic reaction, but the factor that has the most favourable effect upon the outcome of organ transplantation has been close relationship between the donor and the recipient. The best results are achieved when the two members of the pair are identical twins. Functioning kidney transplants between identical twins, for example, have survived eleven years or more.

Preoperative and postoperative care. *Preoperative care.* The principal work of preparation of patients for major elective operations is to restore them to as near their normal physiologic state as possible. By and large, most preoperative preparation consists of administering blood to the anemic patient with low serum-protein values and giving water and salt by injection into a vein if dehydration be a problem. In some cases antibiotics are given. The bowel, for example, is sterilized by administration of nonabsorbable antibiotics for 24 to 48 hours before an operation for cancer of the intestine is undertaken. Antibiotics are also administered before an operation when any sort of foreign body is going to be left in the body: plastic heart valves, for example. The study of the patient's heart and lung functions and other diagnostic tests are a major preoperative preparatory concern of the surgeon. Digitalis may be given preoperatively to stimulate the heart, for example. An estimation of risk is made, but few persons are denied major surgery on account of age alone, previous "heart attack" (coronary-artery disease), or substantial deficiencies in lung function. Statistical data on life expectancy, the average risk of the operation, and its chance of cure are all known to the surgeon. While concerned relatives of the patient can be informed of risk, type of operation, and other factors, it is finally the patient's own decision for or against surgery that is paramount; that is, if the patient is adult and of sound mind.

Pre-operative use of antibiotics

Preparation for emergency surgery must be done rapidly. Surgical shock, needing much blood replacement, is often a problem. Procuring and maintaining a good airway for breathing is essential. Teamwork and organization of resuscitation procedures have made for a high degree of proficiency in large hospitals and have resulted in saving many lives.

Postoperative care. The only difference between preoperative and postoperative care is quantitative. The objective—the restoration and maintenance of an optimum physiological state in a person who has been severely injured, often deprived of his full respiratory function, frequently dehydrated and anemic (especially if the blood supply is inadequately replenished throughout the operation), and often unable to eat and drink for several days or weeks after surgery—is the same as before the procedure was done. Thus, blood transfusions, intravenous administration of fluids, and the use of measures calculated to prevent common complications such as lung infection and blood clotting in the legs are the principal features of postoperative care. The prevention of postoperative complications with respect to the lungs is done by early but controlled moving of the patient in bed, the encouragement of coughing to expel fluids that have collected in the lung, suctioning of the air passage by use of rubber tubes, and finally, early walking about; walking, by improving the circulation in the legs, helps to prevent formation of blood clots, which can cause death if they migrate to the principal arteries in the lung. One of the most helpful modern postoperative aids is assisted respiration. This is done by machines that, by way of a tube in the bronchus, deliver much more adequate oxygen and ensure more adequate excretion of carbon dioxide in patients with crippled lungs.

Post-operative pulmonary complications

Chemotherapy and antibiotic therapy have a large place in obviously infective diseases. If the disease organism cannot be identified so that an accurate selection of effective drugs can be made, broad-spectrum antibiotics may be used; *i.e.*, antibiotics effective against a wide range of disease organisms.

The surgeon and his assistants are indispensable as monitors; by faithful attendance at the bedside, studying the patient and the charted data of his pulse, blood pressure, respirations, and laboratory data, they do much toward carrying patients through to recovery and rehabilitation. Mechanical monitoring has been discussed in an earlier section.

The increased range of surgical possibilities, not only from new operations but also from the acceptance for major surgery of persons so physiologically crippled as to have been considered unacceptable risks 20 years ago, has made monitoring by precision instruments a necessity today. This is also the case for severely accidentally injured patients who are in shock from blood loss and tissue damage or from overwhelming infection. In large modern hospitals, suitable equipment and trained personnel are available for such special postoperative care.

SPECIALIZATION IN SURGERY

The specialties in surgery are regulated by members of the profession rather than by state action in most parts of the world. In the United States, persons qualify as surgi-

cal specialists by taking graduate medical training of a specified type and for a specified period, by passing oral and written examinations in the specialty, and by presenting evidence of having satisfactorily performed a sufficient number of actual operations in the field. Boards have been established consisting of experts in a given specialty, who in periodic sessions meet candidates personally.

The Royal College of Surgeons of England preceded the United States in recognition of surgical specialties. There are royal colleges in England, Scotland, Ireland, and in many of the Commonwealth nations. In continental Europe specialists serve long apprenticeships in teaching hospitals. The result in Europe has been to limit the number of qualified surgeons. There is also a tendency to hold young men back, since professors often maintain control to an advanced age.

The first year of postgraduate training in the United States is called the internship. It involves a combination of internal medicine and surgery or may consist of exposures to the whole range of surgical specialties. The years beyond the internship are called the resident years, and the person serving them is assistant resident or, on occasion, becomes the chief resident. The chief resident in the United States is equivalent to the registrar in Great Britain and the *chef de clinique* in France.

The situation with respect to placement of a newly qualified surgical specialist in the United States differs from that in Europe. In the United States there are usually openings in community hospitals that are available as soon as a man has become qualified in a surgical specialty. There is a relative dearth of such hospital openings in England and in continental Europe.

The process of specializing within the field of medicine has gone furthest in the United States, where there are 31 medical specialties and subspecialties each governed by qualifying boards. Ten of these specialties involve surgery. A brief look at each of the surgical specialties follows.

General surgery. General surgery is the parent specialty and represents the field with the largest repertoire. It requires the longest training period, since it requires a limited exposure to all of the specialties and often one or two laboratory years in experimental surgery. Essentially, general surgery consists of doing the work that the practitioners of the more recent surgical specialties do not do. The heads of the departments of surgery are traditionally general surgeons, and rightly so. The appellation general surgery is an unfortunate one, for the public thinks of the general surgeon as some sort of general surgical practitioner, and thus general surgery seems a less prestigious pursuit. The principal anatomic and systemic spheres of general surgery are now the stomach and intestines; the breast and, particularly, cancer of the breast; and the peripheral blood vessels, as when the circulation to the extremities is poor. The general surgeon also deals with the endocrine (ductless) glands, such as the adrenal and thyroid glands; with tumours of the soft tissues; and with amputations. He sometimes deals with fractures, although this field has largely been taken over by orthopedic surgeons. When a person has been severely injured, a general surgeon is normally in charge of his case, although a number of other specialists may be involved. Since cancer is a common disease in the general surgeon's domain, and since cancer is more common among older people, much of the general surgeon's practice is among the elderly, although he does perform surgery on infants, also.

Functions of general surgeons

The management of patients at the two extremes of life is not a formal specialty but does require specialized knowledge, particularly with respect to children. Pediatric surgery is not so much a matter of understanding special operative techniques as of knowing the metabolic states in children and of having a background of experience with them. Many congenital anomalies require operations in the first few hours after birth, and the timing of the operation is an important consideration. Fortunately, the newborn tolerate anesthesia and surgery well, except in the case of the prematurely born.

Advanced age itself is not a contraindication to extensive surgery; rather, it is the physical state of the patient, his heart function, and his pulmonary status, in particular, that count.

Plastic surgery. Plastic surgery is largely concerned with the bodily surface and with reconstructive work of the face and exposed parts. The disfiguring facial deformities resulting from warfare gave the greatest impetus to the development of plastic surgery after World War I. Congenital anomalies such as harelip and cleft palate also are operated upon by the plastic surgeon. Removal of lesions on the face, eyelids, and ears with the greatest possible avoidance of functional and visible deformity is best done by these specialists. Skin grafting, particularly the resurfacing of burns, is now largely in the hands of the plastic surgeon. The invention of the dermatome, an electrically or manually propelled drum that can cut large pieces of skin of measured thickness, not only has made the art of the cutting of freehand grafts obsolete but produces larger grafts than free cutting ever could. Sufficient of the underlying skin is left at the donor site to ensure that the skin will grow back. This type of graft is commonly called split-thickness graft. Another development of grafting is the pedicle graft, in which the whole skin and the subcutaneous tissue can be gradually shifted to recipient sites from remote parts of the body where it is not needed. Reconstruction of the nose, eyelids, jaws, and ears is done commonly by the plastic surgeons in cooperation with general surgeons. Some plastic surgeons do the whole job in cancer surgery of the head and neck. The essence of technical plastic surgery is the careful planning of incisions so that they fall in the line of natural skin folds or lines, the use of fine suture material, and the early removal of exposed sutures so that the wound is held closed by buried sutures. Surgery of the hand, restoring function after tendon severance, is a specialty in itself, but much of the work is done by plastic surgeons. The public sometimes identifies plastic surgery with "beauty surgery," but most plastic surgeons are more involved in getting good cosmetic results after burns, removal of tumours, and reconstructive work than in cosmetic or "beauty" surgery for its own sake.

Functions of the plastic surgeon

Cosmetic surgery is largely done by surgeons who confine themselves to this field. Such surgeons usually have their own private hospital facilities, and their numbers are small.

Orthopedic surgery. Orthopedic surgery deals with congenital defects such as clubfoot, congenital dislocation of the hips, and scoliosis (curvature of the spine), but the largest field of orthopedics is in the treatment of fractures of the extremities, which may be either simple (in which there is no open wound) or compound (in which the skin has been penetrated either by the fractured bone or by some object). Compound fractures are more serious because disease organisms can enter and infect the tissues and bones. The basic principle of treating fractures of long bones is to bring the two broken ends into the proper position and hold them there until healing has occurred. Traction on the lower segment of the broken bone is necessary, and, while this can sometimes be done by manipulation, more usually it is necessary to exert traction by a system of weights and pulleys secured to the lower fragment. There are a variety of means of applying traction; usually a heavy wire or pin is driven through the lower fragment, and to this the weights are applied. The next important principle of fracture treatment is immobilization. Plaster-of-paris casts are most frequently used for immobilization of fractures of the extremities.

A direct operative approach to the treatment of fractures, commonly called open reduction with internal fixation, is also a common practice today. By incision the fracture is approached and manually reduced (placed in proper position); a metal plate is used to hold the fragments together. Screws are used to fix this plate to the two segments. Frequent X-ray examination is necessary to verify that the bones have remained in proper position and to observe the progress of healing. Bones heal slowly, and failure to heal is sometimes a problem. Another method of internal fixation of fractures of the long bones

is accomplished by driving a long steel rod down the marrow canal after alignment of the fragments.

Metallic prostheses (artificial parts) are commonly used in orthopedic surgery; useful hip joints can be made of metal. Sometimes metal plates used in the treatment of fractures are left in place permanently, but usually they are removed after healing. Orthopedists also treat back pain and examine patients for insurance claims.

Another field of treatment covered by orthopedic surgery is that of benign or malignant bone tumours. Osteomyelitis, a severe infection of the bone, is less common today than 30 years ago, probably because of the better hygiene now practiced in advanced countries. Since orthopedic surgery includes the treatment of joints, muscles, tendons, and ligaments, athletic injuries are mainly in the domain of this specialty.

Obstetrics and gynecology. Obstetrics and gynecology are combined specialties in the United States as far as board certification goes, but many surgeons confine their practice to just one of the two specialties. The basic operations in gynecology are of two general types. The first is repairing the injuries of childbirth. In all births the vagina is stretched and torn, with varying degrees of damage to the suspensory ligaments of the uterus and bladder; the uterine cervix may be injured, for example. Extirpative surgery of the uterus and ovaries for tumours, benign or malignant, is the second major operation. The diagnosis of the causes of sterility and subsequent treatment by medical means or surgery are a combined obstetrical and gynecological effort that has made significant progress since World War II.

Repair of childbirth injuries

Gynecologists practice medical endocrinology when they treat symptoms of the menopause and abnormal uterine bleeding and when they control ovulation to prevent pregnancy.

Gynecologists make routine examinations of cervical and vaginal secretions to detect cancer of the uterus, fairly common among relatively young women; if these tests suggest the possibility of cancer, specimens of tissue may be taken from the cervix for microscopic examination. These and other periodic examinations and minor office surgery take up much of the gynecologist's and the obstetrician's time. Uterine cancer is often treated by a combined use of radiation and extirpative surgery, surgical removal of cancerous tissue. The application of radium to the cervix for cancer must be done accurately and the dosage carefully calculated by the combined efforts of the radiologist and the surgeon. Cancer of the cervix is more commonly treated by radiation alone, while extirpation is the preferred treatment of cancer of the body of the uterus.

Terminations of pregnancy for social reasons rather than medical are now a substantial segment of gynecologic practice in certain European countries and in some states of the United States (New York State, for example).

The obstetrician's tasks are to bring a woman through pregnancy, to deliver her child, and to give her adequate postnatal care. Prenatal care is important in the detection of possible abnormalities that may complicate delivery, and fetal abnormalities can, in some cases, be diagnosed early in pregnancy. Complications such as premature separation of the placenta require prompt surgery. The recognition of acute abdominal complications, such as intestinal obstruction and appendicitis, is an extremely important aspect of prenatal supervision, since the symptoms of such diseases may simulate labour.

One surgical operation performed by obstetricians, cesarean section, in which the infant is delivered through an incision in the mother's uterus and abdominal wall, is considered necessary in far fewer situations than formerly, and the circumcision of male infants at birth as a hygienic procedure is somewhat less routine than formerly.

Another change in the area of obstetrics has been a shift away from specialized lying-in hospitals to general hospitals, where the services of specialists in all fields are readily available and auxiliary services are complete.

Ophthalmic surgery. Ophthalmologists are trained, for a period of at least three years after an internship, in the medical and surgical aspects of the treatment of diseases and injuries of the eye. Although there is a large field of nonsurgical eye treatment, the surgical field has expanded of late. The correction of certain types of squint (conditions in which the lines of vision of the two eyes are not parallel), the extraction of opaque lenses (cataracts), the relief of high pressure in the eyeball (glaucoma), and reattachment of the retina after it has become detached are among the common operations of the specialty.

New developments in eye surgery include the use of ultrasound in diagnosis. Ultrasound has been tried experimentally for dissolution of cataracts also. The laser is now commonly employed in treating diseases that lead to bleeding in the eye. Occasionally, the laser is used for fixation of detached retinas. Ninety percent of persons who suffer detachment of the retina can now expect restoration of eyesight to some degree. Cryosurgery is often used for lens extraction, treatment of detached retina, and treatment of glaucoma.

Ultrasound and laser in diagnosis and treatment

The transplantation of the cornea from a dead donor to a live recipient is one of the oldest, most successful, and most common of transplantation procedures. The cornea must be implanted soon after death of the donor. Transplantations of the cornea are useful for hereditary corneal diseases of the eye and for corneal burns.

Experiments are being made in the use of synthetic materials to replace corneas. Another new development in eye surgery is the use of special glues as a substitute for sutures. These glues have been highly successful in some instances.

Colon and rectal surgery. Colon and rectal surgery is recognized as a specialty, but only a few surgeons confine their practice to the specialty, and few training centres offer it. Much of the surgery on the colon and rectum is done by general surgeons.

Urology. Urology is concerned with the treatment of diseases of the urinary tract and, in men, of the genital apparatus. The urinary tract consists of the kidneys, the bladder, the tubes between the kidneys and the bladder—the ureters—and the urethra, the tube carrying urine from the bladder to the external opening. Most of the urologist's patients are male, for two reasons: (1) the urinary tract in females is also treated by gynecologists, and (2) much of the urologist's work has to do with the prostate gland, which encircles the male urethra close to the juncture between the urethra and the bladder. The prostate gland is often the site of cancer; even more frequently, it enlarges in middle and old age and encroaches on the urethra, causing partial or complete obstruction to the flow of urine. Prostate enlargement is treated either by total excision of the prostate or by reaming a wider passageway through it. The urologist also may operate to remove stones that have formed in the urinary tract. Surgery performed by the urologist also includes operations for cancers of the kidneys, bladder, and testicles.

Much of the urologist's time is devoted to diagnosis—by X-rays, by use of the cystoscope, and by biopsy. The cystoscope is a tubelike instrument equipped with an electric light on its end. By introduction of the instrument through the urethra, the urologist is able to view the interior of the bladder. In biopsy a piece of tissue is removed for examination under the microscope.

Otolaryngology. Otolaryngologic surgery is performed in the area of the ear, nose, and throat. Previously it was concerned almost entirely with operations upon the tonsils, the sinuses, and the mastoid processes, the bulges in the skull just behind the ears. Much of the surgery for infections in these structures has become unnecessary since the advent of antibiotic therapy. Benign and malignant tumours of the upper air passages, particularly of the larynx, are currently dealt with by the otolaryngological surgeons. Laryngectomy, removal of the voice box for cancer, and the subsequent training of patients to talk are noteworthy accomplishments in the field. More recently, conservative operations for cancer with preservation of the voice have become feasible.

The remarkable progress in surgery for deafness that has occurred since the middle 1950s is one of the most

prominent accomplishments of otolaryngology. Any type of deafness that does not involve the auditory nerve may be within the scope of otolaryngological surgery. Microsurgery has made this ear surgery possible. Cryosurgery has a limited use for palliation of malignant tumours that are inoperable and incurable. Ultrasound, as well as decompression operations, has been used in the control of Menière's disease, a disease of the membranous labyrinth that results in attacks of deafness, dizziness, and other symptoms. At present, eardrums and some of the bones of the middle ear are being transplanted from cadavers with remarkable improvement in the hearing of the recipient. An area now receiving experimental attention is the implantation of suitable electrodes in the inner ear to facilitate the conduction of sounds from external hearing aids.

Neurological surgery. Injury to the brain from skull fractures is the most common disease encountered in the area of neurosurgery. Other lesions treated by the neurosurgeon are tumours of the brain and spinal cord, diseases and injuries of the peripheral nerves, and ruptured intervertebral discs that cause pressure on the spinal cord. The majority of the tumours of the brain operated upon by the neurosurgeon are malignant, but palliation is often obtained. A constantly enlarging field is vascular surgery of the brain. Bleeding from congenital aneurysms (weak spots in blood-vessel walls with resultant dilation or bursting of the vessels) or in patients with high blood pressure frequently requires operation. The great developments of neuroradiology have furthered this field of vascular surgery, and indeed angiography (X-rays of blood vessels) of the cerebral vessels accomplished by rapid X-ray photography after the injection of radio-opaque substance into the large blood vessels of the neck (carotid arteries) has aided greatly in the diagnosis of lesions of the brain.

Neurologic diagnosis

The cerebral vessels have a normal fixed anatomic pattern, and distortion of the pattern by the presence of brain tumours or other lesions shows in the X-rays. Measurement of blood flow is another new, useful diagnostic technique that has greatly added to the accuracy of diagnosis. Most important, stereotactic techniques have been developed by means of which instruments (such as hollow probes) can be accurately placed within the brain. By the use of such techniques, small areas of the brain that are causing abnormal cerebral activity can be destroyed, without the disruption that would be caused by the conventional methods. At present, one of the principal benefits of stereotactic surgery of the brain is the knowledge of brain function that is accumulating, and it is in this experimental area that the future of neurosurgery lies. Microsurgery is being used in the removal of small tumours of the cranial nerves and in surgery on small blood vessels. It promises to have a wider application. The overall operative mortality rate is about 5 percent for all operations on the skull; these cases include a large number of patients who have received irreversible brain damage from accidents. Damage to the brain with or without skull fracture is an associated or principal cause of death in over half of automotive accidents.

Thoracic surgery. Thoracic surgery, an offshoot of general surgery, became broad in scope in the late 1940s. Earlier, the treatment of pulmonary tuberculosis by collapsing the lung, so as to rest it, and emergency operations for bullet and stab wounds of the lung and heart were the principal repertoire. The introduction of the mycin group of drugs (*e.g.*, streptomycin and others) to all practical purposes took pulmonary tuberculosis out of the surgical realm.

Modern thoracic surgery received a great benefit from the development of the endotracheal type of anesthesia. When the chest is open, the lung normally collapses and is unable to oxygenate the blood, a process that is essential to life. When a tube is put in the trachea (the windpipe), pressure from a machine that delivers both anesthetic and oxygen can keep the lung inflated and functioning.

Esophageal surgery, especially resection of the esophagus for cancer, became another new item in thoracic surgery. Removal of all or part of the lung for cancer became successful technically in the 1930s and is now a common procedure. In the 1950s surgery of the heart and the great vessels emerged. It is now a principal field in some clinics; indeed, some thoracic surgeons do nothing else. This work was made possible by a thorough study of congenital anomalies of the heart and the physiologic derangement of such abnormalities. Next, acquired heart disease that results in damaged heart valves was attacked. Cardioangiography—visualization of the heart chambers by X-ray photography—and cardiocatheterization—running small tubes all the way from an arm artery into the heart—have made diagnosis of heart diseases and estimations of operability possible.

A heart–lung machine has been developed that provides artificial oxygenation and pumping of the blood. This has made it possible to open the heart so that abnormalities of the valves and walls of the heart chambers can be corrected. In 1961 replacement of defective valves by metallic and plastic materials was introduced.

An outgrowth of the treatment of congenital heart disease has been the invention of pacemakers, which control the beat of the heart. These electric pacemakers are buried in the patient's body.

Another new and promising field in cardiac surgery is in coronary heart disease, in which heart damage results from inadequate blood flow through the coronary arteries to the heart muscle. The defects in the coronary arteries can now be bypassed with vein grafts taken from the leg (see CARDIOVASCULAR SYSTEM DISEASES AND DISORDERS).

BIBLIOGRAPHY. C. BERNARD, *Leçons sur les Phénomènes de la Vie*, 2 vol. (1878–79), earliest work on the internal environment of the body in relation to homeostasis; C.A.T. BILLROTH, *Lectures on Surgical Pathology and Therapeutics*, 2 vol. (Eng. trans. 1877–78), famous textbook of surgery in the 19th century at the height of German influence; W.B. CANNON, *The Wisdom of the Body* (1932), discussion of homeostasis and the body's protective reaction to injury; E.D. CHURCHILL, "American Surgeon, A. U. S.," *Surg. Gynec. Obstet.*, 84:529–530 (1947), discussion of surgery in World War II; H. CUSHING, "Neurological Surgery and the War," *Boston Med. Surg. J.*, 181:549-552 (1919), account of neurological surgery in World War I; *Christopher's Textbook of Surgery*, 8th ed. by L. DAVIS (1964), a modern popular textbook of surgery, including an essay on the history of surgery by C.S. WELCH; *Directory of Approved Internships and Residencies* (annual), information by the American Medical Association on approved internships and residencies in American hospitals together with a description of all specialties and requirements for certification; F. GARRISON, *An Introduction to the History of Medicine*, 4th ed. (1929), a general text; W.S. HALSTEAD, *Surgical Papers*, 2 vol. (1924), technically important papers by one of the great names in American surgery; J. HILTON, *On Rest and Pain*, 2nd ed. (1879), lectures before the Royal College of Surgeons on the value of rest in the treatment of pain and injury; J.F. PALMER (ed.), *The Complete Works of John Hunter*, 4 vol. (1835–37), collection including works on physiology and the reaction of the body to injury; J.H. JACOBSON, JR., "Microsurgery," *Current Problems in Surgery* (February 1971), text by the foremost innovator and experimenter in microsurgery; K. LANDSTEINER, "Individual Differences in Human Blood," *Science*, 73:403–409 (1931), description of blood group and types; J. LISTER, "On the Antiseptic Principle in the Practice of Surgery," *Lancet*, 2:353–356 (1876), good description of his method of antiseptic surgery; G. KEYNES (ed.), *The Apologie and Treatise of Ambroise Paré: Containing the Voyages Made in Divers Places with Many of His Writings upon Surgery* (1952), presentation of Paré's philosophy of surgery and the treatment of wounds; E.J. PULASKI, *Surgical Infections: Pryphylaxis, Treatment, Antibiotic Therapy* (1954), good general treatise on surgical infections; H. SELYE, *The Physiology and Pathology of Exposure to Stress: A Treatise Based on the Concepts of the General-Adaptation Syndrome and the Diseases of Adaptation* (1950), description of the reaction of the body to injury; *The Surgery of Theodoric* (1955), translation of Theodoric's 13th-century text that describes his rational treatment for that period of history; U.S. ARMY MEDICAL DEPT., RESEARCH AND GRADUATE SCHOOL, *Symposium on Treatment of Trauma in the Armed Forces* (1952); C.S. WELCH and S.R. POWERS, JR., *The Essence of Surgery* (1958), text on the general principles of surgical treatment, including a history of surgery.

(C.S.W.)

Surinam

Surinam, formerly Dutch Guiana, is a self-governing territory of the kingdom of The Netherlands located on the north central coast of South America between Guyana on the west and French Guiana on the east. The territory covers an area of 63,251 square miles (163,820 square kilometres), extending southward about 310 miles from the 215-mile-long Atlantic coastline to Brazil. Of the territory's population of about 385,000, 102,000 persons live in Paramaribo, the capital district.

Historically, Surinam was a tropical colony with a plantation economy, at first based on slave and then contract labour. In modern times, however, small farms have replaced the plantations, and in the early 1970s Surinam had become one of the world's largest producers of bauxite, aluminum ore. The aluminum industry is fully integrated, processing bauxite into alumina and aluminum. These products account for 92 percent of the country's exports. Other products exported are rice, bananas, sugar, plywood and particle board, citrus fruit, shrimps, and lumber. Surinam, through its ties with The Netherlands, is an associate member of the European Common Market. The country is connected, within the federation of the kingdom of The Netherlands, to such international organizations as the International Monetary Fund, the World Bank, and GATT (General Agreement on Tariffs and Trade).

THE NATURAL AND HUMAN LANDSCAPE

Natural environment. *Relief features.* Surinam consists of a narrow, sea-level coastal zone in which diking and draining are necessary for cultivation, a somewhat higher belt of scattered savannas inland from the coastal zone, and hills rising to low mountains in the south. Geologically, Surinam is a part of the Precambrian Guiana Shield (sometimes also known as the Guiana Highlands), which is more than 570,000,000 years old. In the north the shield is covered by younger sediments, with their largest extension to the west. The greater part of the shield is rather flat, sloping toward the sea. To the south the terrain is slightly undulating, or hilly, with rather low mountain chains (Acarai and Toemak Hoemak Gebergte [Tumuc-Humac Mountains]). Other mountainous regions form the divides between several river systems. The Julianatop in the Wilhelmina Gebergte in the southwest is the highest elevation in Surinam, 4,218 feet (1,286 metres). The rivers have numerous rapids and falls, and along most of them river terraces, or remnants, form conspicuous features.

The natural landscape

The sedimentary deposits in the north produce several distinct landscapes. The flat to slightly undulating savanna belt consists of brown and white sands of Pliocene age (2,500,000 to 7,000,000 years old). To the north the belt is bordered by the old coastal plain, which consists of sands and clays of Pleistocene age (10,000 to 2,500,000 years old). Old rivers and gullies now filled with clay and peat form swampy areas surrounded by slightly higher terrain.

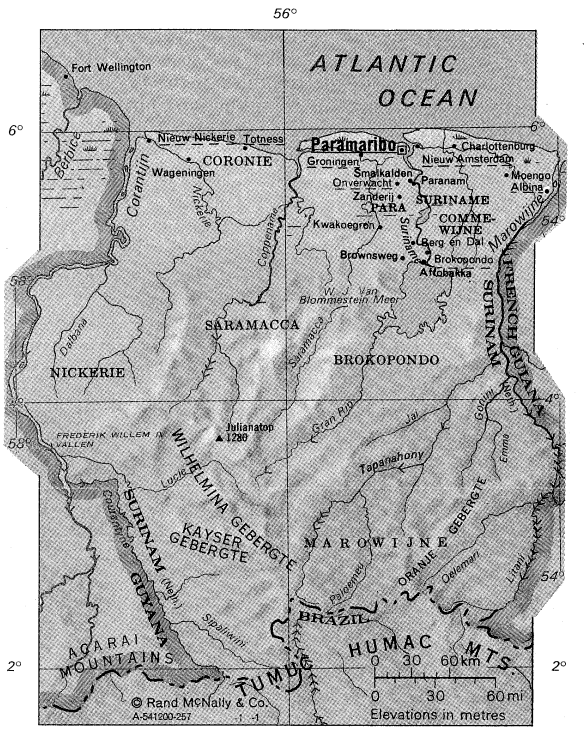
The marine deposits along the ocean form the young coastal plain. The clays, originally supplied by the Amazon River, bear extensive swamps interrupted by old beach ridges, which run mainly east to west and become rarer to the west. Sandy beaches as well as large mud banks tend to creep westward along the coast under the influence of winds and currents. Because of their huge burden of mud, only the larger rivers, with their strong flow, generally empty straight into the ocean, the winds and currents of which deflect the smaller rivers to the west. The seven largest rivers, from west to east, are the Corantijn, the Nickerie, the Coppename, the Saramacca, the Suriname, the Commewijne, and the Maroni (Marowijne).

Climate. Surinam's tropical climate, under the influence of the northeast trade winds, remains warm and moist throughout the year. The day temperature range is between 73° F (23° C) and 88° F (31° C) with an annual mean of 81° F (27° C). The diurnal variation is slight, only about 7° F (4° C).

Temperature and rainfall

Relative humidity is high for most of the year, averaging about 80 percent. Rainfall is highest in the mountains of the interior, and in the coastal areas it varies from east to west, from an annual average of 95 inches in Paramaribo to 76 inches in the west. There are four seasons—a minor rainy season from November to February, a minor dry season from February to April, a major rainy season from April to August, and a major dry season from August to November. In the coastal areas the winds are rather strong but their intensity lessens toward the interior of the country. Surinam lies outside the Caribbean hurricane area.

Vegetation. About 90 percent of Surinam is covered with forests with more than 2,000 varieties of trees. Tropical rain forest covers large areas, but much of it is jungle and inaccessible; and a large percentage of the trees are not commercially useful. Mangrove and swamp forests,



SURINAM

MAP INDEX

Political subdivisions

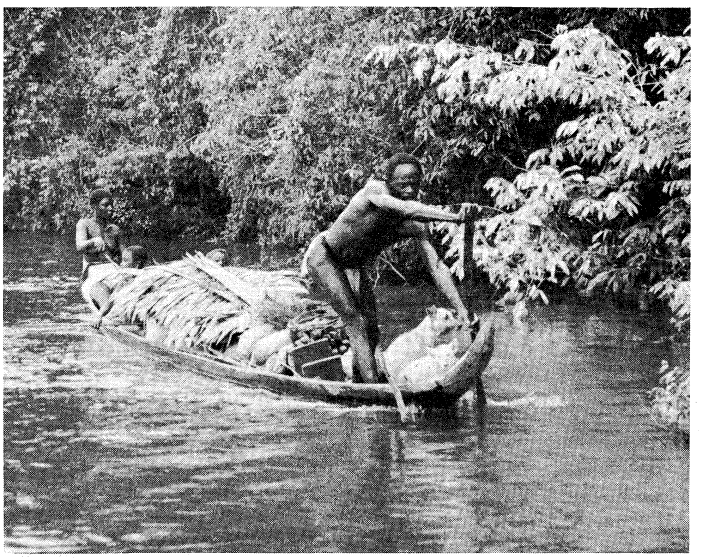
Brokopondo 4·30n 56·30w
Commewijne 5·25n 54·45w
Coronie 5·55n 56·20w
Marowijne 4·00n 56·00w
Nickerie 4·00n 57·00w
Para 5·30n 55·15w
Saramacca 5·00n 56·00w
Suriname 5·30n 55·00w

Cities and towns

Affobakka 5·01n 55·05w
Albina 5·30n 54·03w
Berg en Dal 5·05n 55·05w
Brokopondo 5·03n 54·59w
Brownsweg 5·02n 55·11w
Charlottenburg 5·51n 54·46w
Groningen 5·48n 55·28w
Kwakoegron 5·12n 55·22w
Moengo 5·37n 54·24w
Nieuw Amsterdam 5·53n 55·05w
Nieuw Nickerie 5·57n 56·59w
Onverwacht 5·35n 55·10w
Paramaribo 5·50n 55·10w
Paranam 5·35n 55·05w
Smalkalden 5·36n 55·10w
Totness 5·53n 56·19w
Wageningen 5·50n 56·50w
Zanderij 5·27n 55·12w

Physical features and points of interest

Atlantic Ocean 6·30n 55·00w
Coppename, *river* 5·48n 55·55w
Corantijn, *river* 5·55n 57·05w
Dalbana, *river* 5·01n 57·20w
Emma, *river* 3·50n 54·30w
Frederik Willem IV Vallen, *waterfall* 3·30n 57·40w
Gonini, *river* 4·20n 54·24w
Gran Rio, *river* 4·01n 55·30w
Jai, *river* 3·58n 54·50w
Julianatop, *mountain* 3·40n 56·30w
Kayser Gebergte, *mountains* 3·00n 56·30w
Litani, *river* 3·40n 54·00w
Lucie, *river* 3·35n 57·38w
Marowijne, *river* 5·45n 53·58w
Nickerie, *river* 5·58n 57·00w
Oelmari, *river* 3·24n 54·08w
Oranje Gebergte, *mountains* 3·15n 54·50w
Paloemeu, *river* 3·21n 55·26w
Saramacca, *river* 5·50n 55·55w
Sipaliwini, *river* 2·20n 56·48w
Suriname, *river* 5·55n 55·10w
Tapanahony, *river* 4·20n 54·25w
Wilhelmina Gebergte, *mountains* 3·45n 56·30w
W. J. van Bloomestein Meer, *lake* 4·45n 55·05w
Wonotobo Fall, *waterfall* 4·22n 57·55w



Bush Negroes in a dugout canoe on the Maroni River.
Carl Frank—Black Star

as well as grasses grow along the seacoasts, and various kinds of palms, including coconut, on slightly drier land. Rich stands of baboen (*Virola surinamensis*) in the coastal strip supply the thriving plywood industry. Other commercially important species include greenheart (*Tabebuia serratifolia*, a tree from 30-60 feet high, the wood of which is used to make cogs and rollers in coffee and sugar factories) and wallaba (a valuable timber tree with reddish-brown wood) in the north; mora (a tall, half-evergreen tree, with wood resistant to dry rot and termite injury), and crabwood on swampy sites; and balata and other latex trees, as well as trees yielding handsome cabinet woods in the rain forests and highland forests.

Animal life. There are about 150 species of mammals, including the howler and other varieties of monkey, as well as sloths, anteaters, and armadillos. The largest land mammal is the tapir; deer and pakiras (wild pigs) are hunted by forest dwellers. Jaguar, puma, and ocelot also occur. Rodents include the capybara (the largest living rodent); there are also numerous species of bats. More than 600 different bird species have been observed. Among the reptiles are tortoises and various species of lizard, including the iguana, and caimans. Snakes include the poisonous rattlesnake, the bushmaster, and snakes that strangle their prey, such as the boa constrictor. Among the large variety of frogs and toads are some of incredible size. A large group of insects includes termites and exceptionally large parasol ants, which build canal systems underground.

The landscape under human settlement. *Rural settlement.* With the abolition of slavery in 1863 and later, after having completed their contractual obligations as indentured labourers in the earlier part of the 20th century, a large number of plantation workers settled in the rural areas as small farmers. Types of settlements include plantations exploited by Western methods, government and private settlements of small tenant farmers, and settlements of small farmers who own the land on which they live.

Plantations and small farms

Some small farmers also live at rather great distances apart on their own smallholdings or on rented grounds, especially along the upper parts of the rivers. The economy of the Coronie District, which is comprised of a narrow 50-mile strip along the seacoast west of the Coppename River, and has a predominantly Creole population (of Negroes or persons of mixed origin) is based on the cultivation of coconuts and the production of coconut oil. In all other districts the most important product of the small farms is rice. Other products grown include coffee, citrus fruit, bananas, vegetables, and peanuts. Many small farmers own some cattle. Except in the Coronie District, Asian-Indian and Indonesian farmers predominate.

Urban settlement. By far the largest city, Paramaribo, the chief port, lies about 17 miles inland on the Suriname Rivier and is the seat of government and the main commercial centre. The city is neat and clean with many tree-lined streets; many of its beautiful wooden structures date back to 1835. The marketplace forms a permanent trading centre for tropical produce, Indian pottery, wood carvings, fish, meat products, and other items. In Paramaribo, the early afternoon is siesta time, and many stores and businesses are closed in the afternoons from 1:00 to 4 PM.

Nieuw-Nickerie, the capital of the Nickerie District on the Nickerie Rivier close to the western border with Guyana, has a population of 7,200 and is the second largest city; its port mainly handles rice and banana exports. Many banks and business establishments have branches in the town; there is also a well-equipped hospital. The life-style of the city is much the same as in Paramaribo; the district, however, has a predominantly Asian-Indian population.

THE COUNTRY AND ITS PEOPLE

History. Until the 15th century the only inhabitants of Surinam were Carib, Arawak, and Warrow Indians. Another tribe, the Surinen, who inhabited the country at an earlier time but who were driven away by the Caribs, is considered to be the source of the name Surinam. In 1499, Alonso de Ojeda, a Spanish lieutenant serving the Italian navigator Amerigo Vespucci, landed on the northeast coast of South America, called Guiana by the Indians. In 1593 the Spanish officially took possession of the coast. No actual settlement was attempted, and during the next century several Dutch, British, and French expeditions visited the area.

Derivation of the name Surinam

The first successful settlement in Surinam was founded in 1651 by an Englishman, Francis Lord Willoughby of Parham, Suffolk, who welcomed settlers from other West Indian and South American colonies. Among these were numbers of Jews, mainly from a Dutch colony on the Berbice estuary, west of Surinam, to which they had fled in 1639 from religious persecution in Brazil. They brought capital and skills, and in 1665 erected what is said to be the first synagogue in the Western Hemisphere. The colony developed a prosperous plantation economy based on slave labour imported from Africa, producing sugar, cotton, cocoa, and coffee.

In 1667, after the second Anglo-Dutch War, Surinam was ceded to the Dutch in exchange for Nieuw-Amsterdam (New York). The following century was one of economic decline and political turmoil. Many slaves rebelled and fled to the jungle; banding together, they conducted raids on the plantations. Military expeditions against them resulted in repeated defeats for the army and in the exhaustion of financial resources. In the meantime, as a result of European wars and treaties, the sovereignty of the country shifted between England, France, and the United Provinces of the Netherlands until the Dutch finally gained permanent control under the terms of the Treaty of Paris (1816). When the abolition of slavery in 1863 produced a critical labour shortage, Surinam began to import contract workers (indentured labourers) from China, Java, and India. Many of them stayed—the Chinese mainly as shopkeepers in the rural areas and the Indians and Javanese as independent small farmers.

In 1900, 90 percent of all agricultural products were grown on plantations and 10 percent on small farms. These figures were reversed in the subsequent half century, during which the plantation crops of cocoa, coffee, cotton, and sugar were replaced by rice, bananas, and other fruit. After World War II, the Dutch diverted some of their financial and technical resources from the East Indies to Surinam. In 1950 Surinam obtained some measure of autonomy, and in 1954 it became a self-governing part of the Kingdom of The Netherlands.

Agricultural development in the 20th-century

People and population. *Linguistic, ethnic, and religious groups.* The official language is Dutch, but English is very widely spoken. A lingua franca, known as Sranan (Taki-Taki), is a mixture of English, Dutch, French,

Surinam, Area and Population

	area		population	
	sq mi	sq km	1964 census	1971 census
Districts				
Brokopondo	8,279	21,442	12,000	18,000
Commewijne	1,587	4,111	21,000	19,000
Coronie	626	1,622	4,000	3,000
Marowijne	19,101	49,471	23,000	26,000
Nickerie*	23,598	61,119	30,000	35,000
Para†			15,000	17,000
Saramacca	9,043	23,421	12,000	14,000
Suriname†	1,010	2,617	97,000	152,000
Town district				
Paramaribo	7	17	111,000	102,000
Total Surinam*	63,251	163,820	324,000‡	385,000‡

*The disputed area (6,809 sq mi [17,635 sq km]) adjoining Nickerie district is not included. †Para district was created out of a part of Suriname District on September 8, 1966. Separate area figures for Suriname and Para districts are not available. ‡Figures do not add to total given because of rounding.
Source: Official government figures.

Spanish, and even Hebrew; it is used in interracial communication. Among themselves the Bush Negroes (*i.e.*, those living in the forests) speak variations of African languages; each American Indian tribe has its own language. Hindi, Javanese, Chinese, Lebanese, and other languages are spoken by the various racial groups living in their own communities.

Heterogeneity of the population

One of the most striking characteristics of Surinam is the heterogeneous ethnic composition of its population. The native Indians, or Amerindians, carry on their primordial life almost entirely untouched by outside influence, whereas the Bush Negroes continue their way of life much as their ancestors did in Africa. The Creoles, Asian Indians, and Indonesians, who comprise the three main ethnic groups, maintain their own identities to a large extent. While each retains many of its separate characteristics, the groups nevertheless intermingle in commercial and social activities. There are no minority problems, and assimilation is steadily taking place.

There is complete religious freedom. The Amerindians and Bush Negroes, in so far as they have been converted, as well as the Creoles, are predominantly Christian (either Protestant or Roman Catholic). The majority of the Asian Indians profess Hinduism, and the rest Islām, to which the Indonesians also generally adhere; a small number of both groups is also Christian. Well-known Protestant congregations include the Moravian Brothers Mission, the Netherlands Reformed Congregation, and the Evangelical Lutheran Church. Surinam also has a small Jewish community.

Demography. The 1971 census puts Surinam's total population at about 385,000, consisting of Hindustanis (about 37 percent), Creoles (about 31 percent), Indonesians (about 15 percent), Bush Negroes (about 10 percent), Amerindians (about 3 percent), Chinese (about 2 percent), Europeans (about 1 percent), and others (about 1 percent). In the early 1970s about 45 percent of the population was below 15 years of age; 68 percent was below 30; 6 percent was above 60. The birth rate was 46 per 1,000, while the death rate amounted to 9 per 1,000. The yearly increase is from 3.5 to 4 percent.

ECONOMIC, SOCIAL, AND CULTURAL LIFE

National economy. *Agriculture.* Agriculture is conducted mainly on the coastal plain by a large number of small, independent farmers. The growing of rice occupies about 89,000 acres or 75 percent of the total cultivated area; the largest rice-producing enterprise is the fully mechanized Wageningen project, financed by The Netherlands government, in the district of Nickerie. The sowing periods for rice are from the end of March to May for the major crop and from the middle of October to November for the second crop, each with a growing period of about 142 days. Surinam thus is self-sufficient in rice, its major staple food crop. It also produces sufficient quantities of bananas, citrus fruit, and coconuts for export, in addition to domestic use. As a result of a five-year development plan (1967–71) the production of bananas for export to Europe increased considerably. Citrus production (oranges and grapefruit) is an important factor in Surinam's agriculture, with the bulk of the export going to The Netherlands. Coconut oil is the other principal agricultural export.

Principal crops

Crops grown for domestic consumption include coffee, cocoa, corn, peanuts, soybeans, pulses, sweet potatoes, tomatoes, cabbage, plantains, cassava, and fruits. Sugar, once the principal commodity in Surinam, is now produced on only two plantations.

Industry and finance. The mining of bauxite is the leading nonagricultural economic activity in the country, the bauxite-alumina-aluminum industry accounting for 92 percent by value of 1970 exports. Bauxite is produced at plants at Smalkalden, Moengo, and Paranam; in addition there is an aluminum smelter and an alumina refinery at Paranam. (Alumina, or aluminum oxide, can be reduced by electrolysis to the metal aluminum.) Production of alumina is about 1,000,000 tons annually, and that of aluminum, exceeds 50,000 tons. Surinam is one of the few countries in the world with a fully integrated aluminum industry.

Alumina and aluminum production

To provide the enormous amounts of electricity it requires, the Suriname Rivier has been dammed to create W.J. van Blommestein Meer (lake; 600 square miles), which provides hydroelectric power for installations at Affobakka. Additional large deposits of bauxite in western Surinam are being developed for exploitation.

Surinam has great timber resources, even though many trees are not exploitable. The national forests of Surinam cover more than 37,000,000 acres. The most important potential lies in the so-called high forest. It has the shape of a belt extending from the Maroni to the Corantijn Rivier, running south of the sparsely wooded savannahs and north of the rapids in the rivers and hills region. A large plywood factory produces an excellent product and also manufactures particle board.

Small industries include a shrimp-processing plant, a brewery; shoe, clothing, and paint factories; a cement plant; a flour mill; a packing plant for citrus fruit; and other installations.

The country, with its relatively large export and import trade, has enjoyed a balance-of-trade surplus for many years. Its main trading partners are the United States and The Netherlands. The country has received considerable financial aid from The Netherlands under a ten-year plan (agreed to in 1954) and under subsequent five-year plans; development aid has also come from the European Common Market.

The Centrale Bank van Suriname was established in 1957 as a bank of issue. Surinam also has three commer-

Carl Frank—Photo Researchers



Planting rice, the leading agricultural product of Surinam.

cial banks with branches throughout the country. In addition, some specialized banking institutions handle investing, mortgages, development, credit, and postal savings. A five- to ten-year tax holiday is granted to new enterprises, and other incentives include access to the European Economic Community; all are offered to encourage foreign investment.

The discovery of large bauxite deposits in western Surinam and the plans to develop that region in the 1970s were seen as hopeful signs for the future. The continuation of offshore drillings for natural gas and oil and indications of several mineral deposits, such as copper and nickel, hold promise of accelerated economic growth. An effort is being made to promote tourism. The government has built guesthouses in the interior, some with air-conditioned rooms. Paramaribo has several good hotels and pensions, as well as some first class hotels with casinos and entertainment facilities.

Transportation. Most internal transport is by water over about 750 miles of navigable rivers and canals. Surinam's national steamship line, the Surinam Navigation Company, provides regular service along the coast and on inland waterways, as well as between Paramaribo and New Orleans. Paramaribo, the main port, has custom facilities and modern loading equipment. Other important ports are at Moengo, Paranam, and Smalkalden, all of which handle bauxite.

Roads and highways

There are about 1,500 miles of all-weather roads (300 miles of which are paved) and 810 miles of rural roads. The west–east highway stretches from the Corantijn Rivier, the western border with Guyana, to the Maroni, the eastern border with French Guiana. Excellent highways lead from Paramaribo to Paranam, the centre of the aluminum industry, and from Paramaribo to the international airport at Zanderij. A local airline also operates flights abroad and throughout the country; there are more than 30 airstrips in the interior. A small railroad connects Onverwacht with Brownsweg. Paramaribo's automatic telephone service covers most of the coastal area. International Telephone and Telegraph communications link Surinam with the United States, western Europe, and the Caribbean.

Administration and social conditions. *Government.* Surinam is an autonomous part of the Kingdom of The Netherlands, which also comprises The Netherlands itself and the Netherlands Antilles. This status was established in a constitutional order known as "Het Statuut" in 1954.

The monarch is the head of state and is represented by an appointed governor who presides over the Advisory Council (Raad van Advies) of at least five members, appointed by the governor. Surinam is a parliamentary democracy, with universal suffrage for men and women who have reached the age of 23. By a combination of direct and proportional representation, the voters elect 39 members to the Parliament (Staten van Suriname) for a four-year period. The political party with the majority of seats in Parliament submits to the governor the names of persons to be appointed to the 13-member Executive Council of Ministers, which is headed by a prime minister. Each minister is responsible to the Parliament. The country is divided into nine districts, which are administered by district commissioners who are appointed by the government.

Social conditions. The working population numbers about 100,000, about 26 percent of whom are employed by the government. Agriculture and forestry engage about 24 percent; industry and construction, 17 percent; commerce, banking and insurance, and transportation 12 percent; and mining, 7 percent. The government promotes the hiring of local persons, but grants permission for the engagement of specialists from abroad when they are not available in Surinam. More and more women are being hired by small industries.

The labour union movement in Surinam is young and is responsible in its dealings with management. For the settlement of labour disputes, the law provides for an advisory intermediary council consisting of government representatives, management, and labour. The literacy rate of the country is 70 percent (90 percent in the capital).

Education and public health

Education is compulsory for children up to 12 years of age, and all primary education is free. Denominational schools are subsidized by the government. Of Surinam's approximately 400 schools, 40 educate Amerindians and Bush Negroes living in the interior. In addition, there are high schools, junior colleges, a teacher's college, a university with medical and judicial faculties affiliated with the state University of Leiden in The Netherlands, technical schools, agricultural schools, and an agricultural research centre.

Public health standards are good. Paramaribo has five well-equipped hospitals, and there are other hospitals and various clinics throughout the country.

Cultural life and institutions. Cultural activity generally centres in Paramaribo. Surinam Cultural Centre is active in a variety of fields; it maintains an auditorium with about 400 seats, a reading room, and a library. Theatrical groups perform plays, and classes in ballet, painting, and drawing are held there. The centre also has a film service, which exhibits films in rural areas. The centre also houses a school of music, a chamber music society, a youth orchestra, several choirs and a philharmonic orchestra that has acquired a reputation outside of Surinam. There are several museums in the country, including an open-air one.

Of the country's sports, soccer is the most popular; others include tennis, basketball, volleyball, swimming, and cricket. A modern stadium in Paramaribo is primarily used for soccer, but it has facilities for track-and-field events and bicycle racing.

The one television station is government controlled. Of the five radio broadcasting stations, three are in Paramaribo and the other two in the district of Nickerie. Five daily and three nondaily newspapers and several periodicals are published. Surinam's moviegoers have 16 theatres they can attend, eight of them in Paramaribo, to see films shown from the United States, Europe, and India.

FUTURE PROSPECTS

The development of hydroelectric power resources has resulted in a considerable increase in the gross national product. The discovery and development of new bauxite deposits, coupled with plans for a new hydroelectric power project in western Surinam, point toward continuing activity in the mining sector. Increased exports of rice and bananas, as well as further government efforts to stimulate agriculture, are encouraging signs for the future. If Surinam can maintain the political stability that has hitherto been one of its characteristics and if it succeeds in attracting foreign investors to help develop its resources, the future of the country would appear to be bright.

BIBLIOGRAPHY. The SALZGITTER INDUSTRIEBAN GESELLSCHAFT MBH, *Preliminary Study on the Utilization of Iron Ore Deposits in Surinam* (1966) and *Preliminary Study on Wood Production and Utilization in Surinam* (1965), two studies commissioned by the Surinam government, contain basic information about the country and some of its natural resources. *Investment Factors and General Information About Surinam* and *Facts and Figures About Surinam*, both published by the Economic Division of the Ministry of Trade and Industry (1968), contain valuable information for investors. In addition, *Doing Business in Surinam*, published by the Foundation for the Promotion of Investment in Surinam (1969), provides explicit information about the economy. A.T. VINK (comp.), *Surinam Timbers* (1965), describes the main timber species and deals with the properties and uses of wood as well as with supply. WILLEM VAN DE POLL, *Suriname: een fotoreportage van land en volk* (1949; Eng. trans., *Surinam: The Country and Its People*, 1951) and *Suriname: Foto's en tekst* (1959), give good accounts of the people, vegetation, and animal life. F.E. ESSED, *Opbouw '70* (1971), provides an authoritative analysis of social and economic conditions; a retrospective view is presented by J.H. ADHIN in *Development Planning in Surinam in Historical Perspective* (1961). J. WARREN NYSTROM, *Surinam: A Geographic Study* (1942), is a well-written, informative book. Also of interest are MELVILLE J. and FRANCES S. HERSKOVITS, *Suriname Folk-lore* (1936, reprinted 1969); and MARTHA M. SUMWALT, *Surinam in Pictures* (1971).

(J.G.W.De V.)

Surveying

Surveying is a means of relatively large-scale, accurate measurement of earth surfaces. It includes the determination of the measurement data, the reduction and interpretation of the data to usable form, and, conversely, the establishment of relative position and size according to given measurement requirements. Thus surveying has two similar but opposite functions: (1) the determination of existing relative horizontal and vertical position, such as that used for the process of mapping, and (2) the establishment of marks to control construction or to indicate land boundaries.

Surveying has been an essential element in the development of man's environment for so many centuries that its importance is often forgotten. It is an imperative requirement in the planning and execution of nearly every form of construction. Surveying was essential at the dawn of history, and some of the most significant scientific discoveries could never have been implemented were it not for the contribution of surveying. Its principal modern uses are in the fields of transportation, building, apportionment of land, and communications.

Except for minor details of technique and the use of one or two minor hand-held instruments, surveying is much the same throughout the world. The methods are a reflection of the instruments, manufactured chiefly in Switzerland, Austria, Great Britain, the U.S., Japan, East Germany, and West Germany. Instruments made in Japan are similar to those made in the West.

HISTORY

Egyptian and other ancient devices

It is quite probable that surveying had its origin in ancient Egypt. The Great Pyramid of Khufu at Giza was built *c.* 2700 BC exactly 755 feet long and 480 feet high. It is so accurately square and so perfectly oriented to the cardinal points of the compass that it stands as evidence of the command ancient Egypt had of surveying.

Evidence of some form of boundary surveying as early as 1400 BC has been found in the fertile valleys and plains of the Tigris, Euphrates, and Nile rivers. Clay tablets of the Sumerians show records of land measurement and plans of cities and nearby agricultural areas. Boundary stones marking land plots have been preserved. There is a representation of land measurement on the wall of a tomb at Thebes (1400 BC) showing head and rear chainmen measuring a grainfield with what appears to be a rope with knots or marks at uniform intervals. Other persons are shown, two of high estate according to their clothing—probably a land overseer and inspector of boundary stones.

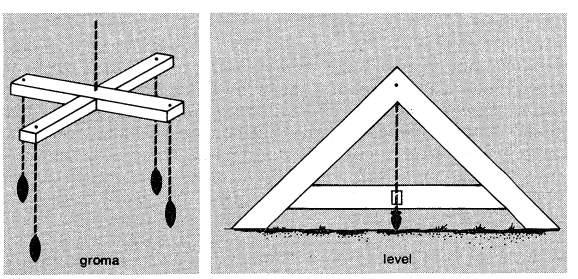


Figure 1: Egyptian surveying instruments.

There is some evidence that, in addition to a marked cord, wooden rods were used by the Egyptians for distance measurement. They certainly had the groma, which was used to establish right angles. It was made of a horizontal wooden cross pivoted at the middle and supported from above (Figure 1, left). From the end of each of the four arms hung a plumb bob. By sighting along each pair of plumb bob cords in turn, the right angle could be established. The device could be adjusted to a precise right angle by observing the same angle after turning the device approximately 90°. By shifting one of the cords to take up half the error, a perfect right angle would result.

There is no record of any angle-measuring instruments of that time, but there was a level (Figure 1, right), consisting of a vertical wooden A-frame with a plumb bob supported at the peak of the A so that its cord hung past an indicator, or index, on the horizontal bar. The index could be properly placed by standing the device on two supports at approximately the same elevation, marking the position of the cord, reversing the A, and making a similar mark. Halfway between the two marks would be the correct place for the index. Thus, with their simple devices, the ancient Egyptians were able to measure land areas, replace property corners lost when the Nile covered the markers with silt during floods, and build the huge pyramids to exact dimensions.

Greek log line and Chinese lodestone

The Greeks used a form of log line for recording the distances run from point to point along the coast while making their slow voyages from the Indus to the Persian Gulf about 325 BC. Still earlier it is said that the Chinese knew the value of the lodestone and possessed some form of magnetic compass, which reached the West from the Arabs in the 12th century. The astrolabe was introduced by the Greeks in the 2nd century BC. An instrument for measuring the altitudes of stars, or their angle of elevation above the horizon, it was in the form of a graduated arc suspended from a hand-held cord. A pivoted pointer that moved over the graduations was pointed at the star. The instrument was not used for nautical surveying for several centuries, remaining a scientific aid only.

During their occupation of Egypt, the Romans acquired Egyptian surveying instruments, which they improved slightly and to which they added the water level and the plane table. About 15 BC the noted Roman architect and engineer Vitruvius mounted a large wheel of known circumference in a small frame, in much the same fashion as a wheel of a wheelbarrow; when it was pushed along the ground by hand it automatically dropped a pebble into a container at each revolution, giving a measure of the distance travelled. It was, in effect, an odometer.

The water level consisted of either a trough or a tube turned upward at the ends and filled with water. At each end there was a sight made of crossed horizontal and vertical slits. When these were lined up just above the water level, the sights determined a level line accurate enough to establish the grades of the Roman aqueducts. In laying out their great road system, the Romans are said to have used the plane table. It consists of a drawing board mounted on a tripod or other stable support and of a straightedge—usually with sights for accurate aim (the alidade) to the objects to be mapped—along which lines are drawn. It was the first device capable of recording or establishing angles. Later, adaptations of the plane table had magnetic compasses attached.

Plane tables were in use in Europe in the 16th century, and the principle of graphic triangulation and intersection was practiced by surveyors. In 1615 Willebrord Snell, a Dutch mathematician, measured an arc of meridian by instrumental triangulation.

In 1620 English mathematician Edmund Gunter developed a surveying chain, which was only superseded by the steel tape in the beginning of the 20th century.

The study of astronomy resulted in the development of angle-reading devices that were based on arcs of large radii, making such instruments too large for field use. With the publication of logarithmic tables in 1620, portable angle-measuring instruments came into use. They were called topographical instruments, or theodolites. They included pivoted arms for sighting and could be used for measuring both horizontal and vertical angles. Magnetic compasses may have been included on some.

17th-century portable instruments

The vernier, an auxiliary scale permitting more accurate readings (1631), the micrometer microscope (1638), telescopic sights (1669), and spirit levels (about 1700) were all incorporated in theodolites by about 1720. Stadia hairs (see Figure 2) were first applied by James Watt in 1771. The development of the circle-dividing engine about 1775, a device for dividing a circle into degrees with great accuracy, brought one of the greatest advances in surveying methods, as it enabled angle measurements to be made with portable instruments far more accurately than had previously been possible.

The start of modern surveying

By the late 18th century modern surveying can be said to have begun. One of the most notable early feats of surveyors was the measurement in the 1790s of the meridian from Barcelona, Spain, to Dunkirk, France, by two French engineers, Jean Delambre and Pierre Méchain, to establish the basic unit for the metric system of measurement.

Many improvements and refinements have been incorporated in all the basic surveying instruments. These have resulted in increased accuracy and speed of operations and have opened up possibilities for improved methods in the field. In addition to modification of existing instruments, two revolutionary mapping and surveying changes have been introduced: photogrammetric, or mapping from aerial photographs (about 1920); and electronic distance measurement, including the adoption of the laser for this purpose as well as for alignment (in the 1960s). Important technological developments benefitting surveying in the 1970s included the use of satellites as reference points for geodetic surveys and computers to speed the processing and recording of survey data.

FUNDAMENTALS OF SURVEYING

The objective of all surveying—whether plane or geodetic—is to determine accurately land or water areas and their boundaries. Related data sought, also determined through surveys, include depths, altitudes, rate of slope (especially useful in mines and tunnel work), and Earth volumes such as required for construction. Thus the surveyor specializes in the measurement of distances and directions, both horizontal and vertical.

Measurements. In contrast to other types of measurement, nearly all those in surveying are made either perpendicular to the direction of gravity (designated as horizontal) or in the direction of gravity (designated as vertical). When not so made, the quantities are immediately reduced mathematically to their horizontal or vertical equivalents. Thus the direction of gravity must be determined whenever a surveying measurement is taken. Surveying measurements are of four types: (1) horizontal lengths or distances, (2) vertical lengths, or differences in height or elevation, (3) horizontal angles measured in horizontal planes, and (4) vertical angles measured in vertical planes.

Horizontal lengths, with few exceptions, are measured in short, straight sections that are together very nearly arcs along the Earth's surface, which again is actually a curved surface because of the spherical nature of the Earth. The surface is calm water, which—because it is continuous everywhere perpendicular to the (vertical) direction of gravity—is a level surface. With a surveyor's level, the difference in height between two points, or the vertical distance between the level surfaces that pass through them (see below *Orthometric Corrections*), can be determined. The level is placed equidistant between the two points, and the graduations are read on a rod held vertically on each point in turn; the difference between the readings represents the difference in height of the two points.

Measuring differences in height

For less accurate results, the instrument need not be placed at the exact midpoint. Variations in the atmosphere affecting the refraction of light restrict the distances permissible from the instrument to the points observed; differences in height between widely separated points, therefore, must be determined by a series of such measurements over temporary points, called turning points. The operation is called levelling, or running a line of levels.

Within any area of the Earth's surface about 12 miles (20 kilometres) square and free of large variations in altitude, the differences between measurements made as described above and the same measurements that might be reproduced on, or from, a truly plane surface would be negligible because the differences caused by the Earth's curvature in a small area are minimal. The results of such surveys can be computed by plane trigonometry; the whole process is called plane surveying.

For surveys covering larger areas, however, the Earth's curvature must be considered. Such surveying is called geodetic; it must be very accurate, and goedetic instruments unusually precise.

Units of measure. The metre is the fundamental unit of length used in surveying throughout the world. In October 1960, the length of the metre was defined as equal to 1,650,763.73 times the wavelength of a specific line in the spectrum of krypton-86 under specified conditions. This stabilized but did not effectively change the length of the metre.

Except in rare instances, the only unit of length used in surveying, in addition to metric units, is the foot, with tenths, hundredths, and thousandths of a foot being used instead of inches and fractions. From 1893 until 1959, the foot used in the United States equalled $^{1200}/_{3937}$ metre or 0.30480061 metre. It was not quite the same length as the foot used in other English-speaking nations, as these were based on the imperial yard, whose relationship with the metre varied slightly from time to time. In 1959 it was agreed by representatives of the English-speaking nations to base the foot on the metre according to the exact ratio: 1 foot = 0.3048 metre. This shortened the foot used in the United States by about two parts in 1,000,000 and lengthened the contemporary British foot by about ⅓ part in 1,000,000.

Setting the exact length of a foot

The U.S. Coast and Geodetic Survey, which maintains the fundamental survey control systems in the United States, measures in metres. Since 1893, this survey has converted to values in feet by the ratio 1.200 : 3.937 in all publications in which the foot appears. To avoid confusion it was necessary to retain this value of the foot until such time as all vertical and horizontal control systems should be readjusted. Accordingly, this foot was named the U.S. Survey Foot.

Two types of angular measure are used throughout the world, the sexagesimal and centesimal:

Sexagesimal system	Centesimal system
1 circle = 360 degrees (°)	1 circle = 400 grads (g)
1 degree = 60 minutes (′)	1 grad = 100 minutes (′)
1 minute = 60 seconds (″)	1 minute = 100 seconds (″)

Directions are given in azimuths or bearings. Azimuths are clockwise angles from the north or occasionally from the south. Equivalent values are the following:

Azimuth from North	50°	120°	200°	310°
Azimuth from South	230°	300°	20°	130°
Bearing	N 50°E	S 60°E	S 20°W	N 50°W

INSTRUMENTS

The fundamental measurements made by the surveyor include distances, angles (both horizontal and vertical), altitudes, and directions. To carry out these measurements, several different types of instruments and supplementary equipment are used.

Distance-measuring equipment. Distances are measured by mechanical devices (chain or steel tape) or by electronic means.

The chain. Gunter's chain, the 300-year-old measuring instrument by which all survey measurement in the English-speaking countries and much of it elsewhere was long done, has been superseded by the steel tape and electronic equipment. Gunter's chain is 66 feet long; 80 chains equal one mile, and 10 square chains equal an acre. The chain is subdivided into 100 links. A rod or perch was 25 links. Each link was a short section of wire connected to the next link by a loop. At each end of the chain was a brass handle. The 66-foot unit is still called a chain and is still in use in property descriptions and in the public land systems.

The steel tape. A graduated steel ribbon, or a flat wire, with handles at each end, is the basic modern means of measuring distance. Its length is taken as the straight-line distance between the two end marks at 68° F (20° C) because the tape's actual length varies with different types of support, different tensions, and different temperatures. In the field, if extreme accuracy is desired, the type of support, the tension, the temperature, and the difference in height of the two end marks are recorded and the measured length corrected accordingly.

In most routine surveys the tape is held by hand at the

ends, kept high enough to clear ground objects, made horizontal by estimation, and placed in the direction of measurement by eye or with an instrument. The desired tension is estimated, and the positions of the graduations are brought to the ground marks by plumb bobs. Sometimes the air temperatures are recorded. For higher accuracy, tripods or other supports are employed, or the tape is fully supported on smooth surfaces. The slopes and temperatures of the tape are determined. Tension is regulated with a spring balance.

Eliminating temperature problems

When especially accurate measurements are necessary, particularly for base lines in a triangulation system, a steel tape must be used at night or when the sky is overcast; otherwise radiant heat would make it impossible to determine tape temperature. Tapes made of Invar avoid this difficulty. Invar is a steel alloy with an extremely low coefficient of thermal expansion (about 0.00000045 foot per foot per degree Fahrenheit). With such a tape, a temperature difference of 10° F (5.6° C) from the design temperature would result in an error of only about 1/20 of an inch (1.3 millimetres) in 1,000 feet (305 metres). Unfortunately, Invar tapes are too easily damaged for general use. Before the introduction of Invar, iron bars immersed in melting ice were used for daytime measurements.

Electronic instruments. Sir Robert Watson-Watt, a Scottish physicist working in England, introduced an electronic pulse-type instrument in 1935. Called hiran, it measures distances over water between land survey stations obscured from each other by the Earth's curvature. An airplane flying high enough to be visible from both stations crosses over the line between them while sending out a series of pulses. These are automatically relayed back to the plane from each station. The minimum sum of distances represented by the timed returns is chosen, and the nearby sums are corrected to be averaged with it. The result is corrected for the altitude of the plane and refraction of the atmosphere, and the final value is taken as the distance between the two stations.

The phase-shift type of system was first developed in 1948 and gives very accurate results. The process is comparable to conventional radio transmission, in which the carrier frequency is modulated by the audio frequencies that the receiver makes audible.

In phase-shift devices, the carrier frequency is either light (generated by a laser or an electric light beam) or an ultrahigh-frequency radio beam. The light beam requires a clear line of sight; the radio beam can penetrate fog, haze, heavy rain, dust, sandstorms, and some foliage. Both types have a transmitter-receiver at one survey station. At the remote station, the light type uses a set of corner mirrors; the high-frequency type utilizes a retransmitter (requiring an operator) identical to the transmitter-receiver at the original station. A corner mirror has the shape of the inside of a corner of a cube. It returns light toward the source from whatever angle it is received, within reasonable limits. A retransmitter must be aimed at the transmitter-receiver.

In both types of instrument, the distance is determined basically by the length of time it takes the radio or light beam to travel to the target and back (see also RADAR; SONAR). The elapsed time is determined by the shift in phase of the modulating signal during its travels. Electronic circuitry detects this phase shift and converts it very accurately into the exact fraction of a small unit distance. By using several frequencies for the modulator signal the total distance can be computed.

Shift in phase of modulating signal

Angle-measuring equipment. The theodolite is the basic instrument for measuring horizontal and vertical angles. It consists essentially of a sighting tube in the form of a telescope, and horizontal and vertical scales so set up that horizontal and vertical angles of the object sighted in the telescope can be read from the scales.

Theodolites and transits. There are two general types of theodolites: direction theodolites (also called simply theodolites) and repeating theodolites (also called American transits). In the direction theodolite, the graduated horizontal circle remains fixed while the telescope is pointed at several objects in turn and the direction to each is read on the horizontal scale. In the repeating theodolite, the same angle is measured a number of times on the horizontal scale, which is rotated by as much as the arc of each measurement. The successive values of each measurement are accumulated on the horizontal circle, the final reading representing the total arc of all the measurements taken. The desired value is then obtained by dividing the final reading by the number of repetitions. The repeating theodolite, or transit, is used in the U.S., Canada, and, to some extent, in Central and South America.

The theodolite is basically a more accurate instrument than the transit, measuring angles to five seconds of arc or less; the transit only measures to about 20 seconds of arc. The transit's repeating characteristic, however, permits the angle to be measured several times and an average reading obtained, in about the same time that it takes to measure once with a theodolite, and the end result is equal accuracy. When angles are established by one setting from one fixed station, as in placing marks for construction, the theodolite is more accurate.

Modern transits and theodolites nearly all have stadia hairs (see Figure 2). When the vertical angle is observed, along with the distance on the rod (staff) intercepted by

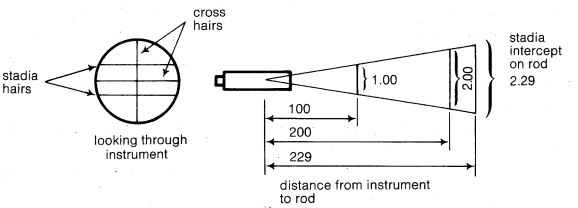


Figure 2: Principles involved in the use of telescope stadia hairs and a staff to determine distance (see text). In the example illustrated, the distance to the rod is equal to 100 times the stadia intercept on the rod.

two of the horizontal stadia hairs, the surveyor can determine approximately the horizontal distance and the difference in height between a point in the telescope and the point on the rod or staff (see below) where the cross hair falls.

Vertical circle attachments. Except in astronomical observations, vertical angles seldom need be measured with the accuracy required for horizontal angles. In all instruments, the vertical circle is attached to the end of the horizontal axle and turns with the telescope. In American transits, it is read with a vernier attached to a standard and thus depends on the accuracy of the levelling of the instrument. In modern European instruments, the vertical circle is read with an optical device governed by a pendulum and therefore is independent of any inaccuracies in the hand levelling of the instrument. The accuracy of the setting of the reading device in both instruments can be tested by reading the same vertical angle with the telescope both direct and reversed.

Testing accuracy of the setting

Plane-table alidades. Modern plane-table alidades consist of a telescopic sight of about 13–16 power equipped with stadia hairs and mounted on a metal graduated straightedge called a blade, with which survey data can be plotted on the supporting drawing board (see below *Maps and charts*). The telescope, on horizontal bearings, does not reverse, and it measures angles of elevation and depression up to as much as 20° to 30°, as indicated on a vertical circle. A prismatic eyepiece makes it possible to observe when looking down into the telescope to avoid the interference of the plane table. Also mounted on the blade are a circular spirit level for levelling the plane table and a trough compass that operates when the alidade is aimed nearly along the magnetic meridian.

Plane-table surveying has been nearly eclipsed by photogrammetry. It is still used, however, for mapping small areas, for filling in where the lack of detail prevents the proper orientation of aerial photography, or where detail is concealed on the photograph. It is sometimes used for finding the position and elevation of photo-control points.

Measurement of direction. There are two major techniques for measuring direction—by means of compasses and by celestial observations.

Compasses. Two needle-types are used: box compasses and circular compasses. Box compasses must be used with angle-measuring instruments, as the compasses operate only when the line of sight points toward magnetic north. The bearing desired must then be measured from this direction.

Circular compasses can be mounted on theodolites and are often permanently mounted on transits. When provided with open sights, they can be used alone, either mounted on light tripods or held in the hand.

Magnetic declination, called variation by mariners, is the angle from the direction of true north to the direction of the needle. This angle varies from place to place, and at any place it changes slowly through large, little-understood swings. This is called the secular change. The diurnal change is a more or less regular swing as follows: quiescent at its mean position from 9 PM to 3 AM, then easterly to 2 to 5 minutes at 8 or 9 AM, and then westerly to 2 to 5 minutes at 1 or 2 PM, thence to mean at 9 PM. There is also an annual swing in the order of one minute, varying according to locality.

Secular and diurnal changes

Celestial observations. Astronomic north, the direction of the celestial North Pole, is determined by observations of the sun or stars with transits or theodolites. The most accurate method of determining direction is to observe the North Star (Polaris) in latitudes where it is visible and not too high to give accurate results. In southern latitudes there are six bright stars about 30° from the Pole, well enough distributed so that one is usually above the horizon. These are observed in the same way as Polaris.

As the first step, the horizontal direction of the star is measured from a mark on the ground. The precise time is recorded at the moment the instrument is pointed at the star. With these values and the latitude and longitude of the instrument, the bearing of Polaris or other star with relation to an Earth station can be computed. If any of these data is unavailable further observations are required.

A second method is sometimes used. The altitude, direction, and the time of observation are noted for a star in the east and one in the west at about the same elevation. By referring to astronomical tables the resulting bearings of the mark are obtained. These are averaged and thus observational errors in the vertical angles are reduced.

Vertical measuring equipment. Theodolites and transits measure vertical (as well as horizontal) angles. Surveying levels and barometric-pressure devices measure vertical levels, or elevations, exclusively.

Surveying levels. There are three types of surveying levels: (1) the wye, or the dumpy, level; (2) the tilting level; and (3) the self-levelling level. The first type consists of a telescopic sight, like that of a transit but usually of slightly higher magnification, to which a long spirit level is attached and adjusted so that the bubble centres when the line of sight is horizontal. The whole is mounted on a vertical spindle in a levelling head so that it can be turned in any direction. The telescope of the wye level can be rotated and changed end for end in the wyes (Y-shaped supports) for adjustment.

The tilting level is similar, but the telescope with the main bubble attached can be separately tilted up or down by means of a micrometer screw, giving it greater accuracy. The spindle is first set nearly vertical with a circular spirit level, and the main bubble is centred with the micrometer screw.

The self-levelling level is similar to the tilting level except that it has no micrometer screw. Instead, it contains an optical part that is operated by a pendulum in such a way that the line of sight is automatically made horizontal. Some types are extremely accurate, and the pendulum eliminates the operation of centring the main bubble, which is especially time-consuming.

For very accurate work, tilting levels are equipped with coincidence-reading bubbles. By an optical device the elongated bubble appears to be split into two sections, with a central portion apparently removed. As the telescope is levelled, the sections of the bubble are brought closer together until their ends coincide; then the bubble is precisely centred. Often tilting and self-levelling levels have three horizontal cross hairs so that three readings of the staff (rod) can be taken and averaged. Optical micrometers also yield greater accuracy in rod readings; they consist of thick glass disks with parallel, optically flat surfaces usually mounted in front of the telescope. When the micrometers are rotated on a horizontal axis, they move the line of sight up or down and parallel to itself. A graduated drum indicates the distance moved from their normal position, so that when the line of sight is placed exactly on a rod graduation, they read the very small additional quantity to be applied.

Coincidence-reading bubbles

The level rod (staff). The rod is used to measure the vertical distance between the point on which it stands and the line of sight. It is made of wood with graduations, and large, clear numbers usually printed on it. It is divided either into centimetres or half-centimetres or into hundredths of a foot or yard. Between divisions it is alternately black and white, so that the exact values occur where the colour changes. For routine levelling, the American rod is made in two sections. The rear section slides upward until, when fully extended, the face reads continuously up to 12 or 13 feet. It has a target that may be set where desired.

For accurate levelling the graduations are painted on a steel or Invar ribbon fastened at the bottom and held taut by a spring at the top. A special thermometer is held with its bulb against the ribbon, and a circular spirit level is attached to aid in keeping the rod vertical.

Aneroid barometers. Vertical measurements can be approximately determined by finding the difference in barometric pressure at the two elevations. Aneroid barometers and hypsometers measure such differences. Aneroid barometers are devices in which changes in atmospheric pressure cause a needle to move over a scale. Instruments of this type designed for surveying are called altimeters. A type that records time along with pressures is usually placed at a known elevation (the base) and the second instrument is taken to the points where measurements of elevation are desired. When the second instrument is read, the time is recorded, so that the simultaneous reading of the base instrument can be selected. The difference of the two readings must be corrected for the unit weight of the air, which is estimated from the barometric pressure, temperature, and humidity.

More accurate results, independent of the unit weight of the air, can be obtained by the two-base method. Recording aneroids are placed at two bases, preferably one higher and one lower than the elevations to be determined. In effect they weigh the air. Each field reading is adjusted in proportion to the relative height above and below the two bases and so that the sum of the two heights equals the known difference in height between the bases. Within a radius of ten miles (16 kilometres), this method gives elevations within about two feet (0.6 metre).

Two-base method

In Australia, some effort has been made to use three bases over a large area and to compensate according to the barometric pressures shown on a weather map.

Hypsometers. The hypsometer is a device that measures barometric pressure by determining the temperature required to boil distilled water. By means of suitable tables, this temperature can be converted to barometric pressure. The bulb of an accurate thermometer is immersed in a very small container of water that is brought to a boil. The results can be handled as described for the aneroid barometer, but usually the boiling point is again determined at the starting point, and if there is a difference in the measured altitude there, the field elevations are corrected in proportion to elapsed time.

CURRENT SURVEYING PRACTICES

Survey control. The accuracy of survey measurements can be improved almost indefinitely but only at increased cost. Accordingly, control surveys are used; these consist of a comparatively few accurate measurements that cover the area of the project and from which short, less accurate measurements are made to the objects to be located. The simplest form of horizontal control is the traverse, which consists of a series of marked stations connected by

measured courses and the measured angles between them. When such a series of distances and angles returns to its point of beginning or begins and ends at stations of superior (more accurate) control, it can be checked and the small errors of measurement adjusted for mathematical consistency. By assuming or measuring a direction of one of the courses and rectangular coordinates of one of the stations, the rectangular coordinates of all the stations can be computed.

Superior horizontal control. A system of triangles usually affords superior horizontal control. All of the angles and at least one side (the base), of the triangulation system are measured. Though several arrangements can be used, one of the best is the quadrangle or a chain of quadrangles. Each quadrangle, with its four sides and two diagonals, provides eight angles that are measured. To be geometrically consistent, the angles must satisfy three so-called angle equations and one side equation. That is to say, the three angles of each triangle, which add to 180°, must be of such sizes that computation through any set of adjacent triangles within the quadrangles will give the same values for any side. Ideally, the quadrangles should be parallelograms. If the system is connected with previously determined stations, the new system must fit the established measurements.

When the survey is of an area where the Earth's curvature is a factor, an imaginary mathematical representation of the Earth must be employed as a reference surface. A level surface at mean sea level is considered to represent the Earth's size and shape, and this is called the geoid. Because of gravity anomalies, the geoid is irregular; however, it is very nearly the surface generated by an ellipse rotating on its minor axis—*i.e.*, an ellipsoid slightly flattened at the ends, or oblate. Such a figure is called a spheroid. Several have been computed by various authorities; the one usually used as a reference surface by English-speaking nations is the (Alexander Ross) Clarke spheroid of 1866. This oblate spheroid has a polar diameter about 27 miles less than its diameter at the equator.

The imaginary Earth, or geoid

Because the directions of gravity converge toward the geoid, a length of the Earth's surface measured above the geoid must be reduced to its sea-level equivalent—*i.e.*, to that of the geoid (see Figure 3). These lengths are assumed to be the distances, measured on the spheroid, between the extended lines of gravity down to the spheroid from the ends of the measured lengths on the actual surface of the Earth. The positions of the survey stations on the Earth's surface are given in spherical coordinates (*i.e.*, geodetic latitude and longitude).

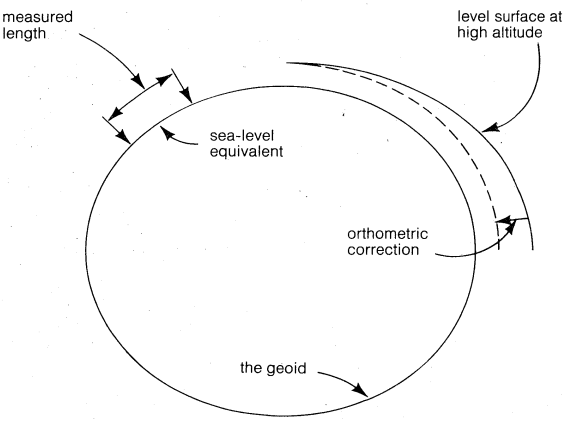


Figure 3: Measured length corrected to its sea-level equivalent (on the geoid) and the orthometric correction for high altitude measurements (see text).

Vertical control. Bench marks, or marked points on the Earth's surface, connected by precise levelling constitute the vertical controls of surveying. The elevations of bench marks are given in terms of their heights above a selected level surface called a datum. In large-level surveys the usual datum is the geoid. The elevation taken as zero for the reference datum is the height of mean sea level determined by a series of observations at various points along the seashore taken continuously for a period of 19 years or more. Because mean sea level is not quite the same as the geoid, probably because of ocean currents, in adjusting the level grid for the U.S. and Canada all heights determined for mean sea level have been held at zero elevation.

Orthometric corrections. Because the level surfaces, determined by levelling, are distorted slightly in the area toward the Earth's poles (because of the reduction in centrifugal force and the increase in the force of gravity at higher latitudes), the distances between the surfaces and the geoid do not exactly represent the surfaces' heights from the geoid. To correct these distortions, orthometric corrections (see Figure 3) must be applied to long lines of levels at high altitudes that have a north–south trend.

Trigonometric levelling often is necessary where accurate elevations are not available or when the elevations of inaccessible points must be determined. From two points of known position and elevation, the horizontal position of the unknown point is found by triangulation, and the vertical angles from the known points are measured. The differences in elevation from each of the known points to the unknown point can be computed trigonometrically.

Maps and charts. The most familiar, everyday experience with surveying for most persons is the use of maps. A complete map or chart shows the horizontal size, position, and elevation of natural and man-made topographic features according to a scale or, if the Earth's curvature is considered, according to a mathematical projection. The shape of the ground in elevation is shown by contour lines that connect points of equal elevation; water depths are shown by spot depths and sometimes by lines of equal depth. Small-scale maps and charts show lines of latitude and longitude; large-scale maps show an indication of geodetic, astronomic, or magnetic north and often the lines of a rectangular coordinate system. Maps or charts made for special purposes omit many of these features.

A drawing of a line that shows to scale the elevation of the ground or other features along a selected route is called a profile. Frequently the profiles of several features are shown on the same drawing.

Taking a profile of surface features

Large areas of the Earth have been surveyed, mapped, and charted, with the results published by government agencies. Government triangulation systems cover most of the world and are connected over oceans by observations of moon occultations, ordinary astronomical observations, and satellite triangulation; the systems are coordinated temporally by radio signals. Smaller bodies of water, but large enough for the Earth's curvature to prevent ground triangulation, are crossed by trilateration—the measurement of the sides of a system of triangles—with electronic devices or triangulation by simultaneous observations of flares dropped from an airplane. Level nets (interconnecting networks of level circuits) based on mean sea level give elevation control as points of reference. Secondary triangulation and levels are established where the mapping will be carried out. Photogrammetry (see below) is used almost exclusively for supplying detail.

The plane table, with its drawing board and alidade mounted on a tripod, is a basic tool of map making. When plotting a distant point, the map maker keeps the rear end of the table's straightedge at the point where he is stationed and aims the front end at the rod held at the distant point to be located. Once it is aimed, he reads the stadia intercept and the vertical angle and records the direction by drawing a line along the straightedge onto paper on the plane table. Points are so located, and a map can be drawn as the work progresses.

Surveys for smaller maps usually are based on traverse and sometimes triangulation with a rudimentary level network to control elevations. Detail is filled in by transit stadia on the plane table. Where elevations are important, the area is laid out in squares and the elevations of the intersections are determined with a level.

When large areas must be mapped but accurate control is lacking, positions are determined astronomically, and elevations are measured by barometric pressure, usually

determined by sensitive aneroid barometers or by hypsometers. Elevations are measured from aircraft by determining the height of the aircraft above the ground with electronic devices and the elevation of the aircraft above sea level by altimeter or trigonometric levelling.

Photogrammetry. Photogrammetry is the measurement of photographs for certain engineering and other purposes. Applied to surveying, it is used to supply detail

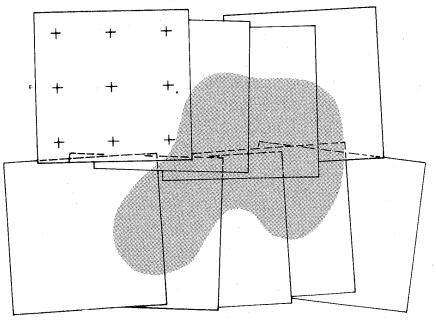


Figure 4: Photogrammetric photographs from two short, overlapping flight strips arranged for supplying mapping details. Photo-control points are shown on only one photograph; shading indicates a typical terrain feature such as a lake (see text).

for mapping. Sometimes it is used for planimetric maps, but usually for topographic maps that show contour lines. Whereas it is possible to map from photographs taken from high ground elevations, the greatest use of photogrammetry is the construction of maps from aerial photographs taken from points almost directly overhead.

Flight strips with 10 percent overlap

Trained pilots fly the airplane or helicopter along previously planned flight strips that are straight, parallel, and so spaced that they overlap by about 10 percent. In each strip the camera automatically changes the film and makes an exposure that causes the photographs to overlap about 60 percent. Thus every ground point appears on at least two and sometimes six pictures (see Figure 4). Ideally, each strip should contain many photo-control points—*i.e.*, three per photograph. The control points that are chosen on the photographs can be precisely determined and are uniformly spaced. The photographs are converted into maps by plotters, of which there are many types of varying accuracy. All of the better ones work on the same principle, which is best illustrated by the multiplex projector (see Figure 5).

That instrument consists of a row of projectors, each with a two-inch by two-inch diapositive (transparency), made from an aerial photograph. The projectors aim downward to the future manuscript map on which the horizontal positions of the photo-control points have been plotted to the desired map scale. Supported on the map surface is a movable tracing table consisting of a white surface about six inches in diameter with a tiny white light at the centre. The table can be raised or lowered as required. Directly under the light and almost on the map surface is a pointer that can be interchanged for a pencil that rests on the map.

The tracing table is placed on each photo-control point in turn and adjusted in height, to the map scale, according to the elevation of the photo-control point. The projector containing the diapositive is adjusted in height, azimuth, position, and aim until the image of each photo-control point falls on the white light, adjusted to the control point position. This reproduces, to scale, the exact geometry that existed when the photograph was taken. A second diapositive is adjusted in the same way.

To create contour lines, the tracing table is set at the elevation of the desired contour. A blue-green filter is placed on one projector and a red filter on the other, and the operator wears coloured glasses arranged in the same way. With the tracing table at the desired elevation of a contour line and placed at B, point C is projected as shown, as the original geometry has been recreated. The operator sees the projection of C at one point with his left eye and at another point with his right eye, as though C were at A. Because B is a white light, he sees it with both eyes in its actual position. Thus the apparent depth of C below B is greatly exaggerated. He moves the table to D where the white light appears to strike the ground. He lets down the pencil and by following the ground surface traces the desired contour. When plotting an actual object he raises or lowers the plotting table so the light appears to follow the ground surface.

Procedure for conducting a survey. The second principal function of surveying after mapping, as has been mentioned, is the establishment of marks to control construction or to indicate land boundaries. A complete survey consists of four steps: (1) reconnaissance; (2) a preliminary survey; (3) a map; then a plan is prepared based on steps 2 and 3; and (4) a location survey based on the plan. Each of these steps may be rudimentary or intricate, depending on the data available, the field conditions, and the project for which the survey is carried out.

Basic steps in surveying

Reconnaissance. A reconnaissance is a study of an area to determine where the preliminary survey or surveys must be made, the extent of the survey, and the information it must obtain. The reconnaissance might be merely a look at the ground, the study of existing maps, a crude survey, or a study of a pieced-together set of controlled aerial photographs. When triangulation is to be used for control, the location of the stations must be determined and the necessary heights of the towers noted.

Preliminary survey. The preliminary survey measures the relative positions in three dimensions of existing objects that control the plan. It usually consists of a horizontal survey control system of triangulation and traverse with connecting survey stations and a vertical control system of precise levelling with connecting bench marks. From these control points, measurements are made to existing objects by tape, plane table, stadia, or photogrammetry, depending on the accuracy required.

Mapping the details. If an office building is to be planned, for example, the architects will need a map that shows such data as (1) property lines, easements, and encroachments with their bearings and lengths; (2) the exact positions and floor elevations of any existing buildings that may be connected with the new building; (3) trees and other objects in the area; (4) contours with a contour interval of one or two feet; (5) any streams or ponds; and (6) the elevations of the low points on the property lines in order to plan the surface drainage system. In addition, the architects will need to know the positions of the adjacent public roads or streets, the eleva-

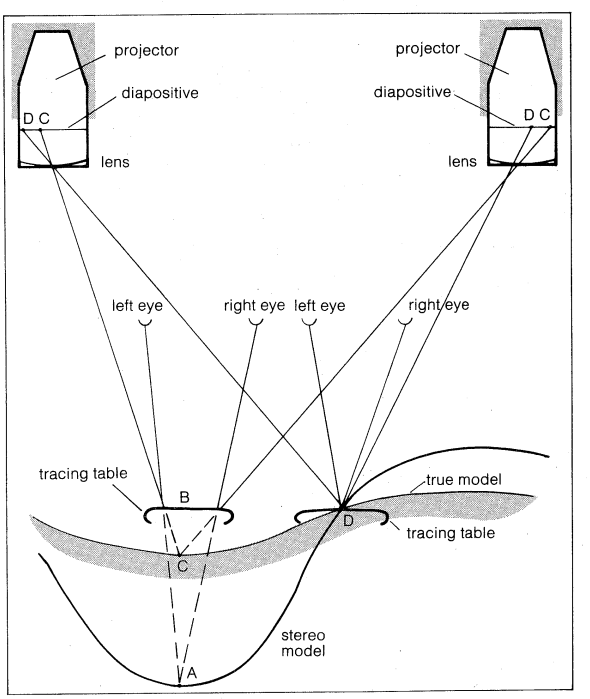


Figure 5: Multiplex projector system of mapping contours and other details from aerial photographs. Only two projectors are shown (see text).

tions of their centre lines and their gutters, the tops of curbs, all at 25- or 50-foot intervals, and facilities, like poles, that might interfere within the street rights-of-way.

Underground facilities must be shown. These include the elevation, position, size, and water pressure of water mains, gas mains, and electric conduits, and the exact elevations of storm and sanitary sewers and their manholes. Frequently, test borings must be made and located.

A land surveyor must be employed to mark the property corners, the easements, and encroachments. His survey usually serves as the control survey. If the project is in a town, levels must be run in from the nearest municipal bench mark to permanent bench marks, preferably three, established in or near the property and safe from disturbance by construction. If the area is fairly large, supplementary control by traverse and levelling is often necessary to reduce the length of measurements to details. Ties to control points are often made with tapes, stadia, hand levels, and by ordinary levelling.

The map is plotted according to the survey measurements. The architect gives dimensions from the property lines and states elevations from which the basic building lines and elevations can be laid out.

Property boundary surveys. Surveys to establish property boundaries constitute a large, highly specialized branch of surveying known as land surveying. Not only must a land surveyor be thoroughly versed in surveying techniques but he must understand real property law, be familiar with property lines within the area where he works, and be able to make decisions that he can justify in court when appearing as an expert witness.

Land surveying

Monuments. In general, property boundary lines must be shown by marks on the ground (landmarks), often called monuments. These marks usually consist of ridges, streams, blazed lines, ditches, walls, fences, trees, fieldstones, stakes, iron pipes or rods, or prepared square concrete or stone posts sunk into the ground.

Land surveying as it is now known was first used to determine acreage. The directions of the boundary lines were determined by compass, and their lengths were determined usually by a Gunter's chain. With the introduction of the transit and tape, it became possible to locate land boundaries accurately by measurement from any monumented line, resulting in an increase in recorded measurements and fewer monuments.

Because the cost of a survey increases with its accuracy, surveys of low-priced land are seldom very accurate. Surveys of parcels of such lands often contain errors, omissions, and inconsistencies. In localities where land is increasing in value, these defects must be eliminated. Whereas the power of decision rests with the court, the land surveyor must reach a solution based on his experience, his estimate of the accuracy of the work of the previous surveyors, his familiarity with the property lines in the vicinity, and his knowledge of the law.

Metes and bounds. A land description in which the boundaries are described by the directions and length of successive boundary lines that the land adjoins is called a metes-and-bounds description. The term derives from measurements and boundaries. A typical good metes-and-bounds land description for the parcel of land shown (see Figure 6) reads as follows:

Beginning at a concrete monument in the northerly line of Day Street distant 896.34 feet westerly from the intersection of the northerly line of Day Street and the westerly line of Main Street, thence: (1) N 75° 55′ E 416.00 feet along the northerly line of Day Street to a concrete monument at the southwesterly corner of John Doe, thence: (2) N 19° 45′ W 285.68 feet along the westerly line of John Doe to an iron pipe at the southeasterly corner of Richard Roe, thence: (3) S 61° 47′ W 187.15 feet along the southerly line of Richard Roe to an iron pipe, thence: (4) N 81° 02′ W 184.32 feet, still along the southerly line of Richard Roe to an iron pipe at the southwesterly corner of Richard Roe and in the easterly line of Mary Jones, thence: (5) S 7° 21′ E 312.92 feet along the easterly line of Mary Jones to the place of beginning. An area of 2.26 acres.

(The adjoiners are named to show that there is no gap between the holdings and to serve as a description by bounds. Also, many other data are often listed.)

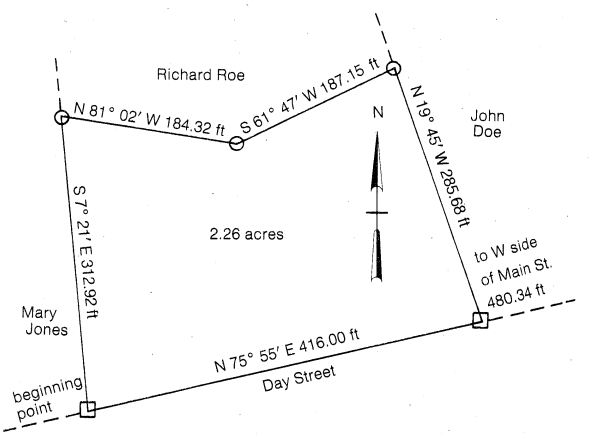


Figure 6: Plot or plat of the metes-and-bounds description (see text).

Special types of surveys. Special surveys for transportation routes, nautical charts, and mine surveying frequently require specialized techniques. All routes of transportation require surveys; particularly accurate must be those for canals, railways, pipelines, and highways. A strip map is made of the most probable route, which is controlled by a traverse and a single line of levels. Photogrammetry usually fills in details. For small surveys, profiles are measured at right angles to the traverse, and contour lines are drawn connecting them. Other details are located by right-angle distances (offsets) from the traverse and by the distance along the traverse to the foot of the perpendicular.

Surveying for transportation routes

The centre line of construction is laid out on the map where the required limits of rate of grade and curvature can be met with the greatest economy. Changes in rate of grade are connected by vertical curves, which are parabolas with vertical axes. Occasionally, in the final layout, spirals are used to connect curves with tangents or other curves. Detailed construction usually is laid out on the ground by measurements from the centre line and the original line of levels. Often the centre line is represented by a base line parallel to it.

The computation of earthwork volumes and the complicated computations required for highway interchanges are nearly always handled by electronic computers.

Nautical charts. Horizontal control for a chart usually consists of triangulation established on land and on stationary positions in the water, such as lighthouses. Traverse is frequently carried out on land, and usually special land targets are built to mark ranges from which the positions of soundings can be determined. These targets, the shoreline, and other objects used for navigation or construction are located from the control.

Mine surveying. Mine surveyors map mine workings, measure the quantities of material excavated, lay out mine tunnels and shafts, and transfer the limiting boundaries on the surface down into the workings.

The fundamental control is almost always located on the surface and is carried out as in ordinary surveys. Wherever possible, ordinary survey methods are also practiced inside the mine. For levelling, shorter than normal rods are used, illuminated by the rodman's flashlight. Where the workings slope, slant measurements are made and their vertical angles recorded, so that horizontal distances and elevation differences can be computed.

Transits are usually equipped with an auxiliary telescope that can be attached to the end of the telescope horizontal axle and adjusted so that its line of sight is parallel to that of the main telescope so that highly inclined sights can be observed.

To measure a downward slant distance, the tape is held at the centre mark of the end of the telescope axle and the other end of the tape at the lower point aimed at by the transitman. Thus the vertical angle of the tape itself is determined. The horizontal angle is measured with the telescope direct and reversed. The average eliminates the effect of the eccentricity of the side telescope.

Because each mine level requires, in effect, a separate survey, it is essential to interconnect these surveys and especially to connect them wherever possible with the surface ground control. When an adit is available, or even a shaft that is not vertical, there is no difficulty. Plumb lines can be used in vertical shafts; a heavy weight, supported on a wire from the top of each shaft, is immersed in oil at the bottom of the shaft to steady it. The greatest difficulty arises when only one vertical shaft is available for surveying operations. In such an event, the azimuth, or bearing, of the surface control must be carried down step-by-step to establish control points at each level. To provide two control points (with only one shaft), two wires with weights are hung as far apart as possible but clear of any contact with the sides of the shaft. Their azimuth is determined as accurately as possible at the surface. At the bottom of the shaft they tend to swing like pendulums, and they are seriously affected by air currents. It follows that many continuous observations are necessary to determine their average position and fix their azimuth with relation to the surface ground control.

The operation can be carried out from the top of the shaft with a transit with an auxiliary telescope by aligning a wire below, so that the line of sight will sweep the wire as far as the shaft permits. The telescope can then be raised and the line of sight to a distant point established.

BIBLIOGRAPHY. H.C. MITCHELL, "Definitions of Terms Used in Geodetic and Other Surveys," *Spec. Publs. U.S. Cst. Geod. Surv. 242* (1948), one of the most complete and accurate works of its kind; J.C. TRACY, *Surveying: Theory and Practice* (1947); PHILIP KISSAM, *Surveying: Instruments and Methods for Surveys of Limited Extent*, 2nd ed. (1956), a text on plane surveying, *Surveying for Civil Engineers* (1956), an advanced and comprehensive text that covers the newest types of survey equipment and techniques, including the theory and use of probability, least squares, and state coordinates, but omitting geodetic reduction, and *Optical Tooling for Precise Manufacture and Alignment* (1962), an excellent text on the theory and use of special instruments for controlling, with extreme accuracy, the construction and adjustment of aircraft and spacecraft jigs, large turbines, satellite launching and control devices, lathes, milling machines, and the like; H.C. IVES and P. KISSAM, *Highway Curves*, 4th ed. (1952), a text treating the theory and practice of all types of highway surveying; C.B. BREED, G.L. HOSMER, and A.J. BONE, *Principles and Practice of Surveying*, vol. 1, *Elementary Surveying*, 10th ed. (1966), and vol. 2, *Higher Surveying*, 8th ed. (1962), well-known texts for all types of surveying; F.R. GOSSETT, "Manual of Geodetic Triangulation," *Spec. Publs. U.S. Cst. Geod. Surv. 247* (1950), description of field procedures; "Formulae and Tables for the Computation of Geodetic Positions," *ibid., 8*, 7th ed. (1933), detail of standard methods using Clark's spheroid of 1866, with complete tables for geodetic reduction; H.S. RAPPLEYE, "Manual of Geodetic Leveling," *ibid., 239* (1948), a discussion of field procedures, and "Manual of Leveling Computation and Adjustment," *ibid., 240* (1948), the most complete text on level nets; K.T. ADAMS, "Hydrographic Manual," *ibid., 143*, rev. ed. (1942), an excellent manual; R.H. SKELTON, *Legal Elements of Boundaries and Adjacent Properties* (1930), a well-known manual including examples and references; W.C. WATTLES, *Land Survey Descriptions*, rev. and suppl. ed. (1956), an outstanding text containing examples and citing legal references; C.M. BROWN and W.H. ELDRIDGE, *Evidence and Procedures for Boundary Location* (1962), an excellent text on land surveying.

(P.K.)

Suvorov, Aleksandr Vasilyevich

Aleksandr Suvorov was the most celebrated and original of all Russian military commanders and the most successful instrument of Russia's territorial expansion in the 18th century. His ability to exploit to the full the patriotism and endurance of Russian peasant soldiers, to fire them with enthusiasm and dash, made him a legend in his lifetime and a national hero who is still held up as a model to inspire the Soviet soldier.

Born in Moscow of a noble family on November 13, 1729, Suvorov was educated at home, joined the Semyonovsky Guards regiment at the age of fifteen, and was commissioned a lieutenant of infantry in 1754. Having gained varied experience during the Seven Years' War (1756–63), in 1763 he became colonel of the Suzdal



Suvorov, portrait by J. Schmidt, 1800. In the State Hermitage Museum, Leningrad.
By courtesy of the State Hermitage Museum, Leningrad

Regiment, for which he wrote a comprehensive manual notable for its realistic emphasis on battle training at a time when the prevailing fashion was for parade-ground manoeuvres in the Prussian style.

His successes during the Russo-Polish conflict of 1768–72 earned him a reputation as a self-willed subordinate and a boldly unorthodox tactician—a reputation that was confirmed when, in 1773, he joined Count P.A. Rumyantsev's army fighting the Turks on the Danube. His two successful descents on Turtukai, his brilliant defense of Hirsov, and, above all, his successful direction of the Battle of Kozludji (1774) showed him to be an incomparable commander of the field, but after he helped to suppress the Russian peasant revolt led by Yemelyan Pugachov in the same year and subsequent service in the Crimea, his career seemed to founder.

A mission to Astrakhan to prepare an expedition into Persia proved abortive, and, even after he returned to a more active posting in the Caucasus in 1782, the establishment repeatedly passed him over. His marriage to Varvara Prozorovskaya, contracted ten years before, broke down in 1784, Suvorov refusing to recognize her child, Arkady, as his son, though he remained devoted to his daughter, Natalya. Returning to St. Petersburg and the court, Suvorov felt consistently ill-at-ease—conscious of his scrawny appearance, rough manners, and inability, as he put it, to "endear myself to my superiors," to which he attributed the slow progress of his career. But in 1787, after almost 15 years, he was at last promoted to the rank of general.

Victories over the Turks

In the Russo-Turkish War of 1787–91, he successfully defended Russian territory on the north shore of the Black Sea from Turkish attack. His victory over the Turks at Kinburn (October 1787) restored him to the limelight, and though his alleged rashness and drunkenness at the siege of Ochakov, in 1788, earned him official displeasure and exclusion from fighting commands, in 1789 he was sent to Moldavia (in present Romania) to conduct joint operations with Russia's Austrian allies. The ensuing victories at Fokshany (now Foçsani, Romania), and his brilliant and bloody storming of the stronghold of Ismail (1790), confirmed his popularity in the army, but his sense of frustration was revived when he was transferred to Finland in 1791. Hostile rumours circulated against him at court, and, as younger, less deserving men overtook him in rank, his embitterment deepened, and he became noticeably more eccentric.

Then, in 1794 he was recalled to crush the nationalist-revolutionary movement in Poland—which he did with ruthless efficiency. The slaughter involved in his storming of the Warsaw suburb of Praga (which he justified as shortening the war and saving lives) shocked Western opinion, but it earned him a reward of 7,000 serfs and the promotion to field marshal he had long coveted.

Subsequently appointed commander in chief of the

southern army, Suvorov set about molding it according to principles he set out in his celebrated military treatise, *Nauka pobezhdat* (*Science of Victory*)—remarkable for its clipped, colloquial style, its novel emphasis on speed and mobility, and the use of the bayonet and accuracy of fire. But on the accession of the Emperor Paul I (reigned 1796–1801), the old Prussian parade-ground emphasis was reimposed, and, refusing to hide his opposition to it, Suvorov was dismissed.

Suspected of treason, he was kept under close surveillance, but a thorough investigation produced no evidence against him, and the emperor restored him to favour. But Suvorov continued to show his dislike of the way the army was being managed and soon retired again.

Italian campaign

Fiercely antirevolutionary, he responded with enthusiasm when he was recalled in February 1799 to command a Russo-Austrian army against the French in northern Italy. A series of rapid victories (April–August 1799) resulted in his capturing Milan and almost completely expelling French forces from Italy. These successes made him a hero to those opposed to the French Revolution but also aroused Austrian misgivings about Russian ambitions in the west. Suvorov had wished to invade France, but was ordered across the Alps to join a Russian force under Aleksandr Rimsky–Korsakov that was being threatened by the French in Switzerland.

Fighting his way across the mountains, he reached Altdorf, only to hear that Korsakov had already been defeated and that the Austrian divisions in Switzerland, intended to support him, had retired. Surrounded by the enemy, short of ammunition and supplies, and with the first snows of winter falling, Suvorov set out to extract his army from an apparently hopeless position. Though old and sick, he succeeded in arousing his starving and dispirited troops, broke out to reach Glarus, and, repulsing a pursuing French force, succeeded in escaping with three-quarters of his men, achieving one of the most remarkable exploits in the annals of war.

Though disillusioned with the Austrians, Suvorov still envisaged fighting another campaign, but in January 1800, having been promoted to the unprecedented rank of generalissimo, he was recalled. He arrived at St. Petersburg, exhausted and gravely ill, only to find that, on account of a petty misdemeanour, the hero's welcome planned for him had been cancelled and that he was excluded from the court. He died on May 6, 1800.

BIBLIOGRAPHY. PHILIP LONGWORTH, *The Art of Victory: The Life and Achievements of Generalissimo Suvorov (1729–1800)* (1965), is the standard biography in English; G.P. MESHCHERYAKOV (ed.), *A.V. Suvorov: Dokumenty*, 4 vol., (1949–53), is the fullest collection of documents on the subject.

(P.L.)

Swamps, Marshes, and Bogs

Swamps, marshes, and bogs are commonly considered together as wetlands, distinct from lakes, ponds, and streams. Wetlands combine characteristics of both aquatic and terrestrial areas. They are covered with emergent vegetation of higher plants and mosses, rather than only algae, which are typically found growing in open water. The soil or root mat from which the plants grow is usually saturated with water for at least a major portion of the year and may be covered with water most of the time.

The distinction between swamps, marshes, and bogs varies from place to place. Bogs are often thought of as places that look dry but that wet the feet of walkers, whereas marshes are places with visible standing water. A useful distinction used in this article is that marshes and swamps are characterized by plants that grow on mineral soils. The area is a swamp if the plants are principally trees, and a marsh if they are principally grasses. Examples are mangrove swamps, cypress swamps, cedar swamps, cattail marshes, and salt marshes. Bogs, often called moors, are wetlands with organic soils or peats. The vegetation in bogs of temperate regions is rounded in outline and cushion-like, in contrast to wet woodlands that are swamps or wet prairies that are marshes.

Swamps and marshes occur on river floodplains behind various sorts of dams and along sheltered intertidal seacoasts in tropical regions. Bogs are characteristically more stagnant areas dominated by the growth of mosses; these usually belong to the genus *Sphagnum*.

Swamps and marshes are found throughout the world wherever there is shallow stagnant water in which higher plants can grow. Bogs are also found throughout the world, although they are much more abundant in temperate and colder regions, including the mountains in tropical climates. Wetlands are frequently formed as a result of the filling in of ponds and lakes, by erosion from the surrounding land or by the accumulation of various organic sediments.

Importance of wetlands

Swamps, marshes, and bogs have been traditionally thought of as undesirable features of the landscape. To a large extent this is because they are breeding grounds for mosquitoes and are thus associated with diseases such as malaria and yellow fever. Wetlands have always been objects of large numbers of drainage projects in an effort to convert them to farmland and sites suitable for construction. It is now recognized that these wetlands are of value to man, the remaining ones especially so because many have been destroyed.

Wetlands, lying in depressions, hold large quantities of freshwater by retarding the rate at which it can run off into rivers and so eventually to the sea. This retardation allows water to percolate into the ground and thus recharge the groundwater. This is of particular importance in countries such as the United States, where highly mechanized industry and farming are threatening to exhaust groundwater supplies in many areas.

The retardation of runoff is also important in the prevention of flooding. Preservation of inland wetlands is a less expensive flood-control measure than the construction of dams and gives added benefits in the field of conservation.

Coastal salt marshes and mangrove swamps, when allowed to remain as a barrier between the sea and human construction, provide a self-repairing protection against damaging storm waves.

Commercial uses of wetlands can be direct or indirect. Peat is mined from bogs in various parts of the world and then dried and used as fuel. Peat fires, used to smoke malt, give the characteristic flavours to Scotch and Irish whiskies. Peat is also mined for horticultural purposes, as is *Spaghnum* moss itself.

Peat production in the United States is between 500,000 and 600,000 tons a year and about half that in Canada. East and West Germany combined produce 2,000,000 to 4,000,000 tons a year, Ireland produces 5,000,000 to 6,000,000 tons a year, and the Soviet Union from 150,000,000 to 200,000,000 tons a year. The large Soviet production is used partly in the generation of electricity.

Swamp forests are sources of timber. Although cypress can grow in upland soils, it is found in large stands only in swamps to which this tree is particularly well adapted. Marshes support fur bearers such as muskrat, mink, and otter.

Coastal marshes are of indirect commercial value because they provide nursery grounds for many species of commercially valuable fish and shellfish. Wetlands are of commercial value also to industries connected with some aspects of recreation. Marshes, both coastal and inland, furnish nesting and feeding areas for waterfowl. Salt marshes support many species of sport fish.

The Florida Everglades region, in spite of widespread drainage, is still one of the world's largest combinations of marshes and swamps, and it makes an important contribution to Florida's tourist industry. Finally, wetlands have a purely recreational value in preserving variety in the landscape and in preserving species of animals and plants that cannot survive elsewhere.

This article treats the physical and biological characteristics of swamps, marshes, and bogs and contains a discussion of wetlands of the geological past. In the latter regard, see COALS and articles on the CARBONIFEROUS PERIOD for relevant information. See also COASTAL FEATURES; LAGOONS; and ESTUARIES for additional coverage of nearshore environments.

BOGS

Types and characteristics

Bogs are spongy, poorly drained, wet areas with peaty soils in which drainage typically gets worse as the bog grows. This results from the water-retaining nature of the plants that grow in bogs or of the peat that accumulates there.

Bogs can be divided into three types: (1) typical bogs of cool regions, dominated by the growth of bog mosses, *Sphagnum*, and heaths, particularly *Chamaedaphne* (northern bogs with trees growing on them are often called muskegs); (2) fens, dominated by grasslike plants, grasses, sedges, and reeds; and (3) tropical tree bogs, in which the peat may be formed almost entirely from tree remains. Typical, or *Sphagnum*, bogs are highly acid with a pH (index of acidity–alkalinity) of less than five (seven being neutral) and are associated with waters containing no more minerals than are contained in rainwater, often the only source of water for a bog. Fens are watered with groundwater that has some dissolved minerals and that has a pH above five; that is, it is only moderately acid. Fens and bogs are often associated in one area that usually is called a bog. Tropical bogs occur only in areas where the water is very low in minerals. They are less common than swamps but still cover extensive areas in Malaya, Indonesia, tropical South America, and Africa.

Typical bogs have simple floras. Other than the *Sphagnums* and heaths, there are a few sedges and grasses, such as cotton grass; insectivorous sundews; pitcher plants; and many orchids. Desmids, a group of unicellular green algae divided into symmetrical halves, are characteristic of bogs. Animals are not common on bogs.

Sphagnums are large mosses with large empty cells with pores opening to the outside that lie between the chlorophyll-bearing cells of the leaves. These empty cells readily absorb and retain water, giving a spongelike quality to the moss. *Sphagnum* absorbs minerals (cations) from the water, replacing them with acid (hydrogen ions), and thus makes the water around itself more acid. *Sphagnum* also possesses mild antibiotic properties. (These characteristics made *Sphagnum* popular as a wound dressing; it was used extensively in World War I.)

The saturation of the moss with water retards passage of air, so that parts of a mass of *Sphagnum* more than a few inches from the surface are usually anoxic. The combination of lack of oxygen, lack of minerals, and highly acid condition greatly retards the action of bacteria and fungi, the usual decay organisms. With the retardation of decomposition of the dead moss, a *Sphagnum* peat develops under the living plants. This is particularly the case in areas where there is a mean annual temperature of below 10° C (50° F), which also retards decay.

Bogs are most common in parts of the world glaciated during the Pleistocene Epoch (10,000 to 2,500,000 years ago). They cover vast areas in the tundra and boreal forest regions of Canada, northern Europe, and the Soviet Union. Areas of high rainfall farther to the south, such as the wetter parts of the British Isles, also contain extensive bogs. Glacial ice created many local depressions by scouring underlying rocks and spreading an uneven deposit of till on the ground. With the melting of the ice, these depressions filled with water. If the mineral content of the water was low, the ponds so formed were colonized by *Sphagnum*, which converted them to bogs.

Once bogs are formed, they retard the development of efficient drainage by inhibiting water movement and slowing erosion of the soil or rocks on which they rest. Thus, bogs tend to be long-lived if temperatures remain low and sufficient excess of rainfall over evaporation exists to prevent their drying out. If they do dry out, upland plants will colonize the former bog.

In some areas in Alaska, an unstable equilibrium between bog and forest exists. Bogs that begin by filling ponds tend to spread as the water that collects around their edges is warmed by the sun and causes the permafrost under the surrounding forest to melt. This undermines the forest, which then falls into the bog. The bog then spreads outward and gradually fills both with peat and with depositions of silt from river flooding, until finally it becomes high enough for trees to grow on the surface. As the trees grow, the ground is shaded, the peat compacts, water movement through it slows, and the ground becomes permanently frozen again. The expansion that occurs with the formation of ice elevates the ground to complete the cycle.

Development from lakes

Small lakes in glaciated regions have also often developed into bogs if they were not drained by headward erosion of stream beds or completely filled with terrestrial sediments. The lake begins to fill by the development of a floating mat of vegetation next to the shore. This may become a fen, with grasses and sedges principally involved, if the lake water has sufficient mineral content. A lake on resistant rocks, poor in minerals, will develop a floating bog mat wtih *Sphagnum* and *Chamaedaphne*. In larger lakes, wave action will not allow such floating mats to develop.

As the mat grows into the water, it floats at the surface supported by air in the plant tissues. Upward growth shades the lower parts and these die, forming an increasingly thick floating mat, the top of which remains only a few inches above the level of the lake surface. Slow decay and mechanical disturbances break off waterlogged bits from the bottom of the mat. These sink and collect on the lake bottom so that the lake is filled both from the bottom up and the top down. The material collecting on the lake bottom is very fine, only a little denser than the water. It does not compact into a firm layer but forms a false bottom layer through which heavy objects fall to the solid original lake bottom.

From the surface down, the layers are (1) floating bog, (2) clear water, (3) false bottom, and (4) true bottom. With the continued thickening of the mat there is less influence of the lake water on plant growth, and *Sphagnum* usually begins to invade the surface of the mat even if it was previously dominated by grasses. With the growth of the moss a true bog is formed and various heaths invade the mat, especially *Chamaedaphne*. With continued thickening, trees may begin to grow, the first usually being larch (*Larix*). Black spruce may invade in the last stages of bog development. From a distance it may be difficult to detect the original boundary between the upland and the now filled lake.

During much of this process the vegetation is floating. The bog is called a quaking bog to indicate the instability of the surface, which will sink slightly beneath a weight. It is even possible to break through the vegetation into the water beneath. Both men and animals have drowned this way. Nonfloating bogs may also quake if the peat is thick and spongy.

Creation of raised bogs

Ultimately, by upward and centripetal growth, the bog fills the lake completely. The original glacial basin will then contain a bottom layer of inorganic sediment combined with organic debris derived both from lake production and from terrestrial sources surrounding the lake. On top of that, a layer of false-bottom material, compacted by the weight of overlying peat, may be visible. Bog peat fills the remainder of the basin. Growth of the bog does not necessarily stop at this stage for, if rainfall is sufficient, the water-retaining properties of *Sphagnum* are sufficient to maintain a wet, stagnant environment above the original water table represented by the lake surface. Thus, continued upward growth of the bog plants creates a raised bog. The raised bog is similar to the ordinary bog except that it does not lie in a depression but is raised above the surroundings. A moat containing some open water typically surrounds a raised bog where water drains from the raised bog and surrounding upland. Because the moat receives upland drainage it may be a fen. The raised bog itself receives only rainwater. Because rain is very low in mineral content, the raised bog is more completely dominated by the characteristics of *Sphagnum* than that of the floating bog. Heaths, larches, and black spruce, which grow fairly well on the floating bog, survive only as stunted specimens around the edges of the raised bog.

Raised bogs can develop elsewhere than in old lake basins. They may even exist atop hills. A moister climate is needed than that necessary to support a bog in a basin, but, given enough rain and mist and a drainage sluggish enough for bog mosses to start growth, a bog may result.

Here again, the water-retaining ability of the mosses, and their effect in removing minerals from the environment, favour their own growth and inhibit the growth of taller plants that might shade the mosses and take over the site. These bogs, perched on the tops and sides of hills, full of water, and very poorly anchored to the underlying substratum, are not stable landscape features. Bog slides, which most often occur after heavy rains, do considerable damage in such places as Ireland.

Properties and accumulations of peat

The peat underlying a *Sphagnum* bog is composed largely of partly decomposed moss. There may be some inclusion of windblown particles, pollen, and dust. The water content of peat may be as high as 90 percent. The ash content of dried peat varies from 2 to 20 percent, with lower values more usual because the higher ash amounts come from sand and clay blown into the peat when it was at the surface. Other chemical characteristics of peat underlying bogs are the absence of free oxygen; the presence of carbon dioxide at high pressure, though in small amounts; low electrolyte concentration; and high acidity.

Plant nutrients, or electrolytes, are too low to support the growth of most higher plants. This is the principal reason why few higher plants can grow on bogs. Peat is commonly used as a medium for growing seedlings, but its suitability lies in its ability to hold water and its freedom from weed seeds. Nutrients have to be added to the medium if the seedlings are to be grown in it. The properties of *Sphagnum* in inhibiting growth of fungi have led to the use of *Sphagnum* for germinating seeds because it prevents the fungus attacks known as damping off.

In general, temperatures in lowland tropical areas are high enough for organic decay to be too fast for large quantities of peat to accumulate. But in areas with very high rainfall and with groundwater of very low mineral content, bogs may occur. As is the case in colder regions, the bogs may fill low basins, or they may develop into raised bogs. The plants of these bogs are mainly broad-leaved evergreen trees, which may be as tall as 100 feet (30 metres). Palms and screwpines may also occur. The number of kinds of plants of these bogs is very limited in comparison with the surrounding forests, just as is the case in colder bogs. Toward the centre of the raised bogs, trees are shorter or absent. Grasses and sedges cover more of the ground and open pools of water may occur. *Sphagnum* does not grow in tropical bogs to any extent, and the peat is composed of the remains of seed plants. The limited number of species is caused entirely by the lack of any groundwater supply of minerals and the rapid removal, by heavy rainfall, of those that arrive in dust.

SWAMPS AND MARSHES

Swamps and marshes, wetlands with mineral soils vegetated principally with trees or grasses, respectively, are found throughout the world. They exist in areas with poor drainage and sufficient water supply to keep the ground waterlogged, and they have a high enough supply of minerals in the water to stimulate decay of organisms and prevent the accumulation of organic materials. They are often found in regions of low relief associated with rivers that supply the water. Marsh and swamp soils do not usually develop into peat, although the salt-marsh peats are a notable exception to this rule. In salt marshes the seawater supplies enough minerals to compensate for the lack of a mineral soil, and it supports the growth of grasses. Swamps and marshes also differ from bogs in that they typically grow from the bottom up. Floating mats are less common than in bogs.

Relations to rivers and river deltas

Freshwater restriction. Marshes are common at the mouths of rivers, especially where extensive deltas have been built. The river brings a steady supply of water. The gradient of the river approaches zero at the sea, where flow is sluggish. Because the delta is deposited by sediment settling from the river water, the land that is built will be poorly drained at its driest and will often be underwater. Sediment supplied by the river has often been eroded from the surface soils of the drainage basin and is thus very rich. The combination of water supplied steadily at a low rate over a waterlogged but rich soil creates a perfect environment for marsh grasses.

Fibrous-rooted grasses bind the muds together and further hinder water flow, thus encouraging the spread of both the delta and the marsh. Marshes occur in the deltas of most of the world's great rivers. In Europe well-known river-mouth marshes include those of the Camargue in the Rhône Delta, the Guadalquivir in Spain, and the Danube in Romania, all of which are famous as bird sanctuaries. In the Middle East, both the Nile Delta and the delta of the Tigris–Euphrates have extensive marshes of historical importance. The marsh dwellers of the Iranian marshes have developed a unique culture adapted to life in the wetlands.

Marshes and extensive mangrove swamps occur in the deltas of the Mekong in Vietnam and the Amazon in Brazil. In the United States, the most extensive delta marshes are those of the Mississippi River. Other than building delta marshes, rivers in mature valleys frequently have extensive marshes and swamps along their sides. Floodplains elevated only a few feet above river level, abandoned river channels, and oxbows may have standing or sluggishly flowing water for appreciable parts of the year and thus support swamps and marshes.

River swamps are abundant along the coastal plain in the southeastern United States. They are delightfully described in their primeval condition by William Bartram in his account *Travels through North and South Carolina, Georgia, East and West Florida* written in 1791. These were the swamps that Francis Marion used so successfully to escape the British forces during the Revolutionary War and the source of his nickname, "the old swamp fox."

The great Dismal Swamp, of North Carolina and Virginia, actually a mixture of waterways, swamps, and marshes, is a coastal-plain swamp although it is not associated with a large river. The Mississippi and its lower tributaries, such as the Red River and parallel rivers in eastern Texas, have extensive swamps along the slugggish portions that flow through the coastal plain. The Paraná and Paraguay rivers in South America have extensive swamps and marshes along their courses.

Some low-lying areas with poor drainage at the heads of more extensive drainage patterns contain wetlands. The vast swamps of the Rio Negro, which drains into the Orinoco and Amazon, is one such. Another is the Pripet Marshes and fens that historically have served as the natural boundary between Poland and Russia. In the United States a mixture of swamps and marshes in Georgia called the Okefenokee Swamp is the source of the Suwannee River.

Drainage basin depressions

In some places basin-like depressions in the Earth's surface trap waters and make wetlands. Most such areas are drained someplace along their rim by a river that is impeded at that point sufficiently to dam water at times of high flow and create marshes and swamps. The world's two largest rivers, the Amazon and the Congo, fall into this category. Both of the great basins named after these rivers have extensive wetlands. The papyrus marshes of the upper Nile in southern Sudan lie above dams of resistant rocks of the cataracts. The marshes along the Danube lie above a natural dam called the Iron Gate, at which point the river flows between the Transylvanian Alps and the Balkan Mountains. Marshes adjacent to the Sahara surround Lake Debo on the upper Niger River in Mali. They were formed by a similar natural obstruction in that river.

The Okavango Marshes east of the Kalahari desert in Botswana are perhaps the best example of marshes formed in an interior, closed basin that has no drainage. Other basins without outlets like that of the Great Salt Lake in Utah have accumulated too much salt for marsh growth.

The Florida Everglades constitutes a unique marsh–swamp combination growing on a limestone base. Because the region is near sea level the water from the abundant rains does not drain but remains on the surface. The Everglades is similar to a huge, shallow, slowly flowing river. The area is an ideal marsh habitat but the Everglades is different from usual marshes. The soils are alkaline because of the limy base, and the water is clear.

Large cypress swamps occur to the northwest of the Everglades.

There are also enormous numbers of small swamps and marshes. Some areas, such as the northern Great Plains of the U.S., have so many small marshes that they are a characteristic of the landscape. These small marshes formed because the landscape left by the retreat of glacial ice was so irregular and so poorly drained that countless little depressions were filled with water each spring. As snow melted, the depressions supported the growth of temporary marshes, which then dried up during the summer. Larger depressions were occupied by ponds. These gradually became marshy as they filled in with sediment.

The succession of plants in these marshy areas is similar to that of bogs except that there is no floating mat. The ponds fill from the bottom up and have marsh plants rooted in bottom muds. The same filling in converts beaver ponds and abandoned stream meanders to small marshes.

Many of the small marshes of the Great Plains have been destroyed as a result of agricultural drainage schemes. Now, in some areas, they are being revived by destruction of the drainage systems. This has been undertaken for waterfowl and shorebird conservation. These birds depend on small marshes as nesting places. Destruction of drainage systems also aids in the conservation of water, as water can now enter the ground supply rather than run off in the drainage structures. Topography and water supply are the two most important features in determining the distribution of freshwater swamps. The nature of soils and bedrock is of importance in determining the drainage in a region, but wetlands may exist locally on any base from sands to impervious rock.

Plant occurrence and accumulation

The flow of water through wetlands is slow because of low gradients and retarding effects of the vegetation. Dead plant matter settles rather than being washed away. The slow replacement and lack of turbulence in the water results in a low rate of oxygen supply. Decay of the dead vegetation quickly uses up what oxygen is supplied, so that the mud and bottom waters are low or lacking in oxygen content. Under these conditions, the decay of organic matter is incomplete. This causes an accumulation of the more resistant fraction (humates and tannins) in the substratum. The familiar swamp water, varying from yellow to such a deep brown that it resembles strong tea or root beer, is the result.

Many species of plants grow in freshwater marshes and swamps, but relatively few are dominant. The number of species is few compared to the numbers that grow on well-watered but not waterlogged land. Grasses, grasslike sedges, and reeds or rushes are of major importance. Cattails (*Typha*) and common reeds (*Phragmites*) are familiar swamp species around the world. Papyrus, a sedge, is widespread in the tropics. Wild rice is of some commercial importance, but true rice is undoubtedly by far the most important marsh plant and supplies a major fraction of the world's grain. Bald cypress is an example of a tree adapted to growth in swamps, but gums, willows, alders, and maples are also common. Tropical swamps have many tree species including palms.

Flooding by seawater. In the foregoing discussion of swamps and marshes, the structure of the land has been noted as a factor that restricts the drainage of freshwaters, thus causing a wetland to be formed. Salt marshes and swamps, however, are formed by seawater flooding and draining, which exposes flat areas of intertidal land. Where flooding is irregular, such areas may become too salty for higher plants to grow.

Salt-marsh grasses will not grow on permanently flooded flats, although mangroves may grow in water up to a few feet deep. Growth is also prevented where the flooded land is subject to strong currents and is therefore unstable. But regularly flooded, protected areas develop salt marshes in temperate regions and mangrove swamps in tropical and subtropical regions. Mangroves will grow in pure sand at the edge of the sea. Extensive swamps develop mainly where land runoff is sufficient to bring a supply of sediments that accumulate and extend the swamp. The deltas of the Mekong, Amazon, Congo, and Ganges and the north coast of Australia and of Sumatra have notable and extensive mangrove swamps. Salt marshes, which are particularly extensive along the east coast of the United States, formed after the retreat of the last continental ice sheet. They are also common in the Arctic, northern Europe, Australia, and New Zealand and to some extent throughout the world.

Salt marshes and sea-level change

Salt marshes are covered with species of grass. In eastern North America, coarse-leaved *Spartina alterniflora* grows at midtide level and the fine-leaved *S. patens* (salt hay) at high-tide level. *S. townsendii* occupies a similar place in northwestern Europe, along with other grasses, *Punccinellia* species. The latter are also important in Arctic salt marshes. Species of *Suaeda*, *Atriplex*, and *Salicornia*, all members of the Chaenopodiaceae, are among the other important marsh plants particularly in saltier, less well drained parts of salt marshes.

On the Atlantic coast of the United States, salt marshes began to form about 4,000 to 5,000 years ago when the rapid postglacial rise in sea level slowed. *Spartina* began to grow at midtide levels in sheltered areas. The grass stems retarded water movement. Sediment accumulated along with roots, which raised the level of the marsh. By this means, coupled with the addition of new sediments from land erosion, marshes grew out into the water and up to high-tide level.

The accumulation of peat composed largely of dead *Spartina* occurred only where the peat was anoxic, which depended on its being waterlogged. The marsh did not grow above high-tide level unless river-flood or windborne sediments were deposited. At high-tide level the marsh supported mainly *S. patens* north of Cape Lookout, North Carolina, and a dwarf form of *S. alterniflora* to the south. As sea level rose marshes grew by spreading out over neighbouring uplands and into the bays and coves in which they had formed.

The sea level in Massachusetts has risen about eight metres (27 feet) in 5,000 years. Peat in the oldest parts of the marshes is about eight metres deep, the amount of rise in sea level. Primary drainage patterns developed on the flats on which the marshes grew.

Secondary patterns result from processes within the marsh. Most water-borne sediment settles out at the edge of the marsh, where water first meets resistance to movement from the grass. Levees tend to form along the creeks and dam up water in interior parts of the marsh. When an area retains standing water continuously between tides, the marsh grasses die and pools are formed. These may subsequently be drained, with the formation of secondary drainage patterns by erosion of the levees or headward erosion of the primary drainage creeks. The same general patterns of water depth tend to occur also in river marshes and swamps with times of flood and drought corresponding to high and low tide.

The creeks in a mature salt marsh are in hydrodynamic equilibrium, maintained by the currents necessary to flood the surface of the marsh. Strong currents occur just before high-tide water. At this time the area being flooded increases from the area of the creeks to the entire marsh. There is a similar strong flow in the reverse direction when the marsh is being drained on the ebb. Whereas the flow in freshwater marshes and swamps tends to be unidirectional and sluggish, the flow in salt marshes is back and forth and varies from strong in the marsh along the creeks to sluggish in portions far from them.

In regions with appreciable rainfall, land drainage is impeded by salt marshes so that there may be a progression from the salt marsh through brackish marshes to freshwater marshes and swamps. Among the most characteristic species of plant in the brackish marshes in the United States are the black rushes, *Juncus gerardii* in the north and the larger and stiffer leaved *J. roemarianus* in the south.

WETLANDS IN THE GEOLOGICAL PAST

It is logical to assume that ever since there have been land plants there have been bogs, swamps, and marshes. The deposits of coal, the remains of those preceding wetlands, prove the point. Coal has been formed in every geological

Association with coal deposits

period since the spread of land plants in the Devonian Period (345,000,000 to 395,000,000 years ago), but, during the Late Carboniferous, some 250,000,000 years ago, the first significant amounts of coal were deposited. It is from those deposits that the world's most valuable coal comes.

Carboniferous coals occur in the British Isles, France, Germany, the Soviet Union, China, and North America. There is very little coal known in Africa, South America, or Oceania. In the United States the coal was largely deposited in the east central part of the country in three main fields slightly separated from one another: the Appalachian, which extends from Pennsylvania to Alabama; the Illinois Field; and the Mid-Continent Field, extending from Missouri to North Texas. The coal swamps lay in a basin with the Appalachian Highlands on the east and the Canadian Shield on the north. The basin sloped toward the sea to the south with a very gentle gradient of only one or two feet to the mile. The entire basin was apparently only a few feet above sea level and was submerged at intervals, so that coal seams alternate with sandstones, shales, and marine limestones that were deposited when the swamps were destroyed by river sands or flooding by the sea.

Under the coal seams are clay deposits that appear to be the soils in which the coal swamp plants grew. In some areas, remains of trees are still in position extending up through the coal that represents the accumulation of remains of plants growing in the swamp. These plant remains are preserved in the coals in sufficient abundance to permit determination of the appearance of the swamps.

These remains have no indications of annual or seasonal growth rings, thus indicating that the climate must have been tropical or at least subtropical. The trees had widely spread roots and often aerial roots, both of which are typical of swampy conditions. The alternation of marine and coal deposits, as well as the small slope, indicates that water movement must have been very sluggish. If there were a very rainy climate as well, conditions would have been like those found in the bog forests of the tropics where, under conditions of very low mineral water, peat is accumulating today from the remains of partly decomposed trees.

The sulfur content of much of the Mid-Continental coals is high. When the coal is burned, sulfur is put into the air and forms one of the irritants in air pollution. Sulfur is believed to have been deposited with the coal, pointing to the fact that most of the coal swamps were brackish rather than completely fresh. Sulfur is not common in freshwater but sulfate is one of the common constituents of sea salt. The coals of Nova Scotia and New Brunswick and of eastern Pennsylvania are low in sulfur; this is an indication that the swamps in those particular regions were probably fresh.

Plants of the former swamps

The Carboniferous coal swamps, more correctly called coal bogs, were vegetated with primitive vascular plants related to modern club mosses, horsetails, and ferns. These are spore-bearing plants whose spores accumulated in parts of the bogs in almost pure deposits. The deposits formed cannel coals, a very fine coal that burns with a bright flame.

Though their modern relatives are usually small plants, the swamp-forming species were trees up to two metres (six feet) in diameter and 30 metres (100 feet) in height.

As well as the Carboniferous coal swamps, extensive deposits of coals were formed in the Jurassic Period (136,000,000 to 190,000,000 years ago), the Cretaceous Period (65,000,000 to 136,000,000 years ago), and the Paleocene Epoch (54,000,000 to 65,000,000 years ago). By Cretaceous time, coal swamps were inhabited by modern types of plants, often belonging to the same genera that occupy swamps today.

Things that occurred during the times the coals were being formed can often be determined by examining the sequence of rocks in the coal beds. In a similar manner it is possible to look at the sequences of deposits laid down in modern bogs and other sediments, such as lake muds, in which organic matter is preserved and determine the history of the bog and its vicinity by noting changes in plant parts and pollen preserved in the peat. This method is subject to uncertainties, however, because the abundance of a particular type of vegetation does not correspond exactly to the abundance of its pollen in peat.

Pollen of pines and grasses is wind-borne and overrepresented compared to plants with heavy insect-borne pollens. Herbaceous-plant pollen may be largely excluded from a bog by a surrounding ring of bushes. Relative changes in any one type of pollen with depth in the peat, however, will yield information about changes of climate and plant cover with time.

Peat formation in Massachusetts

Studies of some peat deposits and radiocarbon dates from Boston, Massachusetts, show that the earliest remains are from a freshwater swamp growing on glacial clay, 7½ metres (25 feet) below present sea level. This was followed about 5,700 years ago by a reed swamp (*Phragmites*), covered in turn by a short-lived salt marsh. Above this are remains of Indian fish weirs from 4,500 years ago.

Cores from a salt marsh on Cape Cod, Massachusetts, were taken to nine metres (29 feet) below sea level in an upland soil. The first definite saltwater peat was found at a depth of eight metres (27 feet) and was laid down about 5,000 years ago. The climate was that of a postglacial warm, dry period, judging from pollen from oak, hickory, birch, and grasses. At a depth of about five metres or 17 feet (1,900 years ago) these pollens became less abundant. Those from pine and hemlock increased, indicating a change to a colder, moister climate. At a depth of about 53 centimetres or 21 inches (400 years ago) there was an enormous increase in grass pollen, a decrease in pine, and an increase in oak, all a likely result of the deforestation of Cape Cod by settlers during the colonial period. Finally, at a depth of 30 centimetres (12 inches), corresponding to the middle of the 19th century, there was a decline in grass and oak pollen and an increase in pine pollen. This represents the time farms on Cape Cod were abandoned as new farmlands were opened in the Midwest. Pine colonized the former fields.

In regions that escaped being covered by ice during the last stages of Pleistocene glaciation, the pollen records in bogs extend much further into the past and give a record of conditions in the vicinity of the ice. One such bog, Humptulips Bog on the Olympic Peninsula in Washington, was only two to eight miles from the edge of the ice at various times. The record extends back more than 50,000 years. Sediments in the core show intervals during which the present bog was filled with silt, lake sediment, fibrous peat, and peat formed from *Sphagnum*. This reflects the great changes that took place in climate and erosion near the edge of the ice.

During most of the period represented by this core, vegetation around the bog was that of tundra or tundra interspersed with trees of types that grow in colder climates. By comparing the fossil pollen in the core with pollen presently being deposited at successively colder temperatures along the Pacific coast, a temperature record was made to correspond with the pollen record in the core. Throughout the 50,000 years, it was colder than at present. The July average was 6° C (11° F) colder during the ice maximum. During an interglacial period, from 20,000 to 30,000 years ago, climatic conditions approached that of the present, but the temperature was still about 2° C (3° to 4° F) lower than at present. The postglacial warm period referred to in relation to the Massachusetts bogs appeared to have been about 1° C (2° F) warmer than the present.

For a more detailed discussion of the utility of pollen studies in such interpretive problems, see POLLEN STRATIGRAPHY.

LIFE IN WETLANDS

Relatively few species of organisms have adapted to the special conditions of wetlands. Raised bogs impose the most rigorous conditions and have the fewest species of both plants and animals. Salt wetlands are next, with highly saline inland-basin marshes having fewer species

than wetlands flooded by ocean tides. Permanent freshwater wetlands are more favourable, but the greatest variety of inhabitants are found in temporary freshwater wetlands where the soil dries out for a part of the growing season.

Effects on plant roots

Influence of the substrata. In all wetlands, plants must contend with substrata low in oxygen and usually completely anoxic. The substratum is also typically soft and offers relatively poor support or anchorage. Plant roots survive by getting a supply of oxygen from the aerial parts of the plant and by being able to metabolize without oxygen, producing ethanol as an end product rather than carbon dioxide. The oxygen-transport mechanism depends on air spaces within the tissues (aerenchyma) through which oxygen passes by diffusion to the underground parts and carbon dioxide returns. The air spaces in grasses communicate with the air through the leaves, but in many swamp trees the roots send up special pneumatophores into the air. The knees of cypress are such, as are the striking fingerlike projections from black mangrove roots.

In nonarboreal marsh plants, the air spaces that typically occupy the entire centre part of the roots are also mechanically important to the plants. With a large empty space in the centre, root tissue is more spread than it could otherwise be, and provides more friction with the soft soil that holds the plant in place. The hollow plant is also more resistant to bending, which in turn, lessens damage by water movement.

Tropical swamp trees often develop buttresses that apparently help support them, though buttresses are also common on many upland trees in the tropics. Generally, all swamp trees lack deep-growing roots. Roots tend to stay near the surface, probably because of the lack of oxygen in the soil. Being near the surface gives a large lateral spread to the root system which also gives the tree support against lateral stresses from winds or water flow. The shallowness of the roots also often produces the appearance of buttresses in temperate swamp trees because the beginnings of the roots usually hidden in the soil are visible at the surface.

Different species of trees that grow in temporary swamps differ considerably in their resistance to submersion—*i.e.*, to lack of oxygen. This is serious during the submersion period but is not a problem in drier periods, at least in the shallow layers of the soil. In North American swamps alders and willows will survive or even thrive on land immersed for periods as long as one month, whereas red gum survives only about two weeks. Cottonwood begins to show the effects of submergence after only two days and survives only one week.

Animal adaptations

Animals have adapted to the limited supplies of oxygen in somewhat similar ways. Rat-tailed maggots (*Tubifera*), for example, survive in shallow marshes by means of a telescoping, tail breathing tube that they extend to the water surface for air. Some larvae of shore flies (Ephydridae) and some nematodes take advantage of the air spaces in plants and obtain oxygen from that source. Many small marsh animals have great resistance to lack of oxygen; for example, many nematodes can live indefinitely in the complete absence of oxygen. This ability is essential for such minute animals that would otherwise be limited in distribution to a thin layer a fraction of an inch deep at the mud surface.

The soft substratum is a problem for large animals whose weight may not be adequately supported. It is also unsuitable for sessile animals. These are rare in freshwaters in any case. In salt wetlands, sessile animals are restricted by the substratum, with the exception of such animals as mussels, which can move their position slightly to avoid being covered by mud.

Effect of salt concentration. High salt concentrations make the soil physiologically dry, and salt marshes and swamps are thus difficult habitats for species (halophytes) that grow there. In many cases these plants grow better in well-watered upland soils than they do in salt wetlands, but they cannot compete and are crowded out by upland plants. Upland plants cannot withstand high salt concentrations and leave such areas to the halophytes by default. All halophytes can concentrate salts in their cells. This enables them to resist the wilting action of saltwater.

Halophytes lose water by transpiration (evaporation from the leaves). In the process, salt is left behind. Some halophytes exclude most of the salt at the roots by means of semipermeable membranes. The excess salts that do get into the plant are secreted by special salt glands on the leaves. Other halophytes do not exclude the salts at the roots but secrete them or store them in leaves and stems. At the end of the growing season, the leaves and stems are dropped and with them the salt.

Salt-marsh animals living at or in the ground are largely derived from marine ancestors and have a problem in resisting freshwater from rains rather than salt. Some, such as the worms, merely hide in the mud until the freshwater has run off the marsh surface. Others, such as fiddler crabs, have developed the ability to control their osmotic concentration in freshwater for periods of up to several days. Insects are the principal land animals found on marshes. Although they can withstand short periods of saltwater immersion, they often avoid saltwater by moving up the plants or flying away.

The tops of mangrove trees shelter a diverse group of animals that are unaffected by seawater, because they never contact it. In general, swamp animals are little affected by wetland conditions except as those conditions affect the species of trees growing in the swamp. Whereas the diversity of plants is limited by the stresses that the presence of water entails, the animal diversity is more a reflection of the plant diversity than the nature of the water supply.

Other physical factors. Physical factors of the environment are important in determining the relative distribution of trees and grasslike plants in both fresh and salt swamps and marshes. The growing tips of grasses, sedges, and rushes are at or just under the soil level. In this position they are protected from certain kinds of damage. The tops of the grasses can be bent or even broken off without permanent damage to the whole plant. Fire is an event that may damage grasses but fires are relatively rare on wetlands. Water movement may damage plants by breaking their upper parts. Trees and shrubs will suffer more from such damage than grasses that have flexible upper parts. Grassy plants also have an advantage in regions where large masses of ice form in winter, especially if the ice is moved by currents or tides. Moving ice can completely destroy seedlings and even large trees. Winter ice prevents the evolution of a tree that could form salt swamps. Grasses, however, are not destroyed even though sheered off. The rhizomes, protected by the frozen mud in winter, put up new stalks with the commencement of the growing season.

Water and ice movement, nutrients, and temperature

Wetland plants in bogs suffer from low concentration of plant nutrients (minerals). In bogs that lie on sloping land there may be regular courses of water flowing from areas with exposed mineral soil. These tracks are clearly evident in the growth of trees. The presence of alders along such tracks is especially significant because alders fix nitrogen by means of symbiotic bacteria existing in nodules on their roots. Nitrogen fixed by these plants is then available for other species that cannot fix it themselves. The conditions that favour the tree growth also favour the growth of fungi and bacteria. Peat under these tracks is thinner than elsewhere in the bog. The relatively good growth along these tracks emphasizes the nutrient-poor environment.

Bogs are also colder than the surrounding environments. Open bogs growing only bog mosses are colder than bogs that support higher plants as well. In a central Massachusetts open bog, the frost-free season is only two months long, whereas in the surrounding spruce forest it lasts four months. Most of the plants are limited by lack of nutrients, however, rather than by the shortness of the growing season.

Animals are also rare in bogs. A few insects live in and on *Sphagnum* and on other bog plants, but in general bogs are as impoverished in animals as they are in plants. In some cases animals may affect bogs. Deer taking shel-

ter in white-cedar swamps in northern Wisconsin break up and compact peat, reducing its water-holding capacity. This makes the site more suitable for some varieties of upland trees and accelerates the change from wetland to dry forest. In the process the wintering yards for the deer are destroyed.

Wetland productivity. The productivities of various types of wetlands range from poor to good. Productivity is low in bogs where plant growth is limited by a meagre supply of nutrients. Organic matter accumulates relatively rapidly in bogs because the rate of decay is slow. Productivity has been measured in bogs in Alaska in which the *Sphagnum* grows at the rate of 150 to 280 grams per square metre per year (1,340 to 2,500 pounds per acre). This is about the same productivity as occurs on uncultivated short-grass prairies.

Few measurements exist of the productivity of swamps or freshwater marshes. Values will vary depending on the nutrient supply and should fall between the bog values and those of the most productive natural systems. The latter are well represented by salt marshes. Productivities of more than 3,000 grams per square metre per year have been reported for the most productive parts of salt marshes, the tall *Spartina alterniflora* stands growing along tidal creeks.

Role of the tides in weeding and feeding

These values correspond to nearly 30 tons per acre per year and are equal to the highest values that have been achieved in agriculture. Agricultural productivities have been maximized in subtropical areas where the growing season lasts throughout the year by the use of fertilizers and by means of careful cultivation and weed control. Natural production in salt marshes comes about without human effort. Tides accomplish weeding by keeping all but a few well-adapted plants out of the marshes. Tides also provide fertilizer in the form of nutrients dissolved in the water and nutrients absorbed in particles of silt and clay transported by tidal flow. Although salt marshes occur in areas that have a cold winter, the growing season extends throughout the year. Grasses do not grow during the winter, but unicellular algae, which live on the surface of the marsh and in the water that floods it, grow throughout the year.

BIBLIOGRAPHY. JOHN and MILDRED TEAL, *Life and Death of the Salt Marsh* (1969), a popular treatment of the ecology and geology of salt marshes and the physiology of their inhabitants; C.D. SCULTHORPE, *The Biology of Aquatic Vascular Plants* (1967), a comprehensive review of the adaptations of higher plants to life in wetlands, with a good bibliography; R.F. DAUBENMIRE, *Plant Communities: A Textbook of Plant Synecology* (1968), an ecology text dealing with plant communities, not limited to wetlands; C.H. WHARTON, *The Southern River Swamp: A Multiple-Use Environment* (1970), a readable, conservationist's treatment of an important wetland type in the United States; P.W. RICHARDS, *The Tropical Rain Forest* (1952), a general botanical study of rain forests including tropical swamps and bogs; V.J. CHAPMAN, *Salt Marshes and Salt Deserts of the World* (1960), an exhaustive, technical treatment of salt wetlands of the world; N. POLUNIN, *Introduction to Plant Geography and Some Related Sciences* (1960), a general text of the worldwide distribution of plants.

(J.M.Te.)

Swaziland

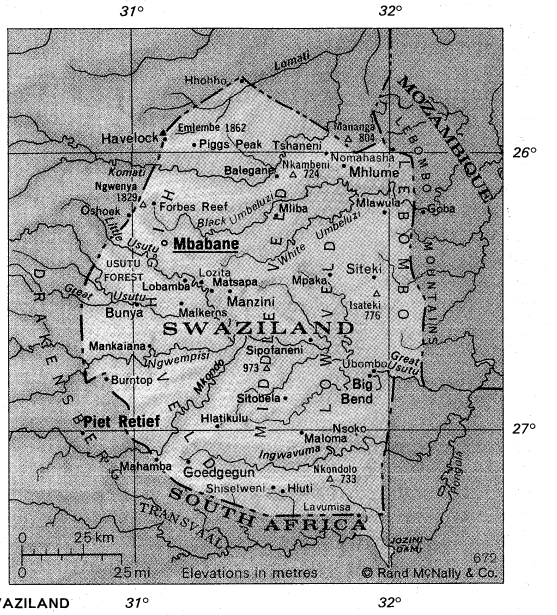
Swaziland is a landlocked, independent kingdom wedged between South Africa to the north, west, and south and Mozambique to the east. It has a total area of 6,704 square miles (17,364 square kilometres). Small and compact, the oval-shaped country is less than 120 miles from north to south and 90 miles from east to west. It is named for the Swazi nation, which comprises almost all of the country's 446,000 people. The national capital is at Mbabane.

Formerly a British High Commission Territory, Swaziland gained its independence in 1968 under the leadership of its king, Sobhuza II. Surrounded by white-ruled countries, the small black nation is faced with delicate problems of economic development. It is rich in agricultural land and mineral resources, but its mines are controlled by foreign interests, and its commercial agricultural operations remain in the hands of Europeans. Economic progress has, however, been steady. For an associated physical feature, see VELD; for historical aspects, see SOUTHERN AFRICA, HISTORY OF.

The natural landscape. Despite its small size, Swaziland has a rich variety of relief features and soils. There are four well-defined topographic regions, extending longitudinally from north to south in roughly parallel belts. From west to east they are Highveld, the Middle Veld, and Lowveld, and the Lubombo escarpment.

Relief features and soils

The Highveld, which covers an area of 2,000 square miles, is a continuation of the Drakensberg range of South Africa. Locally known as Inkangala (meaning A Cold Treeless Place), it is a granite massif (mountainous mass) with quartzite ridges that are broken up into rugged terrain. The average elevation is between 3,500 and 4,500 feet; the highest elevations are the summits of Emlembe (6,100 feet) and Ngwenya (6,000 feet). The rocks on gentle gradients are covered by deep red, orange, and yellow acidic soils of medium texture.



SWAZILAND 31° 32°

The Middle Veld covers an area of 1,900 square miles and has an average altitude of 2,000 to 2,500 feet. It is a region of hilly country and well-watered valleys. Its foundation is mainly dolomite (limestone rich in magnesium carbonate) and gneiss (coarse-grained rocks containing ponds of minerals). Dolerite (coarse basalts) and quartzite also occur. The principal soils are deep and friable red and clay loams; gray-brown sands and sandy loams rest on mottled, sandy clay or iron concretions.

The Lowveld, or Bushveld (locally known as Lihlanze a Swazi name meaning A Warm Place with Trees) covers 2,200 square miles. It is a gently undulating region with hills that rise from an average of 500 to 1,000 feet to a height of 2,300 feet reached by the isolated hill of Nkambeni and the isolated ridge of Nkondolo. The complex geology of the region ranges from the acidic rocks of the west to the basalt and dolerite of the east. The soils are similar to those of the Middle Veld.

The Lubombo, covering 600 square miles, is an escarpment along the eastern fringe of the Lowveld. It rises from an average height of 2,000 feet to the two highest points of Isateki hill (2,500 feet) and Mananga Beacon (2,700 feet), which is on the border with South Africa. The scarp is intersected by the gorges of the Ingwavuma, Usutu, and Black Umbeluzi rivers. The plateau's foundations are composed of acidic and intermediate volcanic lavas, such as rhyolite, lava-formed granite. Soil is virtually absent in some areas, while in others it is fairly deep, reddish, and medium to heavy in texture.

The rivers

The country is one of the best watered in southern Africa; major rivers flow through all four regions. From the Highveld—where most streams are perennial—the Lom-

ati, Komati, Black Umbeluzi, Usushwane (Little Usutu), Usutu, Ngwempisi, and Mkhondvo rivers flow in a generally easterly direction toward the Indian Ocean. The Usutu, which carries the greatest volume of water, flows into the Pongolo to form the Maputo River of Mozambique.

The environment. The climate varies from almost tropical and subhumid conditions in the Lowveld, where the average annual rainfall amounts to between 20 and 35 inches, to the humid, near temperate climate in the Highveld, which receives between 40 and 90 inches of rainfall. The Middle Veld and Lubombo region are subtropical and drier, receiving only 30 to 50 inches of rain. Rains occur during the summer from October to March. The mean annual temperature is 59° F(15° C) in the Highveld and 72° F (22° C) in the Lowveld.

More than 2,600 species of ferns and flowering plants have been recorded. Among these are 25 species of aloe (members of the lily family), *arum lilies*, Cape primroses, begonias, gladioli, and a rich variety of orchids. Cape heath (*Erica*), a large sturdy shrub, with red, pink, green, yellow, and white flowers, grows in the mountains. Trees include the crimson-flowered Hottentot's bean, the scarlet-flowered umsini, umvangata or kiaat (a leguminous tree that yields durable wood), and umbrella trees.

Although the animal population has been greatly reduced (earlier by hunting, but now by the expansion of agriculture), there are still numbers of hippopotamus, blue wildebeest, zebra, and antelope (including the koodoo, impala, reedbuck, duiker, klipspringer, and waterbuck). Lions, baboons, monkeys, and galagos (small African lemurs known as bush babies) are less numerous. There are crocodiles in the Lowveld rivers. Birdlife is abundant and includes a few rare species with northern affinities, such as the European stork, the sacred ibis and hadedah. Other more conspicuous species are the hammerhead, the gray heron, the lisakabuli (widow bird), the hornbill, and the lilac-breasted roller.

The landscape under human settlement. Nine-tenths of all Swazis live in villages that are divided into family homesteads, or *umuti*. Each village is led by a headman, who is the village patriarch. A unique feature of Swazi life is the role of the headman's mother, who acts as village "mother" and is in control of everything affecting the homestead. Each such homestead, at the minimum, consists of a great hut, the cattle kraal (enclosure), and the bachelors' quarters. Flask-shaped pits for storing grain are built into the wooded palisades surrounding the kraal.

Rural life

The pattern of village life is strictly set by the seasons. With the onset of the rains in spring (August to October), women plant small gardens along the riverbanks; later, when the heavy rains come in summer (November to January), with some help from the men, they hoe and sow maize (corn) and millet in larger fields; at this time all able women and children abandon their homesteads for the fields, and the men also join in the planting and weeding. The summer months are, on the whole, the hungry months.

Autumn (February to March) is the time of harvest; in winter (April to July) the last of the crops are brought in. Activity then moves to the homesteads, where women and men thresh the grain, the best of which is stored and the remainder consumed at once. Winter is a time for relaxation; the men hunt and entertain from their full food bins and the women visit their parents.

The central points of Swazi political life are the royal villages of the king at Lozita, with a population of 50, and the queen mother at Lobamba, which had 75 to 100 inhabitants in the early 1970s and was to become the new legislative capital. The two principal commercial centres are Mbabane and Manzini. Mbabane, attractively situated on wooded hills, contains both modern and colonial-style buildings. It is the country's largest town, with 17,850 inhabitants, of whom 2,000 are Europeans. Manzini is a growing industrial and agricultural service centre; its population of 7,580 includes 500 Europeans.

The Swazi nation

People and population. The Swazi nation of 382,000 people is composed of over 70 clans, of which 70 percent are Nguni and 30 percent Sotho. The Dlamini, the dominant clan, forms the nucleus of the nation and constitutes the royal family. Its customs, laws of property and person, and dialect, known as Siswati, prevail.

The Swazis comprise about 90 percent of the total population of about 446,000. Other Africans, mainly Shangane from Mozambique, number about 25,000. Europeans, largely English- and Afrikaans-speaking South Africans, total less than 10,000, and Eurafricans (persons of mixed European and African descent) and Asians number about 5,000. Eurafricans and Asians are mainly in commerce and industry as shopkeepers, clerks, and artisans. Most Europeans are either large-scale farmers, miners, technicians, teachers, public officials, or traders. The majority of the Europeans speak English, which is one of the official languages, together with Siswati.

A majority of Swazis are Christians. About 43 percent hold strongly to their traditional animist religious beliefs.

The population, which totalled about 446,000 in 1972, was expected to double by 1985. The birthrate is 48 per 1,000 and the mortality rate 22 per 1,000. Life expectancy is 40 to 45 years of age. Population varies from 37.9 in parts of the Lowveld to 572 persons per square mile in the Ezulwini Valley of the Middle Veld.

Swaziland, Area and Population

	area		population	
	sq mi	sq km	1966 census	1972 estimate*
Districts				
Hhohho	1,378	3,569	96,000	110,000
Lubombo	2,296	5,947	82,000	117,000
Manzini	1,571	4,069	101,000	112,000
Shiselweni	1,459	3,780	96,000	95,000
Total Swaziland	6,704	17,364†	375,000‡	446,000§

*Resident African population only. †Figures do not add to total because of rounding. ‡Includes 126 persons in transit through the country at the time of the census. §Includes an estimated 12,000 non-Africans.
Source: Official government figures.

The national economy. Swaziland is principally an agricultural country, with 90 percent of the population engaged in subsistence agriculture. Mixed farming is carried out on the Middle Veld, where rice and fruits, such as bananas and citrus, are grown under irrigation; maize, cotton, beans, tobacco, and pineapples are raised on dry lands. The Lowveld is ideally suited to cattle raising, which is little developed commercially; this region is under irrigation, and there is intensive cultivation of cash crops such as sugarcane, rice, tobacco, and citrus fruits, which are mainly grown by European farmers for export.

The nation's main export wealth derives from forestry and mining. The largest man-made forest, the Usutu afforestation estate, covers 100,000 acres of pine and eucalyptus trees planted in the Highveld. It supplies the wood pulp factory, near the town of Bunya, the products of which are a major export earner. Although not extensive, mineral resources are rich. Iron ore and asbestos deposits are located in the Highveld region, and there are also important coal deposits near Mpaka.

Mining operations

Three major mines were in operation in the early 1970s. The most important is the iron-ore mine at Ngwenya, owned by the Swaziland Iron Ore Development Corporation and controlled by South African and British interests. The British-owned Havelock Asbestos Mine exports asbestos fibre. Third in importance are the Mpaka coal mines, mainly owned by a South African firm. Smaller mines, also foreign owned, produce kaolin, barite, and pyrophyllite.

Industry is mainly confined to the processing of agricultural products and to providing small-scale services. Manufactured products include clothing, edible oils, soap, canned fruit, and ginned cotton. There are also a slaughterhouse and a small engineering industry. Handmade woven and wooden articles are also produced.

Less than one-third of the working population is engaged in paid employment. Of these, more than 25 percent are employed in agriculture and related services. It

was formerly traditional for Swazi labourers to engage in migratory labour in South Africa's mines and farms, but by the early 1970s the proportion that still did so had dwindled to about one-quarter of all male workers. Annual recruitment for the South African mines was about 6,000.

The bulk of all wholesale and retail trade is traditionally controlled by white South Africans. All commercial banking is foreign owned.

The country maintains a favourable balance of trade. The need for budgetary assistance ended in 1970, but the country still depends on foreign aid—mostly from the United Kingdom—for capital and development needs. Imports of machinery, fertilizer, clothing, fuel, and vehicles largely come from South Africa. Exports—including iron ore, sugar, wood pulp, asbestos, and agricultural products—are sold to the United Kingdom, Japan, and South Africa.

The main sources of national income are taxes, mining royalties, and customs revenues. Income tax is payable on all incomes derived inside the country. A personal tax is paid by all non-Swazi adult males, and Swazi adult males pay a graduated tax based on the number of their wives. Company taxes are paid at a fixed rate, and mining taxes are on a variable scale.

The country's future economic prospects are mostly encouraging. With improved education and training, the Swazis can be expected to increase their levels of agricultural productivity and to play a much larger role in private industry and commerce.

Transport and communications. Good all-weather roads link the main population centres. The transterritorial highway runs from Oshoek, on the western frontier, to Nomahash, on the eastern border. The road network consists of more than 900 miles of main roads and 1,000 miles of feeder roads.

A railway line running from Kadake to the port of Lourenço Marques, in Mozambique, opened in 1964, is linked to the Mozambique Railways network at Goba across Swaziland's northeast border. There is no rail link with South Africa.

The only airport is at Matsapa, five miles from Manzini. South African Airways operates scheduled services to Johannesburg and Durban, and Mozambique Airways flies to Lourenço Marques. There are more than 20 landing strips.

The government operates postal, telephone, and telegraph services. An international teleprinter service is routed through South Africa.

Administration and social conditions. Executive authority is vested in the king and exercised through a cabinet, which is presided over by the prime minister, who is appointed by the king. The cabinet advises the king and is responsible to a bicameral parliament. The House of Assembly is composed of 30 members—24 of whom are elected by adult franchise and 6 appointed by the king and the attorney general. There is universal franchise for all men and women over the age of 21 who are citizens of Swaziland and who are not insane and who have not been convicted of serious crimes. The Senate has six members elected by the House of Assembly and six nominated by the king. Parliament has no power to legislate in respect of Swazi law and custom, which is the prerogative of the Swazi National Council, consisting of the king, the queen mother, and all adult male Swazis. The civil service is controlled by a public service commission appointed by the king on the advice of the Judicial Service Commission.

The governmental structure

Local urban government operates through town councils and urban area advisory committees. A majority of the council and committee members are elected and a minority appointed by the Minister of Local Administration. Rural administration is conducted in the four districts of Shiselweni, Lubombo, Manzini, and Hhohho by district commissioners responsible to the Minister of Local Administration.

The judicial system comprises a court of appeal, a high court, subordinate courts, and Swazi (traditional) courts. The court of appeal comprises a judge president and three justices of appeal appointed by the king. The high court, presided over by the chief justice, hears appeals in civil and criminal matters from subordinate courts and the Higher Swazi Court of Appeal. Subordinate courts, presided over by magistrates, have wide criminal jurisdiction but not over murder, sedition, or treason. There are 14 Swazi courts, two courts of appeal, and the Higher Swazi Appeal Court; these may hear cases only if all those involved are Africans and if the charges fall within a restricted list of criminal and civil offenses.

Land ownership is the most sensitive issue in Swazi national life. Under the British colonial administration about 44 percent of the land was either sold as concessions to Europeans or alienated by the government. Traditionally, however, all land is vested in trust in the king as Swazi national land, and all Swazis are entitled to a share, which is allocated by the king through the chiefs. The Swazi National Council adopted a policy in 1946 of repurchasing for the nation as much as possible of the 1,873,000 acres still held out of communal ownership by Europeans.

Half of the school-age children, numbering about 71,000, attend the 366 primary schools, of which 322 are mission (primarily Protestant) schools. About 9,000 students attend the 54 secondary schools. Schooling is free except at the Waterford School at Mbabane, which is a multiracial institution for pupils from all over southern Africa. Two teacher-training colleges produce 170 teachers annually. Higher education is provided at the University of Botswana, Lesotho, and Swaziland (UBLS), which is situated at Roma, in Lesotho.

Educational services

Medical services are provided by the government, missions, industrial concerns, and private practitioners. In the early 1970s, the country had 11 hospitals and 44 rural clinics. Chief causes of illness are schistosomiasis (a parasitic infestation of the bladder or intestines), food-deficiency diseases, typhoid and paratyphoid, and respiratory diseases. Malaria has been virtually eliminated.

Cultural life. Traditional ceremonies and music play an important part in Swazi social life. The two major cultural events are the *Incwala* (National Ceremony) held in December or January, and the reed dance (*umhlanga*) held in June or July; both are held at the royal village of Lobamba. The *Incwala*, the most sacred of the ceremonies, represents the king as the source of fertility and the symbol of power and unity. It lasts for six days and is a highly ritualized festival of song, dance, folklore, and martial display. The reed dance, which lasts for one week, encourages girls of the same age-group to work in harmony under discipline and to preserve their virginity.

Major national celebrations

Traditional music accompanies the Swazi from birth to death, beginning with the first lullaby. There are songs for children and adults; for courting, weddings, and mourning; for working; for specific tribes and clans; and for insult. The main musical instruments are rattles, which are tied to the wrists or ankles or shaken by hand. The shield is used for percussion; buckhorn whistles and long red flutes are the principal wind instruments.

The Swaziland Broadcasting service in Mbabane, which is commercial in character, broadcasts in English and Siswati. Radio programs are also received from South Africa and Mozambique. The *Times of Swaziland* is a privately owned English-language weekly. *Umbiki* is a government fortnightly published in Siswati. Both are published in Mbabane.

Prospects for the future. Swaziland is one of Africa's more politically stable countries. With the promise of economic development, based on adequate resources, its future appears assured. Long-term problems arise from the country's sensitive location between Mozambique and South Africa. National policy is to protect the country's independence by not becoming too closely dependent on the political systems of either of its neighbours.

BIBLIOGRAPHY. For physical geography and description, see HARM J. DE BLIJ, *A Geography of Subsaharan Africa* (1964); and ALAN B. MOUNTJOY and CLIFFORD EMBLETON, *Africa: A New Geographical Survey* (1967). Descriptive works include DUDLEY BARKER, *Swaziland* (HMSO 1965); and

DOUGLAS POTT, *Swaziland: A General Survey* (1965). HILDA KUPER, *An African Aristocracy: Rank Among the Swazi* (1947), is the accepted work on the social life and institutions of the Swazi; JACK HALPERN, *South Africa's Hostages: Basutoland, Bechuanaland and Swaziland* (1965), is a useful survey written by a political observer. For administration and political and constitutional affairs, see LORD HAILEY, *The Republic of South Africa and the High Commission Territories* (1963); COLONIAL OFFICE, SWAZILAND, *Annual Reports* (to 1966); RICHARD P. STEVENS, "Swaziland Political Development," *Journal of Modern African Studies*, 1:327–350 (1963); *Lesotho, Botswana and Swaziland* (1967); DENIS V. COWEN, *Swaziland: Report on Constitutional Reform* (1961); and COLIN LEGUM and JOHN DRYSDALE, *Africa Contemporary Record*, 3 vol. (1968–71). J.F. HOLLEMAN (ed.), *Experiment in Swaziland: Report of the Swaziland Sample Survey, 1960* (1964), is a detailed scientific study of the country's physical resources.

(Co.L.)

Sweden

In area the fourth largest nation of Europe—after the U.S.S.R., France, and Spain—Sweden lies in the Scandinavian Peninsula bounded by Norway on the west, Finland on the northeast, and the Gulf of Bothnia, the Baltic Sea, and the North Sea on the east, southeast, and southwest, respectively. A narrow sea channel separates it from Denmark to the southwest. Most of the 8,000,000 inhabitants live in dense clusters in a relatively small portion of its 173,649 square miles (449,750 square kilometres), but in the years since 1815, during which Sweden has remained at peace, the Swedish people have achieved a remarkably stable and democratic way of life. In form, Sweden's government is a constitutional monarchy, with its seat in Stockholm.

Location and general character of the nation

The popular image of Sweden includes such features as the prototype it has provided of the modern welfare state, a "middle way" that is Socialist but non-Communist, and its traditional nonalignment in peace and neutrality in war, which has allowed its people to assume a unique leadership in the world community.

Although more than 90 percent of Sweden's trade and industry is in private hands, its labour movement chose, at the time of industrialization in Sweden (at the end of the 19th and the beginning of the 20th century), to follow national traditions that combined firm and efficient public administration with a keen sense of justice, independence, and loyalty. These qualities perhaps in some way reflected the high degree of homogeneity that the Swedish people had achieved in ethnic, linguistic, and religious identity. Thus, the movement found its place within—or, in the view of many, conquered—the newly evolving social and economic forms by the relatively slow processes of democratic, parliamentary action rather than by upheaval and force. Today, the nation's standard of living is perhaps exceeded only by that of the United States.

As a country that in many ways has had good luck, including a peripheral but in-between location and vast resources of iron ore and timber, Sweden long ago disavowed the military aggressiveness that once made its armies bulk large during Europe's centuries of dynastic warfare. It chose instead to play a balancing role among the world's conflicting ideological and political systems. It is no coincidence that Swedish statesmen have often been sought out to fill major positions of leadership and mediation within the United Nations.

For information on Swedish history, see the article SCANDINAVIA, HISTORY OF. For the nation's place in the larger physical and cultural pattern, see EUROPE and articles on the various arts, such as LITERATURE, WESTERN; MOTION PICTURES, HISTORY OF; and VISUAL ARTS, WESTERN; see also the article STOCKHOLM.

THE NATURAL AND HUMAN LANDSCAPE

The natural environment. *Physical features.* Sweden's four main physical regions are, from the north, the northern mountain and lake region known as Norrland; the lowlands of central Sweden known as Svealand; the relatively low Småland highlands and the small but rich plains of Skåne, the two last areas in Götaland.

From the ancient mountain range with its crest along the Swedish-Norwegian border, the land descends southeast and south. The mountain, or *fjäll*, region has a most dramatic scenery. The highest peaks of Sweden are found in Lappland: Kebnekaise, at 6,926 feet (2,111 metres), and Sarektjåkko, at 6,854 feet (2,089 metres), the latter situated in a national park of great and lonely magnificence. Åreskutan, Helagsfjället and Sylarna peaks lie amid favourite sporting locales in a more southerly location; they are especially popular during the spring and summer.

The four physical regions

The big rivers, or *älvar*, of Norrland flow from the mountains to the sea in the southeast. Many are linked to long, narrow lakes made by dams of moraine, and often they have rapids and falls, many of which have power stations. Several of the larger rivers have huge flows and fall great distances in their course to the sea.

The metallic-ore deposits are found mostly in the lower parts of the highlands, both in the north and farther south. These highlands were not below sea level during the glacial period, and the soil is meagre. Most of the central lowlands were then pressed down below sea level by the glaciers, creating fertile plains of clay, with many lakes. The gravel ridges that are characteristic features of the landscape of the whole region are remains of the last glacial period. In ancient times they were used as highways, but today they serve as gravel and water deposits. Many lakes, big and small, are typical of the central lowlands. Lake Vänern, 2,156 square miles in area, is the largest of Sweden's 90,000 lakes and the largest in Scandinavia. The Göta, which drains Lake Vänern into the Kattegatt section of the North Sea at Göteborg, is the only major river of southern Sweden. It is famous for its falls in Trollhättan, located in a region of fertile plains, forests, industries, and numerous small towns.

South of Lake Vättern the plains are transformed into a new but lower mountain area, the Småland highlands, seldom over 500 feet above sea level. The soil is stony; the agriculture is small in scale; and the people depend upon small industries that utilize forest products.

The plains of Skåne are the most densely populated agricultural areas of Sweden. The ground consists of rocks from later geological periods, and the soil has been enriched by recent marine deposits. Some clay beds in the northwest contain coal, and the clay itself is used for making fireproof materials.

Typical of Sweden is a rocky coastline with hundreds of small, sometimes wooded islands. Ground by the glacial ice in the same direction, they have a common rounded shape. This type of coast, known as *skärgård*, is found both in the east and west, especially around Stockholm and Göteborg. Other parts of the coast are sandy, and the island of Gotland has limestone rocks.

Climate. The Scandinavian countries profit from the warm Gulf Stream, which allows forests and agriculture farther north than anywhere else on the globe. About 15 percent of the country lies within the Arctic Circle, and the climatic differences between northern and southern Sweden are vast. The temperate west winds of the stream produce a continental climate similar to that of the Soviet Union, which is very cold in winter and very warm in summer. Mean February temperatures range from 7° F (−14° C) at Karesuando in the far north to 27° F (−3° C) in Stockholm and 30° F (−1° C) in Lund in the far south; comparable July figures are 57° F (14° C), 64° F (18° C), and 63° F (17° C).

Interaction of geography and climate

The regions covered by the high mountain range have a more continental climate than the west coast, where the prevailing winds dump heavy rains that diminish as the winds approach the east coast. Also there, precipitation is sufficient for agriculture. Annual precipitation ranges from 15 inches in the far north to 22 inches in Stockholm and 23 inches in Lund.

Snow lies in the mountains of Norrland for eight months, whereas three months is the average length of winter in the highlands of Småland, and one month or less in Skåne. Ice seldom covers the waters of the west coast, but the northern Baltic is usually unnavigable from the end of November to the beginning of May.

The shorter growing period in the north, about three or four months as compared to eight in Skåne, is partly compensated for by the long period of constant daylight. In Karesuando, above the Arctic Circle, the sun remains over the horizon between May 26 and July 18. At the summer solstice on June 21, daylight lasts 23 hours at Haparanda, at the northern end of the Gulf of Bothnia, and in Skåne 17 hours. Long twilight hours are typical of Sweden.

Vegetation and animal life. Latitude and altitude produce many local variations. In descending order of altitude and latitude, general vegetational zones are Alpine regions, birchwoods, a coniferous forest region, and a beech wood region.

The alpine region, down to 1,600 feet in the north and 3,000 feet in the southwest, begins with mosses and lichens at the highest points where growth occurs. Dwarf birches become a more extensive belt of alpine birches in the north, together with rowans, bird cherries, willows, and aspens. The undergrowth is rich in herbs. Lichens, especially reindeer moss, prevail in dry areas.

The coniferous forests that cover 57 percent of Sweden are especially abundant in the north and the Småland highlands. Some farms currently are being converted back to forest. Scotch pines and spruces are often mixed with lowland birches, rowans, and aspens in the north and central parts of Sweden and with oaks, ashes, lindens, maples, and elms in the south. The beech forests are found in Skåne.

The wolf, almost extinct, is protected by law, as are bear and lynx, which are only found in the deep forests of the north. Large domesticated herds of reindeer are common in the northern mountains, while the elk, the most valuable game animal, is common throughout the country. Badgers, foxes, hare, and otters are also hunted, as are the roe deer, which are common in southern and central Sweden.

Birds are numerous, with some migratory species breeding in spring and summer. Songbirds, game birds, and forest birds of prey are found, as are cranes in northern marshy clearings; gulls, terns, and eider ducks are found at the coast. The viper is the only poisonous snake. Salmon, trout, char, pike, and perch are common in rivers and lakes. Both freshwater and marine fisheries are important, especially on the west coast, where shrimp, herring, cod, flatfish, mackerel, and sprat are taken. The crayfish, a favourite among Swedes, is found in many lakes of southern Sweden.

The human imprint. *Traditional regions.* Götaland, Svealand, and Norrland are considered the traditional regions of Sweden. The first two take their names from the small, prehistoric clans of Götaland and Svealand who inhabited central Sweden. The Svear and Götar were united into one state about AD 1000. The Götar lived in Västergötland and Östergötland and Småland, the Svear around Lake Mälaren. Certain differences remain in the dialects of these two regions. Skåne and the surrounding regions were taken from the Danish crown in the 17th century, and Skåne is still looked upon as a special region in both language and customs. The food is richer, and the people are said to be warmer and more easygoing than the taciturn northern people.

Variations in cultural character

The vast Norrland was colonized later. It is looked upon as somewhat less important than the old cultural centre of the southern and central regions, and its vast distances are often overlooked. These tendencies have become stronger, with the modern industrial concentration in the southern and central parts.

Rural settlement. Generally, the traditional patterns of villages in the old agricultural areas were broken up in the 19th century to get larger field units and more productive cultivation. In Värmland and Dalecarlia, the traditional red-painted wooden houses still lie side by side in the village centre. In Skåne, a square house with brick and clay applied over a stick frame is the most common form.

Houses built in modern Sweden are basically the same throughout the country. Many small rural holdings have been annexed to neighbouring areas and the empty houses converted to vacation homes. To have homes in both urban and rural surroundings is the dream of many Swedish families, especially those who have had to abandon their rural milieu.

The relatively few castles and manors are found only in southern and central Sweden. The old industrial centres were often concentrated near waterpower and water transportation facilities; the newer ones are situated near railway stations and along the highways.

Urban settlement. The process of urbanization has been rapid during recent decades. About 90 percent of the population lived in the countryside 100 years ago, whereas today less than 20 percent do so—and only about seven percent of the economically active population are farmers.

This shift in concentration has occurred not only in the central cities but also in the areas surrounding them. This is especially true of regions around Stockholm, Göteborg, Malmö, and the cities and towns in the central lowlands. The population has diminished in the interior regions of Norrland, the northwestern portion of Svealand, southeast Götaland, and the isle of Gotland. Often the population shift has been a two-step process, shifting first from the rural areas to small towns and from there to the cities.

Sweden, Area and Population

	area*		population	
	sq mi	sq km	1965 census	1970 census
Counties (*län*)				
Älvsborg	4,435	11,488	390,000	404,000
Blekinge	1,123	2,909	149,000	154,000
Gävleborg	7,024	18,191	293,000	293,000
Göteborg och Bohus	1,973	5,110	666,000	715,000
Gotland	1,212	3,140	54,000	54,000
Halland	1,879	4,867	180,000	193,000
Jämtland	18,343	47,508	131,000	125,000
Jönköping	4,063	10,523	296,000	307,000
Kalmar	4,313	11,171	236,000	241,000
Kopparberg	10,946	28,350	282,000	277,000
Kristianstad	2,335	6,048	262,000	264,000
Kronoberg	3,266	8,460	164,000	167,000
Malmöhus	1,895	4,909	667,000	719,000
Norrbotten	38,188	98,906	260,000	255,000
Örebro	3,340	8,650	268,000	277,000
Östergötland	4,079	10,566	366,000	382,000
Skaraborg	3,029	7,844	255,000	257,000
Södermanland	2,340	6,060	241,000	248,000
Stockholm†	2,511	6,503	1,382,000	1,477,000
Uppsala	2,694	6,977	185,000	217,000
Värmland	6,789	17,584	287,000	285,000
Västerbotten	21,401	55,429	234,000	233,000
Västernorrland	9,313	24,120	277,000	274,000
Västmanland	2,381	6,166	250,000	260,000
Total Sweden	158,873‡	411,479	7,767,000‖	8,077,000‖
	173,649§	449,750		

*Land areas except the second country total. †On Jan. 1, 1968, Stockholm Stad (city) was incorporated into Stockholm Län.
‡Converted area figures do not add to total given because of rounding.
§Includes 38,271 sq km (14,777 sq mi) of inland water not accounted for in county area figures. ‖Figures do not add to total given because of rounding.
Source: Official government figures.

THE PEOPLE OF SWEDEN

Nearly 90 percent of Sweden's population, small in relation to its land area, live in the southern half of the country and about 80 percent in towns or other densely populated areas.

Linguistic, ethnic, and religious groups. Although different groups of immigrants have influenced Swedish culture through the centuries, the population is unusually homogeneous in language, ethnic stock, and religion. The language, which belongs to the North Germanic (Scandinavian) subgroup of the Germanic languages, is closely related to the Danish, Norwegian, Icelandic, and Faeroese languages. It has been influenced at times by German, but it has also borrowed some words and syntax from French, English, and Finnish. The written language showed the influence of Latin as late as in the early 20th century, especially in formal usage. A common standard language has been in use more than 100 years, though

Homogeneity of the populace

dialectal intonation is often retained. The traditionally varying dialects of the provinces are being homogenized rapidly through the influences of education and the mass media.

The Lapps form a special linguistic group, though they usually understand and speak Swedish as well and are taught both Swedish and the Lappish language at school. There are approximately 10,000 Lapps living in Sweden. From time immemorial, they have inhabited the northern parts of the Nordic countries, retaining their own culture and language, the latter belonging to the Uralic group of languages. Immigration to Sweden after World War II created new language groups, such as Finnish, Serbo-Croatian, Turkish, and Greek. Efforts are being made to make these immigrants and their children learn Swedish, while keeping their mother tongue.

The new immigrants have little diminished the ethnic and religious homogeneity. About 90 percent of the Swedish people belong to the established Swedish state church, Evangelical Lutheran. Even most of those who profess to different "free churches" in Sweden still belong to the state church. The Swedish state church is organizationally and economically tied to the state, but it is independent in religious matters. Infants are "born into" the state church unless the parents have withdrawn, and later must withdraw if membership is not wanted. The 10 or so free churches are all Evangelical Lutheran; the two largest each have about 90,000 members. The Roman Catholic Church has almost 35,000 members, but only 8,000 of these are Swedes. The rest are foreigners, some living only temporarily in Sweden. The Jewish faith is practiced by about 16,000 adherents.

Demography. Sweden's birthrates and death rates are among the lowest in the world, and it has had one of the lowest infant mortality rates for many years. The 14.1 births per 1,000 inhabitants in the early 1970s were just over one-half those of the period from 1871 to 1910, while deaths fell from 16.6 per 1,000 to 10.2. As a consequence of these declining rates, the percentages of the population in the age groups below 15 and above 65 have, respectively, dropped and risen markedly during the 20th century; the proportion in the "working ages" 15–64 has remained relatively stable.

Historic patterns of migration

Emigration during the latter part of the 19th century—more than 1,000,000 Swedes emigrated between 1860 and 1910, most of them to the New World—has left its special mark on the Swedish people in respect to population movements across the frontiers. Since 1910, Sweden has had an increasing surplus of immigration over emigration. This indicated a return to an older tradition, when different groups of people moved to Sweden: Hanseatic merchants during the Middle Ages; German, English, French, and Dutch officers and government officials during the 16th and 17th centuries; and German, English, and Dutch merchants and industrialists during the 18th and 19th centuries. After World War II, immigration consisted mainly of refugees from the Baltic states—about 30,000—and later of workers from Italy and other countries in which wide unemployment existed. During the 1960s more unskilled labour was allowed to enter Sweden.

Sweden's high rate of emigration, approximately one-half the rate of immigration, continues, but the bulk of this is made up of immigrants returning to their home countries or moving on to other states. A very small percentage of native Swedes emigrate.

The large internal population movements that have taken place since the beginning of industrialism have accelerated during recent decades. As elsewhere in the world, these movements have increased concentration of people in the larger towns as well as in Sweden's three major metropolitan areas—Stockholm, Göteborg, and Malmö. Between 1910 and 1970, the population of Stockholm grew from 342,000 to 747,000; of Göteborg, from 168,000 to 451,000; and of Malmö, from 88,000 to 265,000.

Projections for the future include the fact that the growing number of educated people who will begin gainful employment at a later age probably will alter the percentage of youth and workers. The labour reserve may be maintained partly by continued immigration and partly by increased participation of married women. The percentage of married women who are gainfully employed already has increased from 25 in 1950 to 50 in 1970.

THE NATION'S ECONOMY

Exports make up a bit over 20 percent of Sweden's gross national product, which is the highest per capita in Europe. The greatest proportion of trading, both export and import, is with member nations of the European Free Trade Association (EFTA) and the European Economic Community (EEC). The formation of multinational enterprises has been a traditional part of Sweden's international economic relations. In the early 1970s about 230,000 persons were employed in Swedish companies abroad, while about 70,000 persons were employed in foreign enterprises in Sweden.

Minerals and forest products

Basic economic resources. Sweden's important raw materials are iron ore and lumber. The iron ore now being mined is concentrated largely in the Bergslagen district and in Kiruna-Gällivare-Pajala, an area rich in ore. Kiruna boasts the largest ore deposit in the country and one of the richest in the world. Until about 1960, mining in Kiruna was done by open-pit workings, but today all mining is done underground.

Of great importance also are the sulfide ores, which often contain important metals. The Skellefteå field completely dominates the production of such metals as gold, silver, copper, and iron pyrite (the last important for the pulp industry). Smaller deposits of zinc, lead, manganese, tungsten, and copper exist elsewhere. Exploitation of coal deposits in southern Sweden has ceased because of the low quality of the coal.

Productive forest area covers more than one-half of Norrland, Svealand, and Götaland. Harvesting and reforestation are regulated by law. Corporations and other private owners control at least three-quarters of the nation's forest land and timber value.

The forest industry today is concerned not only with the supply of such traditional products as lumber, pulp, paper, and board but also with such products as rayon, plastics, dyes, resins, and turpentine. The latter require more complex technology and consequently pump more "value added by manufacture" into the economy.

Its many rivers with large waterfalls make Sweden one of the greatest exploiters of hydroelectric power. The total development of this potentiality is expected to be completed by the mid-1970s. Some waterfalls, however, will be kept undeveloped to preserve the natural environment. A series of thermal power plants complement the waterpower, producing about 20 percent of all energy. They are run almost exclusively by imported oil. A number of nuclear power stations are being built. The state is in charge of approximately 45 percent of the total power production.

Sources of income. In the early 1970s the production of power and goods contributed just under, with services contributing just over, 50 percent of Sweden's GNP, representing a slight rise in the role of services over the two preceding decades. Manufacturing remained the major factor during this period, but the agricultural and fishing shares declined precipitously, while public services, trade, and transportation and communications topped the list of services.

To a great extent, Sweden is dependent on foreign trade for its economic growth and development. About one-half its industrial production is sold abroad. The old domination of iron and forest products has decreased gradually in favour of different kinds of manufactured commodities, especially machinery and transportation equipment. Mining products and fabricated metals make up about 50 percent of the Swedish export, while forestry and forest-based industries contribute another 25 percent.

The private and public sectors. In addition to the 90 percent of Swedish industry that is privately owned, cooperative enterprises account for about one-half of the rest; the remainder is publicly owned, either by the state or by local authorities. Cooperative activity is relatively greater in the retail area. By combining the more impor-

MAP INDEX

Political subdivisions

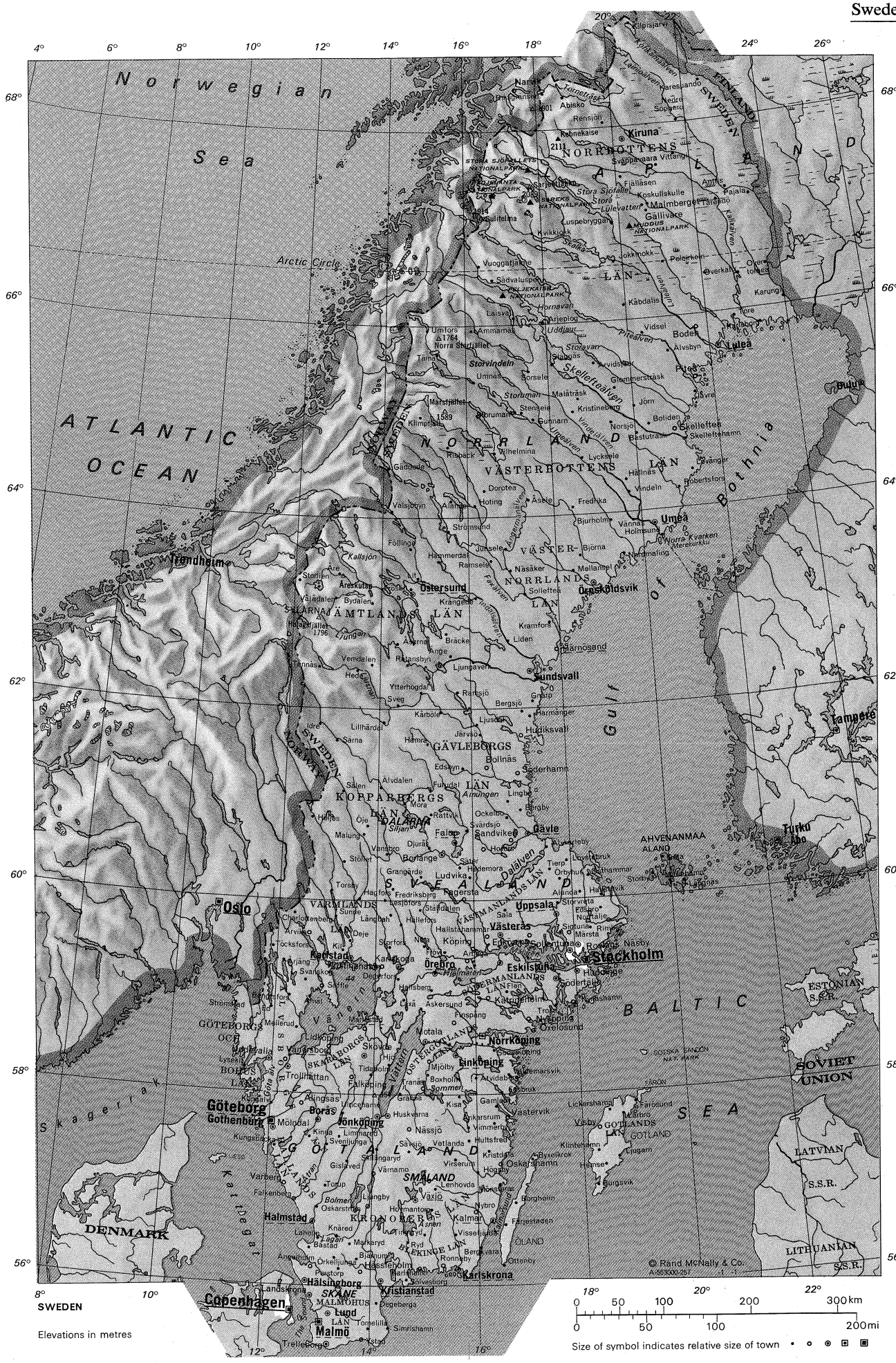
Älvsborgs Län...58·00n 12·30e
Blekinge Län.....56·20n 15·20e
Gävleborgs Län..61·30n 16·15e
Göteborgs Och Bohus Län.....58·30n 11·30e
Gotlands Län....57·30n 18·30e
Hallands Län....56·45n 13·00e
Jämtlands Län...63·00n 14·40e
Jönköpings Län..57·30n 14·30e
Kalmar Län.....57·20n 16·00e
Kopparbergs Län.....61·00n 14·30e
Kristianstads Län.....56·15n 14·00e
Kronobergs Län..56·40n 14·40e
Malmöhus Län...55·45n 13·30e
Norrbottens Län..65·30n 21·30e
Örebro Län.....59·30n 15·00e
Östergötlands Län.....58·25n 15·45e
Skaraborgs Län..58·20n 13·30e
Södermanlands Län.....59·15n 16·40e
Stockholms Län..59·30n 18·20e
Uppsala Län.....60·00n 17·45e
Värmlands Län...59·45n 13·15e
Västerbottens Län.....64·00n 17·30e
Västernorrlands Län.....63·00n 17·30e
Västmanlands Län.....59·45n 16·20e

Cities and towns

Abisko.....68·20n 18·51e
Alanäs.....64·10n 15·42e
Alingsås.....57·56n 12·31e
Alunda.....60·04n 18·05e
Älvdalen.....61·14n 14·02e
Älvkarleby.....60·34n 17·27e
Älvsbyn.....65·39n 20·59e
Åmål.....59·03n 12·42e
Ammarnäs.....65·56n 16·09e
Ängelholm.....56·15n 12·51e
Ankarsrum.....57·42n 16·19e
Anttis.....67·16n 22·52e
Arboga.....59·24n 15·50e
Åre.....63·24n 13·04e
Årjäng.....59·23n 12·08e
Arjeplog.....66·00n 17·58e
Arvidsjaur.....65·35n 19·07e
Arvika.....59·39n 12·36e
Åsarna.....62·39n 14·21e
Åsele.....64·10n 17·20e
Askersund.....58·53n 14·54e
Åtvidaberg.....58·12n 16·00e
Bästad.....56·26n 12·51e
Bastuträsk.....64·47n 20·02e
Bengtsfors.....59·02n 12·13e
Bergby.....60·56n 17·02e
Bergkvara.....56·23n 16·05e
Bergsjö.....61·59n 17·04e
Bjärnum.....56·17n 13·42e
Björna.....63·34n 18·33e
Bjurholm.....63·56n 19·13e
Boden.....65·50n 21·42e
Bollnäs.....61·21n 16·25e
Borås.....57·43n 12·55e
Borgholm.....56·53n 16·39e
Borlänge.....60·29n 15·25e
Boxholm.....58·12n 15·03e
Bräcke.....62·43n 15·27e
Burgsvik.....57·03n 18·16e
Bydalen.....63·06n 13·47e
Byxelkrok.....57·20n 17·00e
Charlottenberg..59·53n 12·17e
Degeberga.....55·50n 14·05e
Degerfors.....59·14n 14·26e
Deje.....59·36n 13·28e
Djurås.....60·33n 15·08e
Ed.....58·55n 11·55e
Edsbro.....59·54n 18·29e
Edsbruk.....58·02n 16·28e
Edsbyn.....61·23n 15·49e
Enköping.....59·38n 17·04e
Eskilstuna.....59·22n 16·30e
Fagersta.....60·00n 15·47e
Falkenberg.....56·54n 12·28e
Falköping.....58·10n 13·31e
Falun.....60·36n 15·38e
Färjestaden.....56·39n 16·27e
Fårösund.....57·52n 19·03e
Finspång.....58·43n 15·47e
Fjällåsen.....67·29n 20·10e
Flen.....59·04n 16·35e
Fredrika.....64·05n 18·24e
Fredriksberg.....60·08n 14·23e
Frösön.....63·11n 14·32e
Frövi.....59·28n 15·22e
Furudal.....61·10n 15·08e
Gäddede.....64·30n 14·09e
Gällivare.....67·07n 20·45e
Gamleby.....57·54n 16·24e
Gävle.....60·40n 17·10e
Gislaved.....57·18n 13·32e
Glommersträsk...65·16n 19·38e
Gnarp.....62·03n 17·16e
Göteborg (Gothenburg)...57·43n 11·58e
Grangärde.....60·16n 14·59e
Gunnarn.....65·00n 17·40e
Hagfors.....60·02n 13·42e
Hällefors.....59·47n 14·30e
Hällnäs.....64·19n 19·38e
Hallsberg.....59·04n 15·07e
Hallstahammar..59·37n 16·13e
Hallstavik.....60·03n 18·36e
Halmstad.....56·39n 12·50e
Hälsingborg.....56·03n 12·42e
Hammerdal.....63·36n 15·21e
Haparanda.....65·50n 24·10e
Harmånger.....61·56n 17·13e
Härnösand.....62·38n 17·56e
Hässleholm.....56·09n 13·46e
Hede.....62·25n 13·30e
Hedemora.....60·17n 15·59e
Hemse.....57·14n 18·22e
Hofors.....60·33n 16·17e
Höganäs.....56·12n 12·33e
Högsby.....57·10n 16·02e
Höljes.....60·54n 12·36e
Holmsund.....63·42n 20·21e
Hoting.....64·07n 16·10e
Hovmantorp.....56·47n 15·08e
Huddinge.....59·14n 17·59e
Hudiksvall.....61·44n 17·07e
Hultsfred.....57·29n 15·50e
Huskvarna.....57·48n 14·16e
Idre.....61·52n 12·43e
Järvsö.....61·43n 16·10e
Jävre.....65·09n 21·59e
Jokkmokk.....66·37n 19·50e
Jönköping.....57·47n 14·11e
Jörn.....65·04n 20·02e
Junsele.....63·41n 16·54e
Kåbdalis.....66·10n 20·00e
Kalmar.....56·40n 16·22e
Kårböle.....61·59n 15·19e
Karesuando.....68·25n 22·30e
Karlshamn.....56·10n 14·51e
Karlskoga.....59·20n 14·31e
Karlskrona.....56·10n 15·35e
Karlstad.....59·22n 13·30e
Karungi.....66·03n 23·57e
Katrineholm.....59·00n 16·12e
Kil.....59·30n 13·19e
Kinna.....57·30n 12·41e
Kiruna.....67·51n 20·16e
Kisa.....57·59n 15·37e
Klimpfjäll.....65·04n 14·52e
Klintehamn.....57·24n 18·12e
Knäred.....56·32n 13·19e
Koskullskulle....67·12n 20·50e
Kramfors.....62·56n 17·47e
Krangede.....63·09n 16·05e
Kristdala.....57·24n 16·11e
Kristianstad.....56·02n 14·08e
Kristinehamn....59·20n 14·07e
Kungälv.....57·52n 11·58e
Kungsbacka.....57·29n 12·04e
Kvikkjokk.....66·55n 17·50e
Laholm.....56·31n 13·02e
Laisvall.....66·05n 17·10e
Landskrona.....55·52n 12·50e
Långban.....59·51n 14·15e
Lärbro.....57·47n 18·47e
Laxå.....58·59n 14·37e
Lenhovda.....57·00n 15·17e
Lesjöfors.....59·59n 14·11e
Lickershamn.....57·50n 18·31e
Lidköping.....58·30n 13·10e
Lillhärdal.....61·51n 14·04e
Limmared.....57·32n 13·21e
Lingbo.....61·03n 16·41e
Linköping.....58·25n 15·37e
Lit.....63·19n 14·49e
Ljugarn.....57·19n 18·42e
Ljungaverk.....62·29n 16·03e
Ljungby.....56·50n 13·56e
Ljusdal.....61·50n 16·05e
Lövånger.....64·22n 21·18e
Lövstabruk.....60·24n 17·53e
Luleå.....65·34n 22·10e
Lund.....55·42n 13·11e
Luspebryggan....67·01n 19·51e
Lycksele.....64·36n 18·40e
Lysekil.....58·16n 11·26e
Malåträsk.....65·11n 18·44e
Malmberget.....67·10n 20·40e
Malmö.....55·36n 13·00e
Malung.....60·40n 13·44e
Mariestad.....58·43n 13·51e
Markaryd.....56·26n 13·36e
Märsta.....59·37n 17·51e
Mellansel.....63·26n 18·19e
Mellerud.....58·42n 12·28e
Mjölby.....58·19n 15·08e
Mölndal.....57·39n 12·01e
Mönsterås.....57·02n 16·26e
Mora.....61·00n 14·33e
Motala.....58·33n 15·03e
Näsåker.....63·26n 16·54e
Nässjö.....57·39n 14·41e
Nedre Soppero...68·03n 21·46e
Nordmaling.....63·34n 19·30e
Norrköping.....58·36n 16·11e
Norrtälje.....59·46n 18·42e
Norsjö.....64·55n 19·29e
Nybro.....56·45n 15·54e
Nyköping.....58·45n 17·00e
Nynäshamn.....58·54n 17·57e
Ockelbo.....60·53n 16·43e
Öje.....60·49n 13·51e
Örbyhus.....60·14n 17·42e
Örebro.....59·17n 15·13e
Örkelljunga.....56·17n 13·17e
Örnsköldsvik.....63·18n 18·43e
Oskarshamn.....57·16n 16·26e
Oskarström.....56·48n 12·58e
Östersund.....63·11n 14·39e
Osthammar.....60·16n 18·22e
Ottenby.....56·14n 16·25e
Överkalix.....66·21n 22·56e
Övertorneå.....66·23n 23·40e
Oxelösund.....58·40n 17·06e
Pajala.....67·11n 23·29e
Perstorp.....56·08n 13·23e
Piteå.....65·20n 21·30e
Polcirkeln.....66·33n 21·01e
Ramsele.....63·33n 16·29e
Ramsjö.....62·11n 15·39e
Rätansbyn.....62·29n 14·32e
Rättvik.....60·53n 15·06e
Rensjön.....68·05n 19·49e
Riksgränsen.....68·24n 18·12e
Risbäck.....64·42n 15·32e
Robertsfors.....64·11n 20·51e
Ronneby.....56·12n 15·18e
Roslags Näsby...59·26n 18·04e
Ryd.....56·28n 14·41e
Sädvaluspen.....66·24n 16·51e
Säffle.....59·08n 12·56e
Sala.....59·55n 16·36e
Sälen.....61·10n 13·16e
Sandviken.....60·37n 16·46e
Särna.....61·41n 13·08e
Sävsjö.....57·25n 14·40e
Sigtuna.....59·37n 17·43e
Simrishamn.....55·33n 14·20e
Skellefteå.....64·46n 20·57e
Skelleftehamn...64·41n 21·14e
Skillingaryd.....57·26n 14·05e
Skövde.....58·24n 13·50e
Slagnäs.....65·34n 18·05e
Söderhamn.....61·18n 17·03e
Soderköping.....58·29n 16·18e
Södertälje.....59·12n 17·37e
Sollefteå.....63·10n 17·16e
Sollentuna.....59·28n 17·54e
Sölvesborg.....56·03n 14·33e
Sorsele.....65·30n 17·30e
Ställdalen.....59·56n 14·56e
Stensele.....65·05n 17·09e
Stockholm.....59·20n 18·03e
Stöllet.....60·24n 13·16e
Storlien.....63·19n 12·06e
Storuman.....65·06n 17·06e
Storvreta.....59·58n 17·42e
Strömstad.....58·56n 11·10e
Strömsund.....63·51n 15·35e
Sundsvall.....62·23n 17·18e
Sunne.....59·50n 13·09e
Svanskog.....59·11n 12·33e
Svappavaara.....67·39n 21·04e
Svärdsjö.....60·45n 15·55e
Sveg.....62·02n 14·21e
Svenljunga.....57·30n 13·07e
Tännäs.....62·27n 12·40e
Tärendö.....67·10n 22·38e
Tärna.....65·43n 15·16e
Tierp.....60·20n 17·30e
Tingsryd.....56·32n 14·59e
Töcksfors.....59·30n 11·50e
Tomelilla.....55·33n 13·57e
Töre.....65·54n 22·39e
Torsby.....60·08n 13·00e
Torup.....56·58n 13·05e
Tranås.....58·03n 14·59e
Trelleborg.....55·22n 13·10e
Trollhättan.....58·16n 12·18e
Trosa.....58·54n 17·33e
Uddevalla.....58·21n 11·55e
Ulricehamn.....57·47n 13·25e
Umeå.....63·50n 20·15e
Umfors.....65·56n 15·00e
Umnäs.....65·24n 16·10e
Uppsala.....59·52n 17·38e
Vålådalen.....63·10n 12·57e
Valdemarsvik....58·12n 16·36e
Valsjöbyn.....64·04n 14·08e
Vänersborg.....58·22n 12·19e
Vännäs.....63·55n 19·45e
Vansbro.....60·31n 14·13e
Varberg.....57·06n 12·15e
Värnamo.....57·11n 14·02e
Västerås.....59·37n 16·33e
Västervik.....57·45n 16·38e
Växjö.....56·42n 14·49e
Vetlanda.....57·26n 15·04e
Vidsel.....65·51n 20·24e
Vilhelmina.....64·37n 16·39e
Vimmerby.....57·40n 15·51e
Vindeln.....64·12n 19·44e
Virserum.....57·19n 15·35e
Visby.....57·38n 18·18e
Vissefjärda.....56·32n 15·35e
Vittangi.....67·41n 21·36e
Ystad.....55·25n 13·49e
Ytterhogdal.....62·12n 14·51e

Physical features and points of interest

Amungen, *lake*...61·09n 15·39e
Ångermanälven, *river*.....62·48n 17·56e
Åreskutan, *mountain*.....63·26n 13·06e
Åsnen, *lake*.....56·38n 14·42e
Ätran, *river*.....56·53n 12·30e
Baltic Sea.....56·30n 19·00e
Bolmen, *lake*.....56·55n 13·40e
Bothnia, Gulf of..63·00n 20·30e
Bottenhavet, *gulf*.62·00n 19·30e
Bottenviken, *gulf*.64·30n 23·00e
Dalälven, *river*...60·38n 17·27e
Dalarna (Dalecarlia), *historic region*...61·01n 14·04e
Fårön, *island*.....56·56n 19·08e
Faxälven, *river*...63·13n 17·13e
Götaälv, *river*....57·42n 11·52e
Götaland, *historic region*.....57·30n 15·03e
Gotland, *island*...57·30n 18·33e
Gotska Sandön, *island*.....58·23n 19·16e
Gotska Sandön National Park...58·23n 19·16e
Helagsfjället, *mountain*.....62·55n 12·27e
Hjälmaren, *lake*..59·15n 15·45e
Hornavan, *lake*...66·10n 17·30e
Indalsälven, *river*.62·31n 17·27e
Kalixälven, *river*..65·50n 23·11e
Kalmarsund, *sound*.....56·40n 16·25e
Kattegat, *strait*...57·00n 11·00e
Kebnekaise, *mountain*.....67·53n 18·33e
Könkämäälven, *river*.....68·29n 22·17e
Lainioälven, *river*.....67·22n 23·39e
Lapland, *historic region*...67·30n 21·00e
Ljungan, *river*....62·19n 17·23e
Ljusnan, *river*....61·12n 17·08e
Luleälven, *river*...65·35n 22·03e
Mälaren, *lake*....59·30n 17·12e
Marsfjället, *mountain*.....65·05n 15·28e
Muddus National-park.....67·00n 20·16e
Norra Kvarken, *strait*.....63·36n 20·43e
Norra Storfjället, *mountain*.....65·52n 15·18e
Norrland, *historic region*.....65·00n 18·00e
Öland, *island*....56·45n 16·38e
Padjelanta Nationalpark....67·28n 16·41e
Peljekaise Nationalpark....66·18n 16·58e
Piteälven, *river*...65·14n 21·32e
Sareks National-park.....67·15n 17·30e
Sarjektjåkko, *mountain*.....67·25n 17·46e
Siljan, *lake*.....60·50n 14·45e
Skalka, *lake*.....66·50n 18·46e
Skåne, *historic region*.....55·59n 13·30e
Skellefteälven, *river*.....64·42n 21·06e
Småland, *historic region*.....57·18n 15·03e
Sommen, *lake*....58·01n 15·15e
Sound, The.....55·50n 12·40e
Stora Lulevatten, *lake*.....67·10n 19·16e
Stora Sjöfallet, *waterfall*.....67·29n 18·21e
Stora Sjöfallet Nationalpark...67·44n 18·16e
Storavan, *lake*....65·42n 18·12e
Storuman, *lake*...65·14n 16·54e
Storvindeln, *lake*.65·43n 17·05e
Sulitelma, *mountain*.....67·08n 16·24e
Svealand, *historic region*.....60·30n 15·30e
Sylarna (Sylene), *mountains*.....63·02n 12·13e
Torneträsk, *lake*..68·20n 19·10e
Uddjaur, *lake*....65·55n 17·49e
Umeälven, *river*..63·47n 20·16e
Vänern, *lake*.....58·55n 13·30e
Vättern, *lake*.....58·24n 14·36e
Vindelälven, *river*.....63·54n 19·52e


Norwegian Sea
ATLANTIC OCEAN
Arctic Circle
Gulf of Bothnia
BALTIC SEA
SWEDEN
FINLAND
NORRBOTTENS LÄN
VÄSTERBOTTENS LÄN
NORRLANDS LÄN
JÄMTLANDS LÄN
GÄVLEBORGS LÄN
KOPPARBERGS LÄN
DALARNA
VÄRMLANDS LÄN
ÖREBRO LÄN
VÄSTMANLANDS LÄN
UPPSALA LÄN
STOCKHOLMS LÄN
SÖDERMANLANDS LÄN
ESKILSTUNA LÄN
GÖTEBORGS OCH BOHUS LÄN
ÄLVSBORGS LÄN
SKARABORGS LÄN
ÖSTERGÖTLANDS LÄN
GÖTALAND
SVEALAND
SMÅLAND
KRONOBERGS LÄN
KALMAR LÄN
BLEKINGE LÄN
SKÅNE
HALLANDS LÄN
GOTLANDS LÄN
GOTLAND
ÅLAND
AHVENANMAA ÅLAND
Kiruna
Gällivare
Malmberget
Jokkmokk
Arjeplog
Vilhelmina
Lycksele
Vindeln
Boden
Luleå
Piteå
Skellefteå
Umeå
Örnsköldsvik
Ostersund
Sundsvall
Härnösand
Hudiksvall
Gävle
Sandviken
Falun
Borlänge
Ludvika
Hagfors
Karlstad
Kristinehamn
Örebro
Lindesberg
Uppsala
Västerås
Eskilstuna
Strängnäs
Nyköping
Oxelösund
Norrköping
Linköping
Motala
Vadstena
Mjölby
Skövde
Lidköping
Vänersborg
Trollhättan
Uddevalla
Strömstad
Göteborg
Gothenburg
Mölndal
Borås
Jönköping
Nässjö
Värnamo
Växjö
Kalmar
Oskarshamn
Västervik
Visby
Halmstad
Varberg
Falkenberg
Ängelholm
Helsingborg
Landskrona
Malmö
Lund
Trelleborg
Ystad
Simrishamn
Kristianstad
Karlskrona
Hässleholm
Olofström
Ronneby
Ljungby
Alvesta
Tingsryd
Nybro
Oskarshamn
Borgholm
Färjestaden
Ottenby
Hoburgen
Copenhagen
DENMARK
Skagerrak
Kattegat
Öresund
Trondheim
Oslo
Kilpisjärvi
Karesuando
Abisko
Kebnekaise
Gällivare
Malmberget
Jokkmokk
Arjeplog
Vilhelmina
Lycksele
Vindeln
Boden
Luleå
Piteå
Skellefteå
Umeå
Örnsköldsvik
Ostersund
Sundsvall
Härnösand
Hudiksvall
Gävle
Falun
Borlänge
Ludvika
Karlstad
Örebro
Uppsala
Västerås
Eskilstuna
Stockholm
Norrköping
Linköping
Motala
Vadstena
Mjölby
Skövde
Lidköping
Vänersborg
Trollhättan
Uddevalla
Strömstad
Göteborg
Gothenburg
Mölndal
Borås
Jönköping
Nässjö
Värnamo
Växjö
Kalmar
Oskarshamn
Västervik
Visby
Halmstad
Varberg
Falkenberg
Ängelholm
Helsingborg
Landskrona
Malmö
Lund
Trelleborg
Ystad
Simrishamn
Kristianstad
Karlskrona
Hässleholm
Olofström
Ronneby
Ljungby
Alvesta
Tingsryd
Nybro
Oskarshamn
Borgholm
Färjestaden
Ottenby
Hoburgen
Copenhagen
DENMARK
Skagerrak
Kattegat
Öresund
Trondheim
Oslo
Kilpisjärvi
Karesuando
Abisko
Kebnekaise
Kiruna

tant publicly owned industries into one concern, Statsföretag AB (State Enterprise Ltd.), the state has, however, increased its importance and influence in Swedish industry. Mixed enterprises, with half-public and half-private ownership, also exist.

Patterns of taxation and public expenditures

The role of government, however, cannot be measured only by its direct management of industrial or business enterprise: a large percentage of the GNP goes into the public sector in the form of taxes and social charges, and much of this is returned to the private sector through insurance payments of different kinds. Taxes and social charges paid to the state and to the local authorities constitute slightly more than 40 percent of the GNP, the largest amount of any nation, and this figure is likely to approach 50 percent during the 1970s. About one-third of this money returns to the private sector, especially to private households.

About one-fourth of Sweden's GNP finances public activities—of which about one-half goes to local authorities—and a good deal of this money is returned indirectly to the citizens in the form of services, the largest of which in terms of expenditure are education, health, defense, and welfare.

The taxes are charged partly as a strongly progressive national income tax, partly as a local proportional income tax, partly as a tax on net wealth and as inheritance and gift taxes. To this should be added the profit taxes on businesses and the indirect taxes, particularly the value-added tax. For the corporations the tax is about 50 percent of net profits, in addition to which a tax on dividends to shareholders is imposed.

Labour and management. Sweden has been called the most thoroughly organized country in the world. Approximately 80 percent of all wage earners belong to a trade union.

The workers' trade unions, united in the Swedish Confederation of Trade Unions (LO), have 1,733,000 members; the Central Organization of Salaried Employees (TCO) has 720,000; and the Swedish Confederation of Professional Associations (SACO) has 120,000. Employers are united in the Swedish Employers Confederation (SAF) covering about 24,100 companies with 1,250,000 employees.

The Swedish labour market long has been known as peaceful, with a good mutual understanding between employers and employees and with a tradition of negotiation that has guaranteed the proper functioning of agreements. The extended wildcat strike that broke out at the state-owned mining company, Luossavaara-Kiirunavaara AB (LKAB) in 1970 shook the Swedish labour market rather roughly, though the direct economic losses were quite modest.

Economic policy making and outlook. Since World War II, Swedish economic policy has remained in the liberal tradition in the sense that the government intervened only in general matters and has tried to keep regulation to a minimum. The theory has been, as far as possible, to allow companies independent economic responsibility and then, through the growing public sector, to provide increasing resources for social expenses and other forms of public consumption. In recent years, however, the Social Democratic government has begun a policy involving more detailed regulations, bringing about strains in the formerly cooperative relation between government and industry.

Symptoms of economic change

The long-range outlook for the Swedish economy in the early 1970s seems somewhat less bright than earlier. Certain Swedish advantages in the world marketplace are no longer so pronounced. The LKAB conflict hurt the peaceful image of the Swedish labour market, while Sweden's long lead in welfare is being diminished by several nations, such as France and West Germany. Growing research-and-development costs make it more difficult to pioneer on the world market. The Swedish refusal to join the EEC as a full member, which, it was felt, might impeach its neutrality, could produce further unfavourable consequences. Overall, the Swedish economy is not in a bad position, but the advantages of yesterday no longer can be expected to pertain as strongly as before.

TRANSPORTATION

Trucks are the main means of transporting heavy goods, while the private automobile accounts for about 85 percent of all passenger traffic. As a result, demands for expressways are increasing.

By the middle of the 19th century, the railways had begun to compete with older means of transportation. Three main principles ruled the construction of the railways: they must not follow existing shipping routes, whether coastal or canal; they should, except for terminal points, be built through sparsely populated or poor districts rather than through rich areas that often could support their own transportation networks; and a connected and coordinating network of railways should be constructed. Through this policy, new towns grew up at the junctions.

The present national traffic policy, dating from 1963, requires that every kind of transport or communications bear its own expenses. This policy has affected the railways in particular: they have suffered from low profits followed by long-term shutdowns and have been able to maintain themselves only by the movement of goods, yet with a reduced share of even this market.

Domestic shipping along the coasts and on lakes and canals has been extremely important throughout Swedish history. Today, intranational shipping has a limited share of all transports. International shipping, on the other hand, is very lively, with at least nine ports engaged in large-scale commerce with foreign nations.

The air services are dominated by the Scandinavian Airline System (SAS), which is owned chiefly by the states of Sweden, Denmark, and Norway. The interests of SAS are concentrated on international aviation, but a few domestic scheduled services are maintained in Sweden. The rest of the domestic traffic is run by Linjeflyg (LIN), half of which is owned by SAS. The most important airports are in Stockholm, Göteborg, and Malmö.

State ownership of airlines and railways

The interests of the state in transportation and communications are wide. The railways are 95 percent owned and run by the state, which maintains bus traffic on a large scale and has two subsidiaries that dominate the trucking industry.

Local transportation consists mainly of buses, though Stockholm has a subway. Streetcars were abolished in several cities in connection with the switchover to right-hand traffic that took place in 1967, but Göteborg and Norrköping retained trolley service. Generally, local traffic is controlled by local authorities.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. The crown of Sweden is inherited in straight lineal descent on the male side. For decades, however, the monarch's position has been almost exclusively symbolic, a status to be confirmed by fundamental law among other far-reaching constitutional reforms that have been in preparation since the 1950s. These reforms already have led, through partial implementation, to important changes concerning the parliament, or Riksdag, and the electoral system. Other changes in the basic Instrument of Government of 1809 are anticipated.

The national level. Since January 1971, Sweden has had a unicameral parliament with 350 members elected for three-year terms by direct universal suffrage. The parliamentary system, in use during most of the 20th century, is now constitutionally established. The Riksdag can, by a vote of "no confidence" by a minimum of 176 members, force the resignation of the government or dismissal of a Cabinet minister, unless the prime minister appeals to the country within a week for a new election. The terms of office after such an election last only for the remainder of the original parliamentary period.

Interactions of administrative and parliamentary bodies

The government, under the leadership of the prime minister, is administered by 12 departments, each headed by a Cabinet-level minister; there are, in addition, six consultative ministers, or ministers without portfolio. This parliament-appointed government has taken over, in large measure, the former prerogatives of the monarch, though formally they function as his advisers. Although through

this arrangement the government often was able to act independently of the Riksdag, the formal prerogatives of the latter have been strengthened by bestowing the right of initiative on the parliamentary committees, which formerly could only discuss propositions from the government. The work of the committees, including constitutional, finance, and 14 other permanent committees, usually is regarded as the most important part of parliamentary activity. The committees are not open to the public, but their reports are published. In many questions, however, the parliamentary-party groups or conferences between party leaders may play an even greater role than the committees.

Certain characteristics of the former dualism between the king (the government) and the Riksdag have been kept. The latter alone has the right of taxation.

Power to enact laws is jointly vested in the Riksdag and in the King in Council, which under the contemporary constitutional practice is the Cabinet appointed in observance of parliamentary principles. Power to institute ordinances and regulations in matters concerned with administration and economic policy is reserved by the Instrument of Government for the King in Council alone, which gives it a wide range of independent action. In practice, however, many of the principles for these so-called economic laws are put before the Riksdag, the decision of which is regarded as binding. So the dividing lines are, on the whole, uncertain.

All lawmaking takes place on the national level—local lawmaking does not exist, because Sweden is a unitary and not a federal state. Ordinances and regulations are of a lower power than laws; they are easier to enact and easier to change. Certain ordinances and regulations of a local character—e.g., parking regulations—can be established by local authorities.

Different rules pertain as to how Swedish laws may be changed. Alterations in the constitutional laws—the Instrument of Government (1809), the Act of Succession (1810), the Parliament Act (1866), and the Freedom of the Press Act (1949)—may be instituted only by the approval of two successive parliaments, between which elections must occur. The civil and criminal laws, on the other hand, may be instituted or changed by the combined agreement of government and parliament. The government itself has authority to change ordinances and regulations.

County and local government. Implementation of administrative decisions is, in part, the responsibility of the 24 county (*län*) administrations, which represent the national government in regional administration. Each of these is headed by a governor appointed by the government.

Parallel to the county–state administrations are 23 county councils and 270 municipal councils, elected locally to provide local administration. One county, Gotland, comprises a single municipality and thus has a municipal council but no county council. Health and medical care are primarily the responsibility of the county councils, while the local school system and almost all other social services are the responsibility of the municipal councils.

Administration of law and justice. Swedish law is the result of a long historical development marked by continuity rather than by abrupt changes. As the result of extensive collaboration with the other Nordic countries, a significant degree of legal uniformity exists throughout Scandinavia, especially in the field of civil law. The Swedish judicial system is built upon written law; practice, customary (or common) law, and judicial precedent retain a strong influence, but the following of precedent is not obligatory. The voluminous body of printed matter—generated by the traditionally careful and prolonged lawmaking process—in which reasons and intent of the law are formulated, is an important source of interpretation for courts, authorities, and individual jurists. Judges and high-ranking administrators—as well as prosecutors, senior police officers, and attorneys—have the same academic training. Moreover, the majority of law graduates serve in one of the courts of first instance for two and a half years to gain experience before taking an appointment in the civil service or embracing other careers.

The primary responsibility for the application of legal rules lies upon the courts and the different administrative authorities. The difference between adjudicative and administrative authorities is less in Sweden than in most European countries: their organization and the division of powers between them have both evolved in response to historical trends rather than to deliberate government planning. Administrative authorities are not partisan representatives of those holding political power at any one time; apart from matters of internal organization and procedure, the lower authorities do not take orders from the higher but follow the generally prevailing rules under their own responsibility. Their decisions may be challenged under appeal provisions resembling those in force for the court system, though in most cases, administrative decisions are not subject to review by a regular court.

Three-tiered hierarchy of courts

Sweden has a three-tier hierarchy of courts: the courts of first instance (*tingsrätter*), the intermediate courts of appeal (*hovrätter*), and the Supreme Court (Högsta Domstolen). In addition to the regular courts of justice, there are also a number of special courts and administrative tribunals.

After a reorganization effective in 1971, which merged the former rural and city courts and considerably reduced their number, Sweden now has just over 100 courts of first instance. These courts, in addition to their primary concern with civil and criminal cases, perform many duties of an administrative nature related to conveyancing, land development, guardianships, probate, and the like. An unusual feature of courts of first instance is the *nämnd*, or panel of lay assessors, who take part in the main hearings in most criminal and some civil cases. Assessors (*nämndemän*) are elected for six-year terms by local representative councils from among eligible local citizens; often they are reelected for successive terms. The *nämnd*, which has medieval traditions and has constituted a significant element of democracy in Swedish public life, is not to be confused with the Anglo-American and Continental jury. Its members are concerned not only with verdicts but also with such matters as the sentences to be imposed in criminal cases. A qualified majority of the lay assessors prevails over the contrary opinion of the judge.

Of the six courts of appeal, or of second instance, the oldest and largest is the Svea Hovrätt in Stockholm, established in 1614. Besides serving as appellate bodies, these courts also are responsible for the administration of the court system within their respective areas and for the further training of judges. In their adjudicative capacity, courts of appeal operate in divisions. No limits are imposed on the right of appeal, and about 5 percent of the total number of cases tried by lower courts pass on to a higher court.

The Supreme Court, that of third and last instance, has 24 justices (*justitieråd*), but it works in three sections. Because of the great value attached to its decisions as precedents, however, the entire court also can meet in plenary session—if, for example, there is cause to depart from previously applied principles on any point. The Code of Judicial Procedure defines the classes of cases that may be considered by the Supreme Court after their trial in courts of appeal. A separate section of the court, comprising three justices, determines whether to accept a case for review, and about one-seventh of all appellate decisions are reviewed by the Supreme Court. Like the courts of appeal, the Supreme Court need not confine itself to points of law but may also consider new evidence. Its decisions are final.

Among the special courts, the Court of Impeachment (Riksrätten), empowered under the Instrument of Government to try Cabinet ministers, has not been convened since 1854. Of greater importance is the Labour Court (Arbetsdomstolen), established in 1928, which has exclusive competence in controversies concerning the interpretation and application of collective bargaining agreements. Of its seven members, two represent labour and

two management. There are also, for example, first instance special water courts (with Svea Court of Appeal and the Supreme Court as appeal instances), special courts for cases relating to land ownership, boundaries and expropriation, a special insurance court, and a court to deal with such matters as undue clauses in standard contracts.

As already noted, the decisions of administrative authorities are not appealable to an ordinary court of justice. Instead, a complainant can take his grievance to the next highest authority or to a specialized tribunal. In theory, a complainant may appeal the decision of a lower administrative authority to the local county administration or to a specialized agency attached to it. From there, two channels are available. In certain cases—of which dissatisfactions with tax assessments are the most numerous—an appeal against the ruling on county level is lodged with one of Sweden's oldest and most notable administrative tribunals, the Fiscal Court of Appeal (Kammarrätten) in Stockholm. The highest instance in the tribunal system is the Supreme Administrative Court (Regeringsrätten). Cases falling within the competence of this court are enumerated by special statute: generally speaking, these are cases in which the legal element prevails, as in taxation or in the institutionalization of the insane, alcoholics, delinquent youth, and the like. The Supreme Administrative Court is also the final court of appeal against the decisions of local representative councils and authorities in those cases in which such decisions are appealable. The second channel, applicable in all other cases, has the King in Council as its highest instance. Like their counterparts in the regular court system, the higher instances in the tribunal system of adjudication are empowered to review questions of both fact and law.

The institution of the ombudsman

The control system—of great importance to guarantee against injustice from either the judiciary or the administrative authorities—is handled to a large extent by the three *justitieombudsmän* appointed by the Riksdag. Their principal task is to oversee the observance of laws and constitutions and to prosecute officials and civil servants who have committed illegal acts. Consequently, the duty of the ombudsman is aimed especially at controlling the ways in which the authorities execute their power and guaranteeing that the rights of the individual are not violated. The Swedish institution of the ombudsman has served as a model for other nations, including the United Kingdom, Denmark, Norway, and Finland, and it has been instituted at various levels of local government throughout the country.

The political process. Sweden's triennial elections result in proportional representation: *i.e.*, each political party receives the same proportion of seats as it receives votes in the general elections. To avoid a proliferation of small parties, however, a party must get either 12 percent of the votes in one constituency or 4 percent of the votes nationwide to be represented in parliament. Of the 350 seats, 310 are elected directly, while the remaining 40 are distributed among the parties so as to achieve proper proportionality.

Proportional representation and citizen participation

Five parties are represented in the Riksdag. With the exception of 100 days in 1936, the Social Democratic Labour Party has formed the government since 1932. At two periods it has formed a coalition with the Centre (former Agrarian) Party. The parties receive financial support from the state in proportion to the number of seats they hold.

All Swedish citizens are eligible and may vote at 20 years of age. In elections, between 80 and 85 percent of the voters usually participate, and about 30 percent of the voters are members of one or another party. During the 1960s these parties endeavoured to broaden their inner democracy by encouraging different conferences of members and holding primary elections to select candidates.

National defense. Sweden's defenses, based entirely on its own financial resources, devolve from its traditions of non-alignment and neutrality. Military and civil defenses, as well as economic and psychological defenses, work closely together—the last two concepts referring, respectively, to the stockpiling of essential goods and materials for wartime and to the maintenance of domestic morale. Service in Sweden's military is compulsory for all able-bodied men between 18 and 47 years of age. Most conscripts have ten months of basic training, which later is followed by five refresher courses. Roughly 50,000 conscripts are called up every year.

Since World War II, Sweden has expended about 4–5 percent of its GNP on defense. Outside its own territory, Sweden takes an active part in the United Nations' peace-keeping operations: since 1956, about 32,000 members of its armed forces have served in different UN missions, particularly in the Gaza Strip, the Congo, and Cyprus.

The social milieu. Peace, economic prosperity, and a tradition of social and political stability are among the factors that allowed Sweden to emerge in the 20th century as the prototypal image of a nation ideally combining free electoral practice with a pronounced socially oriented concern for the well-being of its citizens. Incomes and standards of living generally are looked upon as the highest in Europe, but these have tended to decline relative to other countries in recent years.

Income and other social differentials

Internally, however, differences in level of income and, consequently, of social status and of educational and other attainments are not radically different from the same phenomena in other advanced industrialized nations. Rises in the cost of living have paralleled rises in income, and employment among the young, especially women, tends toward irregularity or the low-income brackets. Social conditions in various regions of Sweden also restrict choice of work, education, or access to types of social service from the state.

Welfare and health. Sweden long has been regarded as a leader in social welfare, and, in spite of inequities, the rights of the citizen for social assistance are manifold. Compulsory health-insurance programs reimburse all but small portions of doctors' fees, hospital treatment, and medicines. The national pension system is paid, beginning at 67 years of age, to all citizens and to foreigners resident in Sweden for a certain period. A supplementary pension, introduced in 1960 and based upon average yearly income, will be paid in full to all persons born after 1914.

Children's allowances are paid for each child under 16 years of age, and every mother receives a stipend at each childbirth. Unemployment insurance is based on voluntary premiums, but it is subsidized by the state. Between 1965 and 1970, state expenditures for social security were doubled.

The relatively excellent health conditions of the Swedish people can be attributed to the rugged but basically healthy climate and to the many factors that have made it possible to offer free medical care to every citizen. Medical practitioners of all kinds are able to remain in the vanguard of research, and hospital facilities have been kept among the most advanced in Europe. Preventive medicine, in terms of information to the public and early diagnosis, is in an advanced state. Tuberculosis, the disease that once claimed the most lives, is far down on the scale of causes of death. The main causes of death are, today, those of the other advanced nations: circulatory diseases, cancer, and the like. In the more remote areas, health care is provided mainly through a system of smaller hospitals and of provincial doctors, nurses, and midwives employed by the county councils, each covering a specific area.

Enrollments in the six medical faculties—which, like all faculties of higher education in Sweden, are supported by the government—swelled during the 1960s, promising a significant increase in the number of registered doctors. Students are eligible to apply directly for medical training after completion of their course in the gymnasium, or pre-university school. There is no entrance examination or requirement of a premedical course, but competition is severe. A new system of medical education introduced in 1969 shortened the basic curriculum of undergraduate training, mainly by doing away with the nine-month internship period. The basic curriculum is followed by

postgraduate training, the first part of which provides for a compulsory "general service period" lasting 21 months. After that the doctor takes either specialist training, usually embracing residencies totalling four to five years, or a three-year period of training for general practice.

Housing. The average Swedish household consists of about three persons. Most of them live in densely populated areas and about one-third of them in one of the three biggest cities and their metropolitan areas. Fewer than one-half of them reside in individual houses. Swedes spend an average of 17 percent of their income on housing. In spite of large-scale construction, a shortage of dwellings exists around the largest cities: Stockholm, for example, had more than 100,000 people waiting for homes in the early 1970s.

Housing shortages

About one-half of Sweden's housing is privately owned, one-quarter owned by housing cooperatives, and most of the remainder owned by semipublic bodies. Since about 1950, however, private ownership has fallen sharply, due in large measure to the fact that state assistance for housing is not equal for all types of investors: private builders of apartment houses get, at most, 85 percent state loans, whereas cooperative and semipublic bodies may receive up to 95 or even 100 percent.

Education. School attendance is compulsory for nine years, beginning at age seven. The nine-year comprehensive school is divided into three-year stages: lower, middle, and upper. All pupils follow the same curriculum for six years. The first foreign language, English, is introduced in the third year. In upper classes, training in German and French can also be chosen, while in the gymnasium, for example, Russian, Spanish, Finnish, and Italian can be chosen as well. From the seventh year the curriculum is divided into different lines chosen by the pupils themselves from subjects or groups of subjects. Special education is given to all pupils suffering from physical or mental handicaps.

An important principle in Swedish education is that everyone who wants to continue studying after the comprehensive school shall have the opportunity. About 30 percent of pupils choose courses that will qualify them for university study. Other, less theoretical curriculums offer a vocational training.

Sweden's five universities are in Stockholm, Göteborg, Uppsala (the oldest, founded in 1477), Lund, and Umeå (the youngest, founded in 1963). The first four have branches in other cities. Other institutions for higher education offer training, for example, in technology, medicine, art and music, and journalism. More than 120,000 students were studying at universities and other institutions in the early 1970s; the corresponding figure in 1951 was 21,000.

Adult-education programs

Strong emphasis is given to adult education. Since 1968 free evening and daytime classes have been sponsored and administered by both local authorities and state government. The instruction follows the same curriculums as the ordinary school, and the adult pupils can either follow classes in separate subjects or choose the entire alternative. The number of students is increasing, though irregularly, hovering around 100,000 per year in the early 1970s.

The folk high schools (*folkhögskolorna*) also offer adult education for those 18 years of age and above. Radio and television courses and instruction by correspondence are forms of education with a large geographical range. Study circles arranged by a number of "study societies" cover a large group of subjects from hobbies to purposeful school or university studies.

Retraining courses for new jobs, funded by the state, attracted about 92,000 students during 1968–69. Training in Swedish for immigrants also is paid for by the state and arranged by study societies.

CULTURAL LIFE AND INSTITUTIONS

In the pre-industrial period the cultural manifestations of the small upper class were different and separated from the popular ones. The upper classes took their patterns from the European currents and passed them down to ordinary people. In the other direction, inspiration from popular songs, art motifs, and handicraft work enriched the higher culture.

The changing cultural milieu

Today, popular and upper class cultures are no longer separate in this way. The popular political movements of the 20th century advanced the notion that education and the understanding of the arts were the rights of every man, but this idealistic hope has not been met fully. New forms of entertainment—television, weekly magazines, and popular music of the international type—have taken their places. Poetry, avant-garde literature, and classical music still belong to small specialized groups.

Improved means of communication—both of such media as television and of the greater travel of foreigners to Sweden and of Swedes abroad—have had other effects. As in many other Western nations, differences between rural and urban life-styles are growing smaller as the population becomes more homogeneous and more internationally influenced. Foreign foods appear increasingly alongside traditional Swedish dishes, and wine has tended to decrease the consumption of hard liquor. The Swedish youth culture has its "pop" music and musicians and its contemporary international mode of dress. Drugs have posed a problem in the growing number and decreasing age of addicts, though under Swedish law many substances—*e.g.*, amphetamines—are prohibited as narcotics that are not so classed in most other countries. Since World War II, however, Swedish youth have tended to lead the way toward greater awareness of, and participation in, national and international political activities, solutions to the problems of the world's developing nations, and similar areas of concern—often in response to the atmosphere created by the informative reports and documentary materials widely available through public media, but often, also, with a consequent lessening of interest in the traditional fine arts.

The movement for equality between men and women is of long duration in Sweden. Positive results include greater female participation in the labour market, in higher education, and in politics; broader access to the professions (including the clergy); and more equality in wages and salaries. The traditional male–female relations likewise have moved toward greater equality, accounting in part, perhaps, for the reputation of sexual permissiveness on the part of the Swedish women—an image created in large measure through misunderstanding and exaggeration. Related to this diminution of double standards of morality is the production and sale of allegedly pornographic materials, which is probably no larger in quantity than elsewhere but which is covered by liberal legislation except in such matters as window display.

The state of the arts. The postwar period has been lively: the "Monday group" of composers was inspired by the antiromantic Hilding Rosenberg in the 1920s and drew also upon leading modern composers from abroad. Their repertoire includes orchestral, operatic, and religious works. The vital Swedish folk song has been developed further by a number of musicians, while Swedish jazz became another phenomenon. A number of Swedish opera singers gained renown throughout the West.

Since the 1920s, Harry Martinson has been a favourite poet in Sweden. His motifs range from the romantic Swedish countryside to global and cosmic visions. Among novelists, Vilhelm Moberg is widely known for the epic and dramatic force in his novels of the Swedish emigrants in the United States and of the farmer-soldiers of old Sweden. Younger novelists tend to treat the social and moral aspects of society.

A multifacetted theatre, opera, and ballet life is found in Sweden. Contemporary international plays are staged throughout the country, and the premieres of several of the posthumously performed plays of the American Eugene O'Neill took place in Sweden. Birgit Cullberg has attained international fame as director of the Swedish Royal Ballet in Stockholm. The Swedish stage and film director Ingmar Bergman gained critical acclaim outside Sweden with his film *Wild Strawberries* (1957). Subsequently, his earlier and later films have been shown throughout the world, and he has been hailed as one of the major film makers of all time.

Modern Swedish art was inspired by early-20th-century romantic nationalism, originating with such painters as Carl Larsson, Anders Zorn, and Bruno Liljefors. Carl Milles, who dominated monumental sculpture in the 1920s, is as famous abroad. At the Paris World's Fair in 1925, an important connection was established between Swedish industry and designers who had both academic art education and popular handicraft tradition. Since then, this combination has been vital, and superb results have been achieved in glassware, ceramics, woodwork, textiles, furniture, silver, and stainless steel.

Folk culture. The genuine rural folk traditions are disappearing with settlement in urban areas. Among those remaining in Gotland, Dalarna, and other areas are special national costumes, dances, folk musicians, and the like, and many traditions are found even in urban settlements. Spring is celebrated on the last night of April with big fires across the country. May 1 is the day of workers' demonstrations. The bright Midsummer Night is celebrated with a big cross placed into the ground and laid with flowers and leaves; children dance around it. Some celebrations have a religious background: Advent, Lucia, Christmas, Easter, and Whitsuntide. The Lucia candlelights are a rather new but very popular custom performed on the dark morning of December 13.

Traditional festivals

Cultural institutions. The traditional cultural institutions were initiated and protected by the king. The Drottningholm Theatre; the Swedish Royal Opera; the Royal Dramatic Theatre; the Royal Academy of Music; the Swedish Academy; the Royal Academy of Letters, History, and Antiquities; the Royal Library; and the Royal Academy of Sciences, all in Stockholm, date from the 18th century. The universities in Lund and Uppsala have traditions from the Middle Ages, when those cities were major European centres of learning.

In modern Sweden organizations and institutions for cultural activities usually can depend upon state and municipal contributions. The state theatre, the state concerts, and the state exhibitions travel around the country. Idealistic organizations collaborate with the labour groups in establishing the 240 *folkets parker* ("peoples' parks"), a type of out-of-doors entertainment hall for dancing, theatre, shows, and concerts in the summer. These organizations and others, for religious or temperance purposes, have educational branches concerned with adult educational programs involving the folk high schools and evening classes. Such activities are financed by government.

The Nobel prizes

Of special note are the Nobel prizes, initiated in 1901 through a bequest by Swedish millionaire Alfred Nobel. Awarded annually in the fields of chemistry, physics, physiology or medicine, literature, peace, and (since 1968) economics to persons throughout the world for specific or general achievements in their fields, they are administered and bestowed by various Swedish academies and institutes. The peace prize is under the aegis of the Norwegian parliament.

Press and broadcasting. Swedish radio and television are owned and run by a company consisting of different organizations, including those of the press. The company, licensed as a monopoly by the government, is regulated in accordance with a special radio law. Three radio networks present contrasting programming. Television operates as two independent stations, each with its own manager and supposedly in competition with each other. Commercial advertising is forbidden in both media, which are supported by annual license fees paid by owners of receivers.

The formal conditions of the Swedish press are good. The Freedom of the Press Act of 1949 regulated working conditions and outlawed all forms of censorship or administrative intervention. Only special courts can accept cases pertaining to the act. The press has a special ombudsman and an honour committee to assure that ethical norms are followed in the industry.

The greatest problems for the press are economic. Since the early 1930s the number of daily newspapers has decreased from some 230 to about 115. This decrease has meant that many areas and towns have only one paper. Measures to prevent further declines have been started by political parties, which use some of their state financing to support the press, and by the state itself, which has made special grants to second and third papers in various areas.

The political affiliations of the daily press are in almost inverse relation to the strength of the parties. The trade-union movement also has weekly magazines, copies distributed to all members—about 1,500,000—functioning largely as a Social Democratic Party newsletter.

A number of political and cultural magazines also play a role in the general debate. They probably have less importance in Sweden than corresponding magazines in many other countries, since the larger daily newspapers cover the cultural and political fields quite well.

BIBLIOGRAPHY. General materials include fact sheets and booklets on various subjects, published by the Swedish Institute, which are both informative and comprehensive; the NATIONAL CENTRAL BUREAU OF STATISTICS, *Statistical Abstract of Sweden* (annual), a comprehensive statistical reference; and the ROYAL MINISTRY FOR FOREIGN AFFAIRS AND THE SWEDISH INSTITUTE, *Sweden in Europe* (1971), articles on Sweden and on its relation to Europe.

Landscape: AXEL SOMME (ed.), *A Geography of Norden* (1968), an informative introduction to Nordic geography.

The people: EDMUND DAHLSTROM (ed.), *The Changing Roles of Men and Women* (1967), a discussion of a subject of current interest led by a major Swedish sociologist; ROBERT MURRAY, *A Brief History of the Church of Sweden*, 2nd rev. ed. (1969); ISRAEL RUONG, *The Lapps in Sweden* (1967), descriptions of an ethnic minority with interesting traditions and culture.

Economy: MARIE and CHRISTIAN NORGREN, *Industrial Sweden* (1971), a good account, with an up-to-date introduction; and MARTIN SCHNITZER, *The Economy of Sweden: A Study of the Modern Welfare State* (1970), on the Swedish economy, with American comparisons.

Administration and social conditions: MINISTRY OF EDUCATION AND CULTURAL AFFAIRS, *Adult Education in Sweden* (1970), a brief report on an important issue; NILS ANDREN, *Modern Swedish Government* (1968), a well-known work on the Swedish political system, mainly its formal construction; ALFRED BEXELIUS, *The Swedish Institution of the Justitie-Ombudsman* (1965), coverage of the Swedish ombudsman system; SWEDISH BOARD OF EDUCATION, *The Comprehensive School in Sweden* (1970), a description of the basic educational system; and PIERRE VINDE, *Swedish Government Administration* (1971), a lucid description of how the government works in practice.

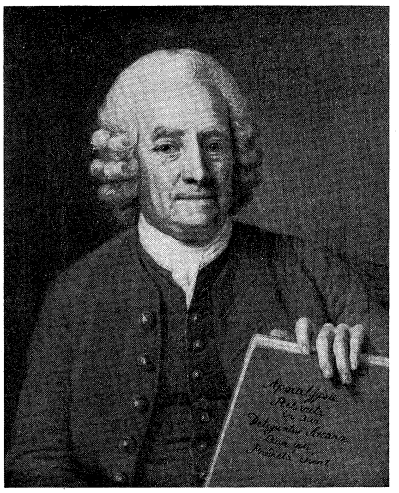
Cultural life and institutions: NIKLAS BRUNIUS, GORAN ERIKSSON, and ROLF REMBE, *Swedish Theatre* (1967), a descriptive review; LEIF FURHAMMAR and BERTIL LAURITZEN, *Film in Sweden* (1966); AKE STAVENOW and AKE H. HULDT, *Design in Sweden* (1961), on the wide range of Swedish design and handicraft; BENGT PLEIJEL (ed.), *Swedish Music: Past and Present* (1967); and ROLF SODERBERG, *Modern Swedish Art* (1963), a good introduction to Swedish 20th-century art.

Swedenborg, Emanuel

Emanuel Swedenborg, a significant Swedish scientist, mystic, philosopher, and theologian, has exerted a considerable influence on modern literature and religion, though the church (Church of the New Jerusalem) and the Swedenborgian societies based on his thought are small in membership. His scientific theories, though novel during his lifetime, were soon surpassed by the rapid developments in science in the 18th and 19th centuries.

Early life and works. Swedenborg was born on January 29, 1688, at Stockholm. His father, Jesper Swedberg, was a prominent member of the Swedish clergy, court chaplain, professor of theology in the University of Uppsala, and later bishop of Skara. When the family was ennobled in 1719, it took the name Swedenborg. In his memoirs Swedberg appears to be an egocentric of the first order, with a remarkably naïve belief in guardian angels and spirits. Emanuel Swedenborg thus received from his psychical heritage as well as his childhood environment a predisposition for the supernatural experiences that occurred in his adult life.

After graduating from the University of Uppsala in 1709, Swedenborg spent five years abroad. Early in his



Swedenborg, oil painting by Per Krafft the Elder (1724–93). In Gripsholm Castle, Sweden.
By courtesy of Svenska Portrattarkivet, Stockholm

Scientific studies and interests

education he had become fascinated by mathematics and natural science, and for that reason he first travelled to England in order to meet some of the representatives of the new natural sciences and to learn practical mechanical skills. These studies, which were continued in France and Holland, gave him an initiation in modern science that surpassed that of his compatriots. He returned to Sweden in 1715 and soon began to publish that country's first scientific journal, called *Daedalus Hyperboreus*, in which he wrote numerous reports of his projects and discoveries and of the inventions of Sweden's foremost mechanical talent of the time, Christopher Polhem. King Charles XII made the young scientist an assistant to Polhem by appointing him assessor *extra ordinem* ("extraordinary") at the Royal Board of Mines. In this capacity he worked for the next five years publishing several reports and treatises on various scientific and philosophical problems; *e.g.*, cosmology (the order of the universe), corpuscular philosophy, mathematics, and human sensory perceptions. Now and then Swedenborg also found time to write poems in Latin, some of which were published in his lifetime.

Development of a philosophy of nature. His favoured position came to an end when Charles XII died in 1718. Though he was ennobled in 1719, such practice was mere routine in the case of children of a bishop. After a second journey abroad in 1721–22, during which he published two Latin volumes on natural philosophy and chemistry, Swedenborg wrote little or nothing for more than ten years. When he set out for a third European journey in 1733, it became obvious that these years had been filled with reading and reflection in addition to his ordinary work as a civil servant. In 1734 he published in Leipzig his *Opera Philosophica et Mineralia* ("Philosophical and Logical Works") in three folio volumes, the first of which, the *Principia Rerum Naturalium* ("Principles of Natural Things"), contains Swedenborg's mature philosophy of nature. As a student at Uppsala, he had studied philosophy and physics according to the theories of René Descartes, a major French philosopher, and his years with Polhem increased the influence of the great French thinker. In fact, the three main themes of Swedenborg's philosophy of nature were inspired by Descartes: (1) matter consists of particles that are indefinitely divisible; (2) these particles are in constant vortical (swirling) motion; and (3) the Earth's planetary system sprang from the solar mass. The third theme, the nucleus of Swedenborg's cosmology, was strongly influenced by the English writer Thomas Burnet's *Telluris Theoria Sacra* (1681; "Sacred Theory of the Earth") and is a precursor of the Kant–Laplace nebular theory (*i.e.*, the Sun and planets come from a common nebula).

According to Swedenborg, the planetary system marks the end of the particle series. The beginning of the series is described with the aid of a metaphysical concept *i.e.*, the mathematical point in which the Creator introduced a conatus (an effort). This mathematical point is multiplied to ever greater powers in a universal system of series and degrees. Because the concept was an abstraction, it did not satisfy Swedenborg. For the next few years he eagerly looked for a language (symbolic or scientific) in which the relations between the Infinite and the finite creation could be expressed with mathematical exactitude.

Physiological and psychological investigations. After he published the *Principia* and a small work on the Infinite in 1734, Swedenborg returned home. His father died in 1735, and in the next year he was granted a new leave of absence from his office as assessor. This time he went to France, Italy, and Holland. In Amsterdam he completed and published a new work in two great volumes, called *Oeconomia Regni Animalis* (1740–41; *The Economy of the Animal Kingdom*, 1846), and in November 1740 he was back in Stockholm.

Psychological interests

The *Oeconomia Regni Animalis* represents a new stage in Swedenborg's scientific career. The term "animal kingdom" in Swedenborg's thought is to be understood as the human body; *i.e.*, "the kingdom of the soul (*anima*)." Here Swedenborg made a thorough study of human anatomy and physiology, with special attention to the blood and the brain. He had been interested in anatomical matters ever since his youth, but now this study had become a real passion to him. The title, however, indicates that Swedenborg was not studying the human body as a subject *per se*, but as the "kingdom of the soul." What he intended was to fulfill a program that he had formulated in 1734; *i.e.*, to prove the immortality of the soul to the senses themselves. He believed the soul to be the inmost life of the blood and located in the brain, specifically in the cellular cortex.

Swedenborg's method as a researcher consisted mainly in collecting facts from microscopists and experimentalists and drawing his own conclusions from them. His energy and his erudition were overwhelming, and he has been given credit for some contributions to the localization of the mental processes. His anatomical works, however, remained almost unnoticed by contemporary science; and when they were discovered by some 19th-century scholars, science had moved beyond them. They did, however, play a decisive role in Swedenborg's personal development.

Oeconomia Regni Animalis was followed by a whole series of sketches and small treatises in which Swedenborg attempted to fulfill his psychological investigations. Some of these became a part of the three volumes of the *Regnum Animale* (*The Animal Kingdom*, 2 vol., 1843–44)—planned to be 17 volumes—that Swedenborg himself published in 1744–45. One of the most interesting elements in this prolific and chaotic production of ideas was the problem of a universal language. Ever since the *Principia*, Swedenborg had been complaining about the lack of linguistic instruments capable of describing psychic phenomena with mathematical exactitude. In the beginning of the 1740s he made several attempts at a universal language based on the model of Descartes and Gottfried Wilhelm Leibniz, an important German philosopher of the 17th and 18th centuries, but at last he made up his mind to present a substitute for it. This substitute was the famous doctrine of correspondence, *i.e.*, Swedenborg's vision of the universe as a system of symbols permeated by divine light in different degrees, which, since the 19th century, has fascinated many poets. Its earliest systematic presentation was called *Clavis Hieroglyphica* (1784; *A Hieroglyphic Key*, 1792), in which the linguistic origin of the doctrine of correspondence becomes evident. Swedenborg wanted to demonstrate how to transform propositions from the natural sphere into its correspondence in the spiritual sphere and from there to the divine sphere by changing a few central terms in otherwise identical sentences. Later, the doctrine of correspondence was given a very wide range, becoming a kind of universal symbolism according to which everything outward and visible in nature had an inward spiritual cause. But the linguistic origin of the doctrine is reflect-

ed in its function as an instrument for revealing the spiritual meaning of the sacred Scriptures.

Religious crisis and theological concerns

The years of preparation, as Swedenborg in his old age called this period, were concluded by a painful religious crisis from which there exists a unique document. It is usually called the *Journal of Dreams* (1743–44) and was obviously meant to be a journal of his new travels beginning in July 1743; but the rather trivial notices were suddenly interrupted. There follows instead a list of various dreams recalled from earlier years and a detailed report on his spiritual experiences, mostly at night, from March to October 1744. He began to write the journal in The Hague when he prepared the second part of the *Regnum Animale*, and the journal comes to a sudden stop while he was in London, where he was writing his drama of creation, *De Cultu et Amore Dei* (1745; *On the Worship and Love of God*, 1801). His scientific and literary work determined to a great extent both his dreams and interpretations of them. Some of the dreams were of a grossly sexual character and caused many pious readers quite a shock when the journal was published in 1859. Swedenborg himself was obviously not ashamed of these dreams. His feeling of guilt was concentrated on his intellectual pride, his burning ambition to be recognized as a great man of science. On April 7, 1744, he had his first vision of Christ, which gave him a temporary rest from the temptations of his own pride and the evil spirits he believed to be around him.

A definite call to abandon worldly learning occurred in April 1745, Swedenborg told his friends in his later years. The call apparently came in the form of a vision of the Lord when he was awake. Swedenborg left his last work as a secular writer unfinished. It was his drama of creation, *De Cultu et Amore Dei*, that is the synthesis of his scientific writings. The creation of the Earth and its inhabitants, concluded by the birth of the first human couple, is the theme of this modern hexaemeron (creation poem). It is written in a style that wavers between the poetically sublime and the minute concerns of the scientist.

Theological thought. For the remainder of his long career, Swedenborg devoted his enormous energy to interpreting the Bible and to relating what he had seen and heard in the world of spirits and angels. From 1749 to 1771 he published some 30 volumes, all of them in Latin and the major part anonymously. He gave an admirably clear summary of his theological thinking in his last work, the *Vera Christiana Religio* (1771; *True Christian Religion*, 1781), which was written when he was 83.

Briefly, the *esse* ("being") of the Lord cannot be described since it is substance and form itself, but his *essentia* ("essence") can be understood in two of its primary qualities, love and wisdom. In the world of spirits that corresponds to nature, this *essentia* of the Lord is represented by the spiritual sun, the warmth of which is love and the light is *sapientia*, or "wisdom." But the Lord also has manifested himself in other forms, and foremost in his Word. During the ages, mankind has removed itself so far from its origin that the Lord allowed himself to be born into the human world through a mortal woman. Swedenborg thus rejected the doctrine of Christ as the Son of God, as well as the doctrine of the Trinity (*i.e.*, the one God revealed in the Persons of Father, Son, and Holy Spirit).

The maintenance of creation proceeds according to the doctrine of series and degrees. This is one of the many theories and doctrines that Swedenborg kept more or less unaltered in form from his scientific period. Everything in the universe belongs to a series, and in each of these there are three discrete degrees interrelated in terms of correspondence as end, cause, and effect. Each effect becomes an end in the next series at a lower level. The ultimate end of creation can be achieved only through man, and in Swedenborg's anatomist version of the *De civitate Dei* (*The City of God*) by St. Augustine, the great 4th–5th-century North African theologian, the final goal of creation is formed by innumerable angelic societies making a *maximus homo* ("greatest man").

His view of ultimate end of creation

This true order of creation, however, has been disturbed by man's misuse of his free will. He has diverted his love from God to his own ego, and thus evil has come into the world. The ultimate fate of an individual is determined in the intermediate stage immediately after the death of his body. The inner man (*i.e.*, his soul) at that time chooses a society of spirits with which he feels an inner kinship. His social position on Earth will be of no importance when his true affections are revealed in the spiritual world. Instead, new relationships, including marriages, will be established there.

On March 29, 1772, Swedenborg died in London where he was buried in the Swedish Church. At the request of the Swedish government, his body was removed to Uppsala Cathedral in 1908.

Influence. Swedenborg never acted as a preacher but rather relied totally on the effect of his huge Latin volumes. The first Swedenborgian societies appeared in the 1780s and the first independent congregation, the origin of the various Church of the New Jerusalem organizations, was founded in London by the end of the same decade. Swedenborg's influence is by no means restricted to his immediate disciples. His visions and religious ideas have been a source of inspiration for a number of prominent writers, including Honoré de Balzac, Charles Baudelaire, Ralph Waldo Emerson, William Butler Yeats, and August Strindberg. His theological writings have been translated into many languages, and there is a constant flow of new editions.

BIBLIOGRAPHY. The greater part of Swedenborg's manuscripts, preserved in the library of the Royal Swedish Academy of Science in Stockholm, were lithographically published in 10 vol. in 1869–70, and in an 18-volume edition in 1901–16. Unfortunately, these editions cannot stand up to the demands of modern scholarship. Most of the biographical sources were collected, translated, and annotated by R.L. TAFEL in *Documents Concerning the Life and Character of Emanuel Swedenborg*, 2 vol. (1875–77). His letters have been published in English translation in A. ACTON (ed.), *Letters and Memorials*, 2 vol. (1948–55).

For supplementary material, see E.J. DINGWALL, *Very Peculiar People: Portrait Studies in the Queer, the Abnormal, and the Uncanny*, pp. 11–68 (1962); J. HYDE, *A Bibliography of the Works of Emanuel Swedenborg* (1906); I. JONSSON, *Emanuel Swedenborg* (Eng. trans. 1971); M. LAMM, *Swedenborg* (1915), in Swedish; CYRIEL O. SIGSTEDT, *The Swedenborg Epic* (1952); and SIGNE TOKSVIG, *Emanuel Swedenborg, Scientist and Mystic* (1948).

(I.Jo.)

Swift, Jonathan

Generally acknowledged as the foremost satirist in the English language and one of the satiric masters of all time, Jonathan Swift was many other things as well: a wit, a social personality, a clergyman and churchman, an eminent public figure. He wrote prose satires widely different in theme and execution; he was a distinguished minor poet in an age of great minor poetry; and he was a political journalist of unsurpassed power. His genius and achievement as a satirist have always been recognized, but in other respects he has been a controversial figure—sometimes provoking severe judgments on his personality, various aspects of his career, and on the intent and effect of his satiric writings, especially his most famous work, *Gulliver's Travels*. There are episodes in his private life that, for lack of full and certain factual information, cannot be wholly understood; but these fascinate all the more for that reason and have given rise to endless conjecture—often of a sensational nature and frequently to his discredit. Even quite generally accepted facts about his birth and parentage have, from time to time, been called in question by persons persuaded that the true circumstances were concealed from the world.

Early life and education. Swift was born on November 30, 1667. His grandfather, the Rev. Thomas Swift, had been vicar of Goodrich, Herefordshire, and was remembered for his unswerving loyalty to the Royalist cause during the English Civil War (1642–51). Swift's father, Jonathan the elder, along with three of his brothers, settled in Ireland after the Restoration and became steward of the King's Inns, Dublin. In 1664 he married Abigail Erick, who, though she may have been



Swift, oil painting by C. Jervas (1675?–1739). In the National Portrait Gallery, London.
By courtesy of the National Portrait Gallery, London

born in Ireland, was the daughter of an English Leicestershire clergyman. Their daughter, Jane, was born in May 1666. In the spring of 1667, however, Jonathan the elder died suddenly, leaving his wife, baby daughter, and an unborn son to the care of his brothers.

Unsettled youth

As a result of growing up fatherless, dependent on the generosity of his uncles, with no settled home, it may be that Jonathan Swift suffered from a sense of insecurity. His education was not neglected, however, and at the age of six he was sent to Kilkenny School, then the best in Ireland. In 1682 he entered Trinity College, Dublin. In a fragment of autobiography, written many years later and first published in 1755 as "Family of Swift" (edited by H. Davis, *Miscellaneous and Autobiographical Pieces*, 1962), he represents himself as so discouraged and sunk in spirit by "the ill-treatment of his nearest relatives" that he neglected his studies and was therefore only granted his bachelor of arts degree in February 1686 *speciali gratia* ("by special favour"). In fact, his career as an undergraduate seems to have been neither distinguished nor disgraceful; the manner of his degree being a device often used when a student's record failed, in some minor respect, to conform to the regulations, it was not the "discreditable mark" against his name that he himself called it.

He continued in residence at Trinity College as a candidate for his master of arts degree until February 1689. But the disorders that had begun to spread through Dublin after the Revolution of 1688 caused the college authorities to permit students to withdraw to safety. Swift sought security in England, and, after visiting his mother (who had earlier returned to Leicestershire), he became a member of the household of Sir William Temple, at Moor Park, Surrey, where he was to remain off and on until Temple's death in January 1699.

Years at Moor Park. Temple, a man of wide culture, was engaged in writing his memoirs and preparing some of his essays for publication, and he needed someone to act as a kind of secretary. Swift's residence at Moor Park was an interrupted one; he twice returned to Ireland (from May 1690 to December 1691 and from May 1694 to May 1696), a fact that points to some uneasiness about his career. During the second of these visits to Ireland, Swift took orders in the Anglican Church, being ordained priest on January 13, 1695. At the end of the same month he was appointed vicar of Kilroot, near Belfast.

It may be said that Swift came to intellectual maturity at Moor Park, with Temple's rich library at his disposal. Here, too, he met Esther Johnson (the future Stella), the daughter of Temple's widowed housekeeper. She was a dark-haired child of eight when Swift first came to Moor Park, and his friendship with her began by his taking a share in her education. In 1692, through Temple's good offices, he received the degree of M.A. at Oxford, after incorporation as a member of Hart Hall. It was probably in 1693 that he went to Kensington on a mission from Temple and there had a conference with William III, who had asked Temple's advice on a constitutional problem.

Between 1691 and 1694 he wrote a number of poems, notably six odes. His early verse, which at one time provoked only the harshest of judgments, has come to be much more favourably regarded, perhaps because of William Butler Yeats's expressed admiration. Clearly, however, his true genius did not find expression until he turned from verse to prose satire and composed, mostly at Moor Park between 1696 and 1699, *A Tale of a Tub*, one of his major works. Published anonymously in 1704, this was made up of three associated pieces: the *Tale* itself, a satire against "the numerous and gross corruptions in religion and learning"; the mock-heroic "Battle of the Books"; and the "Discourse Concerning the Mechanical Operation of the Spirit," which ridiculed the manner of worship and preaching of religious enthusiasts at that period.

Early satire

In the "Battle," Swift displayed his loyalty to Temple, who, through an essay entitled "Upon the Ancient and Modern Learning" (1690), had taken part in a long-standing dispute about the relative merits of ancient and modern learning. The essay supported the cause of the ancients and was severely criticized by William Wotton, a defender of contemporary culture, and by Richard Bentley, a distinguished scholar. Swift allegorized this attack on Temple, ridiculing Wotton and Bentley; in the famous episode of the bee and the spider, with which the symbolic battle opens, he expresses his own and Temple's tenets regarding the right and universal standards in taste and literature as first set forth by the ancients: the spider feeds wholly on itself, producing nothing at last but flybane and a cobweb, whereas the bee in its wide flights visits all the flowers of the field and brings home honey and wax, "thus furnishing mankind with the two noblest of things, which are sweetness and light." But the *Tale* itself is the most impressive of these three compositions, outstanding for its exuberance of satiric wit and energy and marked by an incomparable command of stylistic effects, largely in the nature of parody. Swift saw the realm of culture and literature threatened by zealous pedantry, while religion—which for him meant rational Anglicanism—suffered attack from both Roman Catholicism and Dissent. He proceeded to trace all these dangers to a single source: the irrationalities that disturb man's highest faculties—reason and common sense.

Career as satirist, political journalist, and churchman. With Temple's sudden death in January 1699, Swift entered on a period of varied experience and considerable uncertainty. In the summer he returned to Dublin as chaplain and secretary to the Earl of Berkeley, who was then going to Ireland as a lord justice. During the immediately ensuing years he was in England on some four occasions—in 1701, 1702, 1703, and 1707–09—and won wide recognition in London for his personal charm and his wit as a writer. Otherwise, his position remained unchanged. He had resigned his position as vicar of Kilroot while still at Moor Park, but early in 1700 he was preferred to several posts in the Irish church, of which the most important was the vicarage of Laracor, not far from Dublin. His public writings of this period show that he kept in close touch with affairs both in Ireland and England. Among numerous compositions, the most notable are the "Discourse of the Contests and Dissensions between the Nobles and the Commons in Athens and Rome" (a forcible statement of Whig principles and an appeal to both political parties for moderation); "A Project for the Advancement of Religion, and the Reformation of Manners"; and "The Sentiments of a Church of England Man with Respect to Religion and Government." In London he was increasingly well-known. His religious and political essays, serious and ironic; the publication in 1704 of *A Tale of a Tub*, anonymous but generally attributed to him; certain impish works, including the well-known "Bickerstaff" pamphlets of 1708–09, which put an end to the career of John Partridge, a popular astrologer, by first prophesying his death and then describing

it in circumstantial detail; and several amusing pieces in a minor key, such as the short poem "A Description of the Morning," printed by Richard Steele in *The Tatler* in 1709—these brought him to the attention of the general public as well as to a circle of Whig writers led by Joseph Addison. It is apparent, however, that Swift was uneasy about many policies of the Whig administration. He was, on the one hand, a Whig by birth and education, and the political principles that he had been setting forth in his writings were in the 17th-century Whig tradition. But, on the other, he was also passionately loyal to the established church, and he came to view with apprehension the Whigs' growing determination to yield ground to the Dissenters. He was increasingly aware of this conflict of loyalties.

Political alliance with the Tory party

Another momentous period began for him when, on September 7, 1710, he once again found himself in London. A Tory ministry headed by Robert Harley (later earl of Oxford) and Henry St. John (later Viscount Bolingbroke) was replacing that of the Whigs. The new administration, bent on bringing hostilities with France to a conclusion, was also assuming a more protective attitude toward the Church of England. Swift's reactions to such a rapidly changing world are vividly recorded in his famous *Journal to Stella*, a series of letters written between his arrival in England in 1710 and 1713, which he addressed to Esther Johnson and her companion, Rebecca Dingley, who were now living in Dublin. The astute Harley had already befriended the satirical poet and pamphleteer Daniel Defoe, thereby gaining the valuable political services of this writer. Now he made overtures to Swift and, by assuring him that the grant of Queen Anne's Bounty would be extended to the church in Ireland, won him over to the Tories. But though Swift was thus placing church above party, he was not renouncing his deeply fixed and essentially Whigish convictions regarding the nature of government. The old Tory theory of the divine right of kings had no claim upon him. The ultimate power, he insisted, derived from the people as a whole and, in the English constitution, had come to be exercised jointly by king, lords, and commons, between whom it was most important that a balance be maintained lest any form of tyranny arise.

Swift quickly became the Tories' chief political writer and, by the end of October 1710, had taken over the Tory journal, *The Examiner,* which he continued to edit until June 14, 1711. He then began preparing a pamphlet in support of the Tory drive for peace. This, "The Conduct of the Allies," appeared on November 27, 1711, some weeks before the motion in favour of a peace was finally carried in Parliament. His services were not underrated by his friends in the ministry, but his reward was slow to come. When it did—in April 1713—it proved to be the deanery of St. Patrick's Cathedral in Dublin, not one of the English deaneries then vacant. The Queen and many persons of influence are believed to have regarded *A Tale of a Tub* as irreligious, and it may be that they were instrumental in blocking Swift's way to preferment in England. In any event, English deaneries and bishoprics were not being bestowed on those who were, like Swift, of Anglo-Irish birth.

Dean of St. Patrick's, Dublin

Withdrawal to Ireland. But with the death of Queen Anne in August 1714 and the accession of George I, the Tories were a ruined party and Swift's career in England was at an end. He withdrew to Ireland, where, except for two later visits to England in 1726 and 1727, he was to pass the remainder of his life. At first he was the object of scorn and abuse on the part of the Anglo-Irish Whigs, and, bitter over the turn of events and emotionally exhausted, he retired to the seclusion of his deanery. Only gradually did he regain his full energies, but by 1720 he was showing a renewed interest in public affairs and was also returning to verse, which he continued to write through the 1720s and early 1730s. In his Irish pamphlets of this period he came to grips with many of the problems, social and economic, then confronting Ireland. His tone and manner varied from direct factual presentation to exhortation, humour, and bitter irony. He was moved by his perception of the realities of the Irish situation and by his desire to strike at the all-powerful Sir Robert Walpole, and his Whig administration. Ireland's backward state he chiefly blamed on the blindness of the English government; but he also insistently called attention to the things the Irish themselves might do in order to better their lot. Of the Irish writings, the "Drapier's Letters" (1724–25) and "A Modest Proposal" are the best known. The first is a series of letters, purportedly written by M.B., a Dublin linendraper, attacking the English government for its scheme to supply Ireland with copper halfpence and farthings; the latter is a grimly ironic letter of advice in which a public-spirited citizen suggests that economic conditions could be alleviated if the children of poor parents were used as food for the rich.

Mention should be made of certain events in Swift's private life and the theorizing to which they have given rise. The bare events are these. Stella continued to live with Rebecca Dingley after moving to Ireland in 1700 or 1701. Another woman, Esther Vanhomrigh—named Vanessa by Swift—appeared in his life during his London sojourn of 1707–09; in 1714, despite Swift's admonitions, she insisted on following him to Ireland. Her letters to Swift reveal her passion for him (though before her death in June 1723, she apparently turned against him, for he is not mentioned in her will). What lies behind these facts? Were Stella and Swift secretly married? According to one story, long in circulation, there was a marriage, though in name only, in 1716. One version makes Stella Sir William Temple's illegitimate daughter, for which reason Swift was reluctant to marry her openly. Sometimes it is said that Stella, deeply disturbed by Vanessa's presence in Ireland, finally insisted on a marriage. A more sensational variant makes Stella Swift's half sister, each fathered by Temple—a fact obligingly revealed to them immediately after their marriage. And what was the nature of the relationship between Swift and Vanessa? Did Swift remain merely her friend, or did he become her lover? One report affirms that they were the parents of a boy who—add others—was in time taken in and cared for by the compassionate Stella. Denis Johnston's *In Search of Swift* (1959) effects the ultimate turn of the screw: Swift's father was not Sir William Temple but Sir William's father, thus making Swift and Stella uncle and niece; there was no marriage between them, but Swift and Vanessa were more than friends. The point must be made that there is almost no evidence likely to stand up in a court of law to support any of these theories. The actual events, clear to the participants, may very well have been more commonplace.

Stella died in January 1728. Swift's greatest satire, *Gulliver's Travels,* had already been published, the original edition having appeared late in October 1726. It is uncertain when Swift began this work, but it appears from his correspondence that he was writing in earnest by 1721 and had finished the whole by August 1725. He had not been to England since 1714, but in 1726 he returned, carrying with him in manuscript the work—designed, as he said in a letter (September 29, 1725) to Alexander Pope, to "vex the world rather than divert it." He visited Pope at Twickenham, where other friends of former days joined them and where he completed arrangements for publication. *Gulliver's Travels* appeared on October 28, after its author had left again for Dublin. Its success was immediate. Then, and since, it succeeded in "vexing" (but also diverting) all classes of readers. To what extent is it a political allegory, hinting at the story of the Tory ministers Oxford and Bolingbroke, their fall, and Walpole's policies? Is it essentially comic, or is it a misanthropic depreciation of mankind? How seriously is one meant to take Lemuel Gulliver's utopian sketches? Its matter-of-fact style and its air of sober reality confer on it an ironic depth that defeats oversimple explanations.

Publication of *Gulliver's Travels*

Last years. The closing years of Swift's life have been the subject of some misrepresentation, and stories have been told of his ungovernable temper and lack of self-control. It has been suggested that he was insane. From youth he had suffered from what is now known to have been Ménière's disease, an affliction of the semicircular

canals of the ears, causing periods of dizziness and nausea. But his mental powers were in no way affected, and he remained active throughout most of the 1730s—Dublin's foremost citizen and Ireland's great patriot dean. In the autumn of 1739 a great celebration was held in his honour. He had now, however, begun to fail physically and later suffered a paralytic stroke, with subsequent aphasia. In 1742 he was declared incapable of caring for himself, and guardians were appointed. He died on October 19, 1745. He was buried in St. Patrick's Cathedral, and on the wall was affixed a Latin epitaph that he had himself composed and that may be translated thus:

> The body of Jonathan Swift, Doctor of Sacred Theology, dean of this cathedral church, is buried here, where fierce indignation can no more lacerate his heart. Go, traveler, and imitate, if you can, one who strove with all his strength to champion liberty.

MAJOR WORKS

POLITICAL WRITINGS: "A Discourse of the Contests and Dissensions between the Nobles and the Commons in Athens and Rome" (1701); "The Conduct of the Allies, and of the Late Ministry, in Beginning and Carrying on the Present War" (1711); "Some Remarks on the Barrier Treaty, between Her Majesty and the States-General" (1712); "The Importance of the Guardian Considered" (1713); "The Publick Spirit of the Whigs" (1714).

POLITICO-RELIGIOUS WRITINGS: "A Project for the Advancement of Religion, and the Reformation of Manners" (1709); "A Letter . . . Concerning the Sacramental Test" (1709); "Argument Against Abolishing Christianity" (1711); "The Sentiments of a Church of England Man with Respect to Religion and Government" (1711); "A Letter to a Young Gentleman, Lately Enter'd into Holy Orders" (1721).

ON IRISH QUESTIONS: "A Proposal For the Universal Use of Irish Manufacture" (1720); "Drapier's Letters" (1724–35; concluding with "An Humble Address to both Houses of Parliament," 1735); "A Short View of the State of Ireland" (1728); "A Modest Proposal For preventing the Children of Poor People From being a Burthen to their Parents, or the Country, and For making them Beneficial to the Publick" (1729); "An Examination of Certain Abuses, Corruptions, and Enormities in the City of Dublin" (1732).

OTHER: *A Tale of a Tub . . .* (1704), including the *Tale*, "The Battle of the Books," and "Discourse Concerning the Mechanical Operation of the Spirit"; Bickerstaff pamphlets (1708–09); *Journal to Stella* (letters 1 and 41–65 published 1766; 2–40, 1768); *Travels into Several Remote Nations of the World . . . By Lemuel Gulliver* (1726, known as *Gulliver's Travels*).

VERSE: "A Description of the Morning" (1709); "A Description of a City Shower" (1710); "Cadenus and Vanessa" (1726); "The Grand Question Debated" (1729); "On Poetry: A Rapsody" (1733); "Verses on the Death of Dr. Swift" (1739).

BIBLIOGRAPHY. The Swift bibliography in *The New Cambridge Bibliography of English Literature*, vol. 2 (1971), covers both Swift's writings and Swift criticism through 1969. For Swift's works, see also H. TEERINK, *A Bibliography of the Writings of Jonathan Swift*, 2nd ed. rev. and corrected by A.H. SCOUTEN (1963). For Swift criticism, see L.A. LANDA and J.E. TOBIN, *Jonathan Swift: A List of Critical Studies Published from 1895–1945* (1945); and J.J. STATHIS, *A Bibliography of Swift Studies, 1945–1965* (1967).

Major collected editions, editions of separate works, and selections: *The Prose Works*, ed. by H. DAVIS, 14 vol. (1941–68), is the standard edition, though still useful is the older *Prose Works*, ed. by T. SCOTT, 12 vol. (1897–1908). *The Poems of Jonathan Swift*, ed. by H. WILLIAMS, 2nd ed., 3 vol. (1958), is of first importance. The useful one-volume *Poetical Works*, ed. by H. DAVIS (1967), is based on the Williams edition. H. WILLIAMS' editions of the *Correspondence*, 5 vol. (1963–65), and of the *Journal to Stella*, 2 vol. (1948), have superseded all earlier editions. Modern editions of separate works include *A Tale of a Tub . . .*, ed. by A.C. GUTHKELCH and D. NICHOL SMITH, 2nd ed. (1958), which is notable for its scholarly introduction and notes; *Polite Conversation*, ed. by E. PARTRIDGE (1963), with most interesting commentary; and *A Discourse of the Contests and Dissentions Between the Nobles and Commons in Athens and Rome*, ed. by F.H. ELLIS (1967), valuable both for its textual scholarship and for its critical and historical notes. Selections include *Swift: Gulliver's Travels and Selected Writings in Prose and Verse*, ed. by J. HAYWARD (1934); *The Portable Swift*, ed. by C. VAN DOREN (1948); *Gulliver's Travels and Other Writings by Jonathan Swift*, ed. by R. QUINTANA (1958); *Jonathan Swift: Selected Prose and Poetry*, ed. by E. ROSENHEIM, JR. (1959); *Jonathan Swift: Gulliver's Travels and Other Writings*, ed. by L.A. LANDA (1960); *Jonathan Swift: Gulliver's Travels and Other Writings*, ed. by M.K. STARKMAN (1962); *Poetry and Prose by Jonathan Swift*, ed. by H. DAVIS (1964); and *Selected Poems of Jonathan Swift*, ed. by J. REEVES (1967).

Biography and criticism: H. CRAIK, *The Life of Jonathan Swift*, 2nd ed., 2 vol. (1894), for many years the standard biography, is still useful. Of more recent biographies, the most detailed and in many respects the best is I. EHRENPREIS, *Swift: The Man, His Works, and the Age*, vol. 1 (1962), vol. 2 (1967), vol. 3 (in prep.). There is also J.M. MURRY, *Jonathan Swift: A Critical Biography* (1955). *In Search of Swift* by D. JOHNSTON (1959), professing to reveal the nature of the relationship between Swift and Esther Johnson, has proved convincing to some but has been repudiated, either openly or silently, by the majority of Swift scholars and critics. S. LE BROCQUY, *Cadenus: A Reassessment in the Light of New Evidence of the Relationships Between Swift, Stella and Vanessa* (1962) and *Swift's Most Valuable Friend* (1968), the latter concerned principally with Esther Johnson, and both heavily influenced by Johnston's speculations, have won little assent. Swift's personality, aspects of his daily life, and some of the particulars of his career as a public writer are treated in the following: R.W. JACKSON, *Jonathan Swift, Dean and Pastor* (1939) and *Swift and His Circle* (1945); L.A. LANDA, *Swift and the Church of Ireland* (1954); M. FOOT, *The Pen and the Sword* (1957), concerning Swift and the Duke of Marlborough as public antagonists; I. EHRENPREIS, *The Personality of Jonathan Swift* (1958); B.A. GOLDGAR, *The Curse of Party: Swift's Relations with Addison and Steele* (1961); O.W. FERGUSON, *Jonathan Swift As a Tory Pamphleteer* (1967). Reactions to Swift by his contemporaries and by later writers are the subject of D.M. BERWICK, *The Reputation of Jonathan Swift, 1781–1882* (1941); and K. WILLIAMS, *Swift: The Critical Heritage* (1970), the latter reaching to 1819. Modern criticism of Swift has been extensive. Some studies have a broad, general approach to Swift and his writings: R. QUINTANA, *The Mind and Art of Jonathan Swift* (1936; reprinted with additional notes and bibliography, 1953) and *Swift: An Introduction* (1955); K. WILLIAMS, *Jonathan Swift and the Age of Compromise* (1958); D. DONOGHUE, *Jonathan Swift: A Critical Introduction* (1969). More specialized studies, scholarly and critical, include M.O. JOHNSON, *The Sin of Wit: Jonathan Swift As a Poet* (1950); M.K. STARKMAN, *Swift's Satire on Learning in A Tale of a Tub* (1950); J.M. BULLITT, *Jonathan Swift and the Anatomy of Satire: A Study of Satiric Techniques* (1953); M. PRICE, *Swift's Rhetorical Art: A Study in Structure and Meaning* (1953); W.B. EWALD, *The Masks of Jonathan Swift* (1954); J.P. HART, *Swift and Anglican Rationalism: The Religious Background of A Tale of a Tub* (1961); E.W. ROSENHEIM, *Swift and the Satirist's Art* (1963); H. DAVIS, *Jonathan Swift: Essays on His Satire and Other Studies* (1964); and G.P. MAYHEW, *Rage or Raillery: The Swift Manuscripts at the Huntington Library* (1967). M. VOIGT has reviewed the body of modern scholarship and criticism in his *Swift and the Twentieth Century* (1964). The Swift tercentenary, observed in 1967, called forth several collections of essays both old and newly composed: *Jonathan Swift, 1667–1967: A Dublin Tercentenary Tribute*, ed. by R. MCHUGH and P. EDWARDS (1967); *Fair Liberty Was All His Cry: A Tercentenary Tribute to Jonathan Swift, 1667–1745*, ed. by A.N. JEFFARES (1967); *Swift Revisited*, ed. by D. DONOGHUE (1968), consisting of lectures broadcast in Ireland; *The World of Jonathan Swift: Essays for the Tercentenary*, collected and ed. by B. VICKERS (1968). Also to be noted are several studies more closely analytical and focussed upon Swift's stylistic characteristics: R. PAULSON, *Theme and Structure in Swift's Tale of a Tub* (1960); L.T. MILIC, *A Quantitative Approach to the Style of Jonathan Swift* (1967); W.B. CARNOCHAN, *Lemuel Gulliver's Mirror for Man* (1968); A. ROSS, *Swift: Gulliver's Travels* (1968); KATHLEEN WILLIAMS, *Jonathan Swift* (1968), consisting of extracts from Swift's work accompanied by commentary and analysis; and J.R. CLARK, *Form and Frenzy in Swift's Tale of a Tub* (1970).

(R.Q.)

Swimming and Aquatic Sports

Swimming is the action of self-support and propulsion in the water. As a pastime it has few equals, can be enjoyed year round either outdoors or indoors during cold weather, and is reputed by physical educators to be the best exercise for the symmetrical development of the whole body. It is enjoyed by millions and is often a means of protecting and saving lives.

This article is intended for the general reader who may have no knowledge of the activities described. The article traces the history and development of swimming and covers methods of instruction, basic strokes, and growth of competition, along with sections on diving, channel and open water (marathon) swimming, water polo, water skiing, surf riding (surfing), and skin diving. For information on specific rules of competition or finer points of performance the reader should consult the works listed in the bibliography.

SWIMMING

History. The ability to swim was valued in many ancient civilizations, including Egypt, Assyria, Greece, and Rome, where it was practiced widely and stressed in the training of warriors. During the Middle Ages, however, swimming fell into almost complete disuse throughout Europe because of the general belief that outdoor bathing helped to spread the epidemics that so often swept across the continent.

Not until the second half of the 19th century was this medieval prejudice overcome sufficiently for aquatic activities to be resumed. Competition was revived, first in Great Britain, then elsewhere. Swimming was one of the sports included in the first modern Olympic Games at Athens in 1896. Diving events were added in 1904, and women's swimming and diving in 1912.

There have been cases where beginners have been able to swim upon first entering the water, but generally speaking swimming is an acquired skill. With the increased acceptance of swimming as recreation and healthful physical activity, it has grown by leaps and bounds. More and more swimming pools or baths (Great Britain) are found in schools, colleges, communities, and private homes and clubs throughout the world. Teaching of swimming has long been part of the curriculum in many schools. Learn-to-swim campaigns have been fostered by the Red Cross and by national athletic and youth organizations in most countries. During World War II almost all nations included swimming instruction as part of the training of all military personnel. After the war, the gains made in swimming instruction and promotion were consolidated further by national organizations, such as the United States Council for National Cooperation in Aquatics, which is made up of several different organizations and meets annually to further the cause of swimming in all phases.

Instruction. Many methods of instruction have been tried, some more successfully than others. In the main there have been two predominant schools: the American, whose coaches insist that the crawl should be taught first; and the European, where emphasis has been on teaching the breaststroke. In the past twenty years, however, teachers in most countries have stressed learning the crawl first. With children being taught at ever younger ages, learning to swim presents few problems.

The greatest handicaps in learning are failure to relax and haste in performing swimming movements. Fear of the water, experienced by almost all nonswimmers, is overcome to a large degree by playing games in shallow water, ducking the head and looking for objects on the bottom, and learning the fundamental skills of kicking, arm stroke, and floating. Recent innovations aiding the nonswimmer to learn more quickly are swim fins, rubber fins attached to the feet like shoes to help the beginner learn proper kicks; the snorkel, a breathing tube that helps the beginner learn stroke technique without worrying about the breathing cycle; and swim buoys, egg-shaped floats that are attached to the torso by means of an adjustable belt and supply the buoyancy needed to keep the nonswimmer afloat while concentrating on the swimming stroke being taught.

During World War II, when great numbers of men had to be taught and speed in training was necessary, a method of starting a pupil on his back was developed by U.S. Navy personnel. In this position floating was facilitated, breathing made easier, thus developing confidence in the water more quickly. A very simple propulsive stroke was taught, a frog kick synchronized with a short arm stroke, the arms remaining underwater throughout.

Another simplified approach called drown proofing emphasizes breathing and relaxation. This method was adopted by the U.S. Marine Corps for its survival-training program. Eschewing the safer techniques of learning in shallow water, beginners are introduced directly into deep water. By the use of natural buoyancy, breath control, and a minimum of effort, beginners are able to stay afloat as well as propel themselves slowly in the water. Drown proofers with little instruction can stay afloat for hours.

Drown-proofing approach

Swimming strokes. The most popular strokes or styles of swimming are the crawl (freestyle), sidestroke, backstroke, and butterfly. In the early years sidestroke and breaststroke were most frequently used, but the thirst for speed in swimming led to the development of the faster, more efficient crawl. The older strokes remain, however, the basic strokes in most lifesaving courses.

In early competition the most effective racing stroke was sidestroke, originally swum with both arms submerged. The first advance was made when swimmers conceived the idea of minimizing water resistance by recovering (bringing forward) one arm above the water. This was called the single overarm. The double overarm soon followed, with each arm recovering above the surface and the body rolling from side to side on its long axis. This stroke was made famous in the 1890s by an Englishman, J. Arthur Trudgen, who told of learning it in South Africa. The stroke took his name and eventually became the basic action of the crawl.

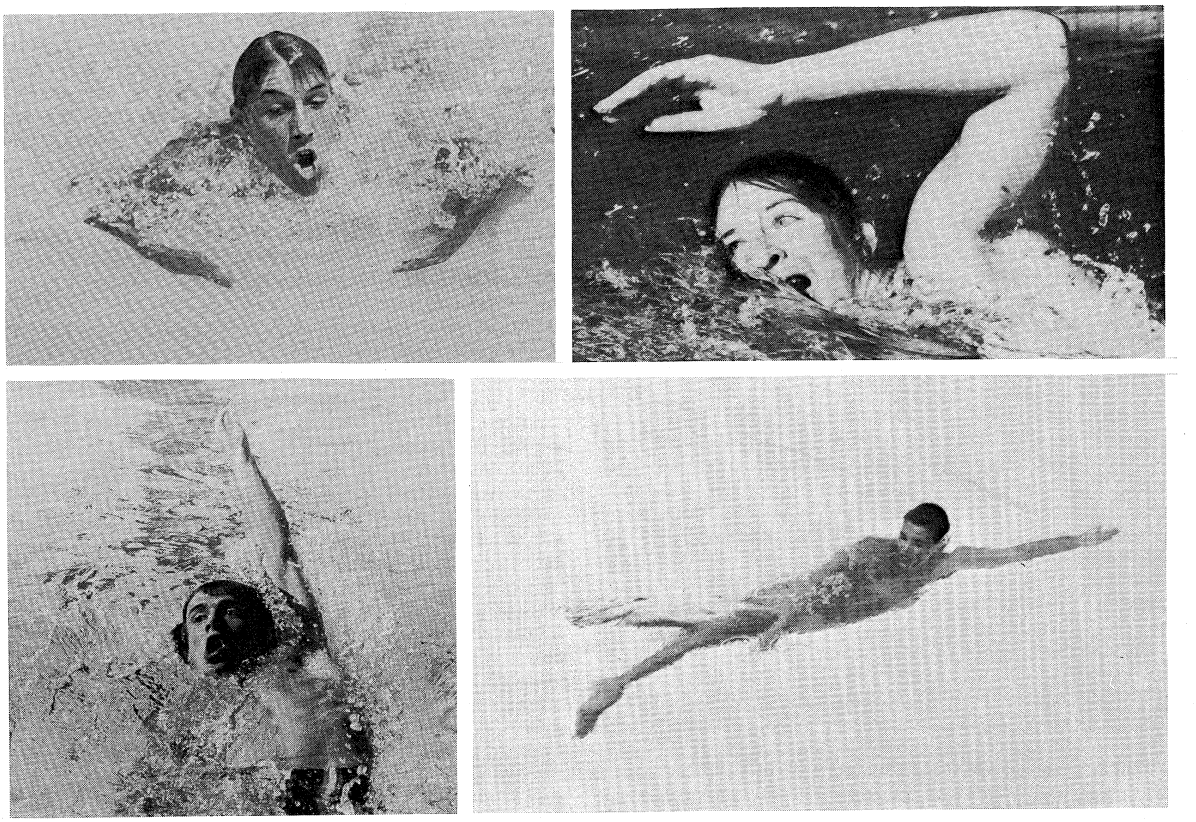
Sidestroke. The sidestroke is valued today chiefly for lifesaving purposes, since it is the most efficient of all in rescue situations. It is also preferred by the average untrained person for open-water pleasure swimming, since it is faster and more easily learned than breaststroke, and the face remains at or near the surface at all times for ease in breathing. The body stays on its side, as the name implies, and the arms propel alternately. The lower arm starts from a comfortably extended position overhead, pulls straight down below the body to waist level, then bends upward and extends again. The topmost arm pushes somewhat outside the body line with a sweeping movement from a point slightly beyond the head all the way to the thigh and recovers either over or under the surface. The kick used in sidestroke is called the scissors kick. It is performed by opening the legs slowly, the underleg moving backward and the upper leg forward, both knees bent a little and toes pointed. The scissoring action of the legs coming smartly together after opening creates the forward propulsion of the kick.

Scissors kick

Breaststroke. The breaststroke is believed to be the oldest of strokes and is of great practical value in lifesaving and recreational swimming, as well as being ideal for swimming in rough water. The French writer M. Thevenot in his book *Art de Nager* (Art of Swimming) of 1696 described the stroke as a wide lateral pull of the arms combined with a symmetrical action of the legs, simulating the swimming movements of a frog. This stroke is now known as the orthodox breaststroke to differentiate it from the newer butterfly breaststroke, or butterfly, a racing adaptation introduced by a German swimmer, Erich Rademacher, in 1926 and later popularized by the American swimmer Henry Myers in 1933.

The breaststroke and the butterfly stroke are both performed lying face down in the water and are primarily distinguished by differing arm actions. In butterfly the arms are brought forward above the surface, whereas in the orthodox breaststroke the arms always remain under water. In breaststroke the two hands are extended together and start their pull about six to eight inches under the surface, with the hands and arms pressing backward and to the sides in an arclike pattern of about 100 degrees. When they reach their maximum spread, the hands are brought together under the breast, completing the arc, and are then extended again quickly, palms down, to start another stroke.

The leg action for the breaststroke, known generally as the frog kick, may be of two types—wedge kick or whip kick. The older wedge is characterized by a very wide



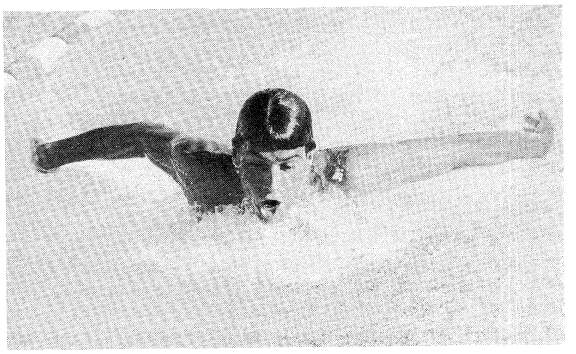
Four basic swimming strokes.
(Top left) the breaststroke; (top right) freestyle or crawl; (bottom left); the backstroke; (bottom right) the sidestroke.
(Top left) Swimming World Publications, (top right) Wide World Photos, (bottom left) Don Wilkinson, (bottom right) Russ Kinne

Wedge and whip kicks

action in which the feet are drawn up, knees bent, and the kick produced by a vigorous rounded and outward sweep of the legs in a fully extended position, until the feet come together again. The wedge kick is often described as "squeezing" the water. The more efficient whip kick features a much narrower action. The knees are kept much closer together, yet the feet reach out to engage the water by rotation of the upper thigh.

Formerly, breaststroke swimming featured a momentary glide at the completion of the leg thrust. Modern techniques eliminate the glide entirely, aiming to maintain continuous propulsion. This results in keeping the body high in the water, reducing resistance, and helping to produce more speed. Breathing techniques also have changed. In the old breaststroke, inhalation came at the beginning of the arm stroke. Inhalation now comes at or near the end of the arm pull, just before the hands come together under the shoulders.

Don Wilkinson



Butterfly stroke.

Butterfly. The butterfly stroke is a refinement of the orthodox breaststroke. It was brought to the attention of the United States officials in 1933 in a race that included Henry Myers. He insisted that his method of swimming breaststroke (bringing the arms forward over the water) conformed to the rules of breaststroke as then written. This precipitated much discussion and consternation in the U.S. and abroad, the eventual result being international recognition of a new competitive stroke called breaststroke butterfly in 1953. In its short history it has become, next to the crawl, the fastest competitive stroke and can be considered a double freestyle or crawl. It is definitely a crawl-like stroke in that the arm stroke contributes most of the forward motion and each arm executes a pattern similar to the crawl arm action. Instead of pulling alternately with the arms as in crawl, however, the butterfly swimmer pulls and then recovers both arms simultaneously.

When the butterfly stroke was first swum the frog kick was used. The effect of combining this kick with the simultaneous movement of the arms was a series of jerky forward movements. This was eliminated as the up and down movement of the fishtail (dolphin) kick was perfected, mainly through trial and error by the swimmers themselves. Present butterfly-stroke rules demand up and down movement of the legs only, whereas breaststroke rules still require the rounded outward sweep of the legs in the lateral plane.

Fishtail (dolphin) kick

In the early stages of development of the fishtail (dolphin) kick, excessive undulation of the body was thought necessary to provide speed. This was replaced by a kick that reduces undulations of the torso to necessary reactions to the movement of the arms and legs in certain phases of the stroke. The kick is actually a double crawl kick. Most swimmers execute two dolphin kicks to one arm pull.

In breathing, inhalation usually takes place when the arms are in the final stages of the propulsive action. Exhalations come immediately afterward during the overwater recovery of the arms and the initial part of the down stroke. In sprint competitions (100 metres or less) it is common to inhale every second or third stroke to develop more speed. Lifting the head too often and too high when breathing creates unnecessary drag.

Backstroke. There is indication in the 1903 edition of *Swimming* by Archibald Sinclair and Willam Henry that a backstroke similar to the trudgen, and therefore similar to modern backstroke, was then in use. The style of backstroke then in vogue, however, was aptly described as inverted breaststroke, since it consisted of a simultaneous arm action combined with a frog kick performed while swimming on the back. To U.S. Olympic champion (1912) Harry Hebner goes the major credit for revolutionizing the stroke that first became known as back crawl. The stroke featured alternate arm pulls, as in the crawl stroke, and was so revolutionary that Hebner was almost disqualified for swimming improperly. Intervention by American officials citing the rule that required the swimmer to remain on his back throughout the race preserved his Olympic victory. The new back crawl, now called backstroke, was much faster and far more efficient and drove the old stroke into discard. Since then the stroke has undergone subtle changes, all for the better, if improvement in race times is the criterion.

In the backstroke the body position remains as flat and streamlined as possible. A pronounced bend at the hips creates a "sitting" effect, a common error that produces too much resistance to make the stroke effective. The arms reach alternately above the head and enter the water at a point directly in line with the shoulder, palm facing outward and little finger entering the water first. After the hand enters the water the arm is pulled to the thigh. The first third of the pulling action is done with a straight arm, the middle third with a flexed arm, and the last third with a final thrust or push to full extension. Authorities differ slightly on how much elbow flexion is desirable during the middle third of the pull as well as the degree to which the shoulders and body should roll. Most agree, however, that body roll is necessary for efficient backstroking but should not be more than 45 degrees. At the completion of the pull, the arm is raised in a straight position from the water directly overhead with no lateral deviation. The hand is then ready to start the next pull.

The kick has seen many changes, the most important of which was the switch from the frog kick to alternating up-and-down movements of the legs, as in the crawl. Some swimmers favour continuous kicking; others prefer a broken tempo.

The head should remain still throughout the stroke. Breathing presents few problems, since the face is almost always out of water. Breathing, however, is cyclical as in all other strokes, inhalation usually coming as one arm reaches the high point of its recovery.

Crawl stroke. Crawl, often called freestyle, has become the fastest of all known strokes. In its infancy crawl swimming was considered fit only for sprinting, as it was thought too physically exhausting for the longer competitive distances. Presently, crawl is the almost unanimous choice of swimmers the world over for any distance. Many marathon swims, including crossings of the English Channel, are accomplished using crawl exclusively.

There is evidence that a type of crawl was practiced in the Solomon Islands of the South Pacific and elsewhere before the Cavill brothers of Australia, Syd and Charles, popularized it in Europe in 1902 and in America a year later. The introduction of the crawl in Australia, however, is generally credited to Harry Wickham. When demonstrating it in Australia around 1893, Wickham's alternating arm action was described by coach George Farmer as crawling through the water. The name crawl gained immediate favour, later becoming the Australian crawl, since the Cavill brothers, along with their countryman Cecil Healey, became most proficient after learning it.

Australian crawl

The chief difference between the crawl and the older trudgen is in the leg action. Trudgen relied on the wider scissor- or breaststroke-type kick, creating drag. The crawl kick eliminated the wide lateral movements by fluttering up and down.

The earliest knowledge of the crawl obtained in the United States was derived from hazy newspaper accounts and descriptions of the Cavills' stroke. In attempting to imitate the action, United States swimmers unwittingly developed a faster leg drive, executing four kicks per stroke instead of the two performed by the Cavills. Further experiments led to the belief that the four-beat leg thrash was more effective when composed of one comparatively wide and three very narrow scissoring motions. This style, termed the four-beat trudgen crawl or the four-beat single-rhythm crawl, became the accepted racing stroke. This combination of movements led to such an improvement in average performances that experts conceived the idea of obtaining still greater speed from a six-beat drive set to a double rhythm—two major and four minor kicks—which would minimize resistance and afford more even distribution of propelling power.

Six-beat crawl

Late in 1917 two young swimmers of the New York Women's Swimming Association, Charlotte Boyle and Claire Galligan, determined to give the six-beat kick a trial, and by the summer of 1918 they broke records with it over the longer regulation courses, 800 yards and one mile. The six-beat crawl immediately won favour. Within a few years it had become the recognized stroke the world over, not only for racing but for all-around purposes as well. A great many women and some men used an eight-beat kick successfully in the early years of crawl swimming, notably the Japanese Olympic swimmers in 1932.

Despite the fact that the six-beat crawl is the almost universal choice of sprinters, present-day distance swimmers use a variety of kicks, some six-beat, some four-beat, some scissor, some a combination of scissor and flutter. Others practically drag their legs, using a minimal crossover movement primarily to maintain balance, it being well established that the arm stroke is the major propulsive force in crawl swimming. Debbie Meyer (U.S.), 1968 Olympic champion in three events and then women's world record holder at all freestyle distances 200 yards and over, and Mike Burton (U.S.), 1968 Olympic 400 and 1,500 metre freestyle champion, demonstrated wide differences of leg action in their respective strokes. Miss Meyer swam a strong, classic six-beat crawl, while Burton, particularly in distance racing, used his arms predominantly and a minimal drag kick.

Developments in crawl swimming in the 1950s and 1960s were dramatic, due largely to more and better facilities, both indoor and outdoor, increased technical knowledge, and improved instruction, coaching, and training methods. It became increasingly difficult to predict ultimate limits. In the men's 1,500 metre race, the 1956 world record was 17 minutes 52.9 seconds set by George Breen (U.S.); by 1970 the world record for the same race stood at 15 minutes 57.1 seconds, set by John Kinsella (U.S.), a drop of almost two minutes. Comparable times for women at the same distance show a drop of over three minutes: 20 minutes 34.6 seconds in 1958 by Sylvia Ruuska (U.S.) to 17 minutes 19.9 seconds in 1969 by Debbie Meyer (U.S.).

The correct position in crawl swimming is that of a streamlined whole, the body lying flat on the surface, the legs kept slightly under the water. Raising the head high results in an exaggerated arch in the lower back, which increases body resistance. Lateral movements of the body must also be avoided. The arms move alternately, so timed that one will start pulling just before the other has finished its pull, making propulsion continuous. Most crawl swimmers complete about one-half of the arm pull before the recovering hand next enters the water. The hand enters the water with the elbow held high and slightly bent. Throughout the pull, the shoulder is always above the elbow, and the elbow above the wrist and hand. Bending the elbow affords greater strength in pulling. The force of the arm stroke can be thought of as pressing downward and backward during the first third of the stroke, pushing during the middle third, and pushing through and backward but not upward during the final third of the action, completing an arc of roughly 180 degrees in the vertical plane. The direction of pull should be on a line from the shoulder moving under the body and out past the thigh on recovery. On recovery the elbow is bent and carried high. The forearm circles quickly out and around until the hand extends toward the point of entry. The elbow is then extended, although not fully, for the catch of the hand on a line with the shoulder.

Stroke and kick basics

In kicking, the legs remain fairly straight but fully relaxed. The legs whip up and down alternately in an undulating action, so that in both the up kick and down kick there is a backward push directed against the water by the lower leg and more so by the foot. The scope of the thrash varies somewhat according to the individual length of leg and the number of beats used per stroke.

When a swimmer is viewed head on from an underwater window, a distinct roll from one side to the other is apparent. This newer technique has completely outmoded the previous no-roll style.

Breathing is effected by turning the head to either side during recovery of the arm on that side. Inhalation is quick, and the head is then returned to the face downward position, exhaling through the nose and mouth. Normally one breath is taken for each full cycle of the arms. In short-distance sprinting however, breathing may take place after two, three, or even four cycles of the arms. The skilled swimmer can breathe equally well on either side.

Competitive swimming. While swimming was not a part of the ancient Olympic Games, the sport was not unknown to the Greeks and Romans and other early civilizations. In Japan competition is known to have taken place as early as 36 BC, during the reign of Emperor Sujin; an imperial edict in AD 1603 made swimming an integral part of the school curriculum. By 1810 Japanese swimming competition had reached the level of a three-day meet.

Anglo-Saxons, however, are responsible for organizing the modern sport. The first championship meet of modern times was held in Sydney, Australia, in 1846. A group of London clubs met in 1869 and formed the Metropolitan Swimming Club Association, which later changed its name to the Amateur Swimming Association. The United States Amateur Athletic Union, formed in 1888, included swimming along with other sports under its supervision. Soon other Western nations founded similar governing bodies.

Controlling national groups met in London in 1908 and organized the Fédération Internationale de Natation Amateur (FINA), its purpose to establish programs and rules for Olympic Games and other international competitions and to supervise and register world records. From ten original members the FINA had grown to almost 100 by the 1970s. With few exceptions, notably China, practically all national swimming organizations are FINA-affiliated. Aquatic sports over which the FINA assumes authority are swimming (women's competition was recognized in 1921), diving, and water polo. Synchronized swimming (*i.e.*, swimming with music) was formally added for the 1955 Pan-American Games. Major competitions operating under the FINA code in addition to the Olympic Games, include the Pan-American Games, the Maccabiah Games, the European Games, the Asian Games, and more recently the World Student Games.

While most countries adhere to FINA rules, the United States, with its tremendous emphasis on club, interscholastic, and intercollegiate swimming, competes under rules of the Amateur Athletic Union (AAU) and National Collegiate Athletic Association (NCAA) for the simple reason that American swimming traditionally emphasizes short-course competition (pools 25 yards long), while FINA recognizes world standards set in 50-metre pools only.

Ever since Olympic competition was revived in 1896, freestyle (crawl) races have been more predominant than races using any other stroke. This also holds true for NCAA and interscholastic competition in the United States. Only in the past ten years have races in backstroke, breaststroke, butterfly, and individual medley (all four strokes) challenged the popularity of the crawl.

Olympic swimming

A study of Olympic competition reveals that excellence in performance follows no one pattern or nation, as Hungary, Denmark, Australia, Germany, France, Great Britain, Canada, Japan, and the United States have been leaders at different times. Since 1920, however, with the exception of the 1932 and 1956 Olympiads, the swimmers of the United States have won the largest share of Olympic honours. They were particularly strong in the 1960s, winning 50 of 74 Olympic events. (For complete lists of Olympic swimming champions, see ATHLETIC GAMES AND CONTESTS: *Olympic champions;* for world records see SPORTING RECORD in the *Ready Reference and Index.*)

Many factors contributed to the United States' rise to preeminence in swimming, the most important of which was the start of the age-group swimming programs in the late 1940s. Age-group swimming is a program sponsored by the AAU whereby boys and girls compete on local, sectional, and national levels in all competitive strokes at specific age levels: ten and under, 11–12, 13–14, and 15–17 (originally 15–16). Other countries have instituted similar programs. Unselfish help has come from coaches and editors of such swimming periodicals as *Swimming World* (U.S.) and the *Swimming Times* (London). These programs have provided youngsters with the stimulus necessary to reach local and international proficiency. In the United States alone, over 1,000,000 age-group swimmers enjoyed racing annually in the 1970s. In 1970, 500 swimmers representing 12 countries swam in the Seventh Annual Darmstadt (Germany) International Junior Gala.

Other factors contributing to the elevation of swimming to a place of eminence in world sport include increased construction of standard 50-metre pools featuring such technical innovations as wave-killing gutters, racing lane-markers that further reduce surface turbulence, the use of movie cameras to study strokes from underwater viewing windows, large-faced electric clocks that enable swimmers to observe and regulate their own performance in training, and electrically operated automatic judging and timing equipment.

All aspects of swimming are being explored continually. The World Swimming Coaches Association, formed at the 1968 Olympic Games, has as one of its goals the exchange of technical information on swimming. Out-of-water exercises, both in and out of season, body building, running, the use of both heavy and light weights to develop specific muscle groups, and isotonic and isometric exercises are some of the many methods used to increase strength, muscle tone, and endurance.

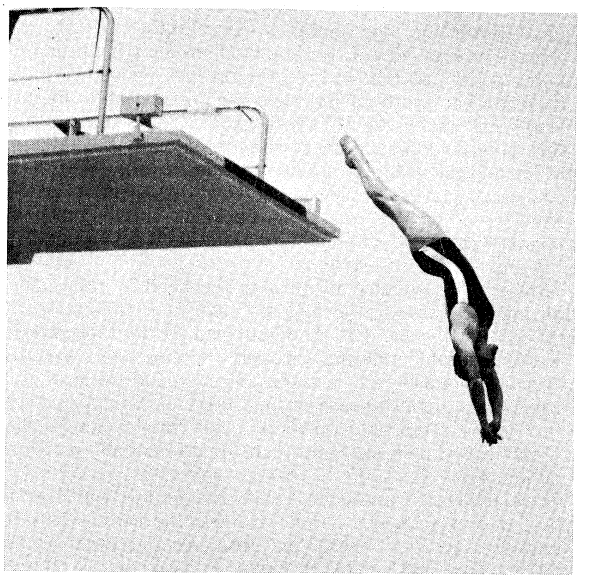
Before interval training and repeat training came into prominence in 1956, swimmers trained mainly by swimming as much distance as they could in any given workout regardless of stroke quality or speed. Interval training consists of a series of swims of the same distance with controlled rest periods. In slow interval training the rest period is always shorter in time than it takes to swim the prescribed distance. Fast interval training permits longer periods of rest with consequently greater recovery of the heart rate to normal. Slow interval training is used primarily to develop endurance, while fast interval training develops speed, with endurance a secondary goal. Repeat training is a type of swimming in which rest periods between swims are long enough to permit almost complete recovery of the heart and respiratory rate. The emphasis in repeat training is on speed and quality. All competitive swimmers, from age-group to Olympic champions, use both interval and repeat training methods.

DIVING

To dive means, most simply, to plunge into the water head first from a pool deck, a float, raised platform, or springboard. In competition dives are executed in a prescribed manner and consist of a takeoff from springboard or platform, a movement or series of movements performed in the air, and a smooth entry into the water either head first or feet first. Diving contests are a part of most aquatic meets.

When diving first became part of the Olympic program in 1904, it was little more than plain high diving from five- and ten-metre fixed platforms. The combination of acrobatic and gymnastic manoeuvres with diving, first called variety diving and later fancy diving, made a uniform method of judging the competitors necessary. An international committee in 1904 composed a table of dives, listed a description of each, and assigned a value for each based on its relative difficulty of performance.





Two types of dives.
(Left) Inward dive in pike position from a 3-metre springboard; (right) back dive in layout position from a 10-metre tower or platform.
(Left, right) Swimming World Publications, (left) Fort Worth *Star-Telegram*, (right) R.F. Lewis

Original tables listed but 20 springboard dives and 14 platform dives. By 1970 there were 77 springboard and 59 platform dives recognized in international competition, governed by the FINA.

Types of dives. Competitive dives, both platform and springboard, are separated into different groups: (1) forward dives, in which the performer, facing the water, dives out from the board and rotates ½ to 3½ turns before entry; (2) backward dives, in which the performer stands with his back to the water and rotates backward ½ to 2½ turns; (3) reverse dives, in which the performer begins in a forward position but changes the direction of his rotation from forward to backward after takeoff, making ½ to 2½ turns; (4) inward dives, in which the performer stands with his back to the water and, after a backward takeoff, rotates inward ½ to 2½ turns toward the board; (5) twisting dives, in which the performer spins about the long axis of his body, twisting ½ to 3½ turns while performing any dive from the other four categories. A sixth category, used only in platform contests, includes the armstand dives, in which the performer must maintain a perfectly balanced handstand at the edge of the platform before executing his dive.

Straight, pike, or tuck

The majority of dives in the FINA diving tables can be performed in three different positions: straight, pike, or tuck. In the straight position the body remains perfectly in line, with no flexion at the hips or knees permitted. In the pike position, there must be a distinct flexion at the hips, but none at the knees. In the tuck position, the body is gathered tightly into a ball with the hands grasping the shins firmly. In all three positions the toes must be pointed.

An indication of the advances made in diving during the 20th century is the report of the IV Olympiad (1908), which suggested the elimination of the double somersault from the tables because it was believed too difficult to control and a possible cause of injury. Modern divers, however, perform up to 3½ somersaults forward, 2½ somersaults inward or backward, and add complicated twisting movements while somersaulting—all with grace and beauty.

Judging and scoring. Winners are determined by a panel of judges—three in dual meets, five in international, and seven in Olympic competition. Judges score each diver from 0 to 10 points in ½ point increments, taking into account the run, the takeoff, the technique and grace of the dive during the passage through the air, and the entry into the water. The panel is supervised by a referee assisted by secretaries and an official scorer who records the award of each judge. When five judges officiate, the scorer eliminates the highest and lowest awards, multiplies the sum remaining by the degree of difficulty of the dive, and enters the result on a scoring sheet. Example: 5 judges (8)-7-7-7-(6) = 21 × 2.0 (degree of difficulty) = 42 (score). When seven judges are used, the score, obtained as before, is divided by 5 and multiplied by 3 to establish a comparable score: 7 judges (8)-7-7-7-7-7-(6½) = 35 total; 35 × 2.0 (degree of difficulty) = 70 ÷ 5 = 14; 14 × 3 = 42 (score). The winner of the contest is the one totalling the greatest number of points for all dives. Should two or more score the same number of points, a tie is declared.

Equipment. The popularity of diving is partially the result of improvement in equipment, most notably the invention of the Duraluminum springboard. A previous invention, the movable fulcrum, enabled the diver to adjust the flexibility of the board to his own liking.

Equipment for diving is defined by rules made by the FINA International Diving Committee, the most important of which follow.

Springboards may be 1 metre or 3 metres (10 feet) above the surface of the water; the board itself must be at least 4 metres long and ½ metre wide. Minimum depth of water is 3 metres beneath the 1-metre board and 3.5 metres (12 feet) beneath the 3-metre board. Originally springboards were made of wood, covered with cocoa matting to prevent slipping, later of fibreglass-sheathed laminated wood. The Duraluminum board, with its built-in nonskid surface, was first used in the Olympic Games at Melbourne in 1956 and has made the wooden board practically obsolete. Regulation springboards must be provided with a movable fulcrum easily adjusted by each diver.

Fixed boards (platforms or towers) may be 7.5 metres or 10 metres (33 feet) above the surface of the water. Platforms 2 metres wide and not less than 6 metres long must be rigid, accessible by stairs, covered over their entire length with rough cocoa matting, and the back and sides surrounded by a handrail. Minimum depth of water under the 7.5-metre (intermediate) tower is 4 metres (13 feet) and 4.5 metres (15 feet) under the 10-metre tower.

Competition. Men's springboard competition at the most proficient levels comprises 11 dives: 5 required (forward, backward, reverse, inward, and forward with ½ twist) and 6 voluntary dives selected from the diving tables. Dual meet diving at the interscholastic and intercollegiate levels consists of contests of 6 dives, 1 required and 5 voluntary. Men's tower diving competition comprises 6 dives, 1 from each category, with a total degree of difficulty that does not exceed 11.2; and 4 voluntary selections, again from different categories but with no limit on the degree of difficulty.

Women's championship springboard contests require

the same 5 required dives as the men, but only 5 voluntary dives. In women's tower competition there are 4 required dives, specifically named (forward, backward, reverse, and inward), and 4 voluntary dives.

In international competition, when there are more than 12 contestants, each contestant performs the required dives and three voluntary dives in a preliminary event. The 12 highest scorers then compete in the finals, in which they perform three voluntary dives. The United States national championship diving contests consist of three parts: preliminaries, semifinals, and finals, because the size of the field has become so unwieldy. Trials for the 1968 U.S. men's Olympic tower diving team alone saw more than 100 divers competing, while 75 or more divers are not unknown at the national intercollegiate championships. This far exceeds the numbers competing in European, Asian, Pan-American, or Olympic competitions.

The tremendous growth of diving in the United States has resulted in the separation of diving from its parent sport swimming, with increasing strength in all phases of the program. The American Diving Coaches Association, organized in 1969 with more than 150 members, has the power through its own Olympic Diving Committee to select divers and coaches for Pan American and Olympic diving teams, as well as for other international events. The committee has also organized age-group, regional, and zone competitions in greater numbers than ever before and has revitalized women's intercollegiate championships.

More frequent contests between divers from various countries at all levels have resulted in dramatic improvements everywhere. Recent Europe-versus-United States team-diving contests, exchange programs with Canadian, German, Swedish, and English divers, and the establishment of diving camps in various areas to give younger divers first-class coaching indicate future strength.

The dominance of U.S. divers grew out of the efforts of outstanding coaches. Both the American and World Diving Coaches Association (formed in 1968) have contributed to the standardization of all equipment and the spread of diving expertise. Their success is born out by the fact that Klaus Dibiasi (Italy) won the 10-metre tower Olympic title in 1968, breaking an unprecedented string of United States victories that began in 1920, complete except for the 10-metre win of U.S.-trained Joaquin Capilla (Mex.) in 1956. The record in ladies' diving is almost identical, with U.S. women winning most springboard and tower Olympic gold medals until 1960 when Ingrid Kramer (Germany) won both events. (For lists of Olympic diving champions, see ATHLETIC GAMES AND CONTESTS: *Olympic champions.*)

OTHER WATER EVENTS

Channel and open-water (marathon) swimming. Long before the advent of swimming pools, distance swimming was practiced by swimmers everywhere. It was not until 1875, however, that Matthew Webb, a 27-year-old Englishman, captured the fancy of the world by being the first person successfully to swim the English Channel, covering the distance from Dover to Calais in 21 hours and 45 minutes. English Channel swimming received its greatest publicity in 1926 when Gertrude Ederle, U.S. world-record holder, became the first woman to conquer the Channel, using the crawl the entire distance and finishing in 14 hours 31 minutes. Webb swam breaststroke.

The Channel has been conquered hundreds of times since then, the fastest crossing belonging to Barry Watson of England in 9 hours 35 minutes. The best time for women, 9 hours 59 minutes, is credited to the Australian professional Linda McGill. Florence Chadwick, (U.S.) was the first woman to swim the Channel in both directions, in 1950 and 1951. Undoubtedly the greatest feat is that of the Argentine professional Antonio Abertondo, who swam the Channel both ways in 1961, 18 hours 50 minutes from England to France and 24 hours 25 minutes for the return, a total of 43 hours 15 minutes. The Channel has even been conquered underwater by Fred Baldasare, an American frogman, using aqualung equipment and swimming in a towed submerged cage.

While the English Channel has received the most popular attention, there are other challenging long distance, or marathon, swims. Most of them are in the 15- to 37-mile (24- to 59-kilometre) range, although some are longer, notably the Lake George marathon (42 miles, or 67 kilometres), and the Lake Michigan Endurance Swim (50 miles, or 80 kilometres), both in the United States.

Marathon swimming

Since the FINA is concerned with amateur swimming at distances up to 1,500 metres only and since marathon swimming is largely professional in nature, an international group of amateurs and professionals met in 1954 to form the Fédération Internationale de Natation de Longue Distance (FINLD). Since the FINLD permitted amateurs to compete with professionals, problems were bound to arise, and in 1963 the professional distance swimmers founded the World Professional Marathon Swimming Federation (WPMSF) to set the standards for the conduct of marathon swimming. Eight sanctioned championship races are usually held each year. Countries sponsoring championship marathons include Canada, Egypt, Italy, Argentina, and the United States. During 1969, 10 international professional marathon races were held, 8 of which were sanctioned by the WPMSF, with 107 swimmers from 21 different countries participating.

Water polo. Water polo was first conceived in England in the 1870s. At first the game was an attempt by British gentlemen to emulate the polo played on horseback while in the water. "Horses" were made from wooden barrels, decorated with horseheads and fancy tails. Double-bladed paddles served both to strike the ball and to provide propulsion and balance. The game eventually gave way to one wherein all players remained in the water as swimmers.

The swimming game was introduced in the United States in 1890, at which time teams of six men played with a partially inflated ball. This "soft ball" game was characterized by very rugged underwater tackling and was called by some the roughest sport on Earth. Its popularity had diminished greatly by 1935, when it was succeeded by the international (hard ball) game in which underwater tackling was prohibited. This made the hard-ball game a surface rather than an underwater game, and swimming, ball handling, passing, and throwing skills became prime requisites.

Hungary has dominated the sport since that time, winning Olympic titles in 1932, 1936, 1952, 1956, and 1964, with Italy winning in 1948 and 1960, Yugoslavia in 1968, and the U.S.S.R. in 1972, also supplanting Great Britain, the Olympic winner four times in the early years of the sport. Hungary's strength was due in large part to the play of Dezső Gyermarti, who at 37 helped his country win a gold medal in his fifth Olympic competition (1964). (For Olympic water polo champions, see ATHLETIC GAMES AND CONTESTS: *Olympic champions.*)

Swimmers and water polo

Water polo has grown by leaps and bounds during the past 25 years, especially since the disproval of a superstition that a world class swimmer would lose his swimming form by playing polo. Due in part to the 1964 Olympic exploits of Roy Saari (U.S.) then world 1,500 metre record holder and a water polo player of international class at the same time, and Charles Bittick (U.S.), national collegiate backstroke champion and member of the 1960 U.S. Olympic water polo team, swimming coaches throughout the world at last permitted swimmers to play the game. The rapid growth of the sport is indicated by the fact that in the United States alone there is organized water polo competition at age-group, secondary school, club, collegiate, and national levels. Many swimming coaches urge their swimmers to engage in water polo as training to improve not only their swimming skills but their endurance as well.

International water polo is governed by the FINA. The game is played in a pool 20–30 metres (65–100 feet) long and 8–20 metres (25–65 feet) wide, by two teams of seven men each, using a rubber fabric composition ball, 68–71 centimetres (27–28 inches) in circumference, fully inflated. The depth of the pool may vary, except in championship play, when at least 1.8 metres (6 feet) is required. Players on one team wear numbered white caps; oppo-

nents wear a contrasting colour. The use of plastic ear guards is also recommended.

The game, four periods of five or seven minutes each with a two-minute interval between periods, is officiated by two referees, a timekeeper, a recorder, and two goal judges. Teams change sides after each period. A goal, scored as 1 point, is made by throwing the ball between the goal posts, which are 3 metres (10 feet) apart and 0.9 metres (3 feet) high, by dribbling the ball through them, or by a penalty throw, an unimpeded shot at a goal guarded only by a goalkeeper. Penalty throws are awarded by the referee for serious fouls or infractions of rules. Fouls may be technical, personal, or major. Technical fouls result in loss of the ball to the opposing team. A penalty throw is awarded to a team after its opponents commit ten personal fouls or one major foul. Any player committing five personal fouls must leave the game, although he may be replaced by a substitute. (K.B.M.)

Water skiing. Water skiing evolved in the 1920s and early 1930s in Europe and the United States concurrently with the rising popularity of motorboats capable of speeds over 20 miles per hour (32 kilometres per hour). Today it is practiced in nearly every country in the world wherever a stretch of calm water is available. Its popularity with people of all ages stems from the thrill of speed enjoyed without its usually attendant dangers.

General purpose and jumping skis are about 68 inches (1.7 metres) long by 6½ inches (16.5 centimetres) wide, with wooden fins on the bottom for stability. The resilient rubber foot binding is put on like a rubber overshoe and releases without injury in case of a fall. For slalom competition, a single ski is used that is tapered toward the rear, where a large metal fin is mounted. It uses two foot bindings in tandem. Trick riding is best done with skis 48 inches (1.2 metres) long by 10 inches (25 centimetres) wide, slightly rockered and without fins of any kind. The standard tow line is about 75 feet (23 metres) long with a wooden handle on a triangular bridle. Minimum boat speed is about 15 miles (24 kilometres) per hour.

The American Water Ski Association, with headquarters in Winter Haven, Florida, sanctions upward of 100 tournaments each year in the United States. Australia, France, Canada, Italy, Great Britain, and other countries have similar organizations. World competition is regulated by the World Water Ski Union in Geneva. All competition is open; no discrimination is made between amateur and professional.

Competition under a rigid set of rules is in three categories: jumping for distance at up to 35 miles (56 kilometres) per hour boat speed from a 6-foot (1.8-metre) high, sloping wooden ramp; slalom through a winding course of buoys at progressively higher boat speeds up to 36 miles per hour; and trick riding, where as many tricks as possible are performed in two 20-second runs at a speed selected by the contestant. Trick riding involves variations of turning from front to backwards on the water and in the air from wake or ramp jumps and holding the tow line with the foot as well as the hands.

Spectacular variations of water skiing include barefoot skiing, where boat speeds of 35 to 40 miles (56 to 64 kilometres) per hour are required to enable the rider to "ski" on his feet on the surface and do many of the manoeuvres including ramp jumping, that normally are performed on skis.

Ski kiting is done with a large fabric kite on a 100-foot (30-metre) or longer line. The kite is attached to the tow line and the skier, using light skis, holds a trapeze-like bar attached to the kite. Lift-off in still air is at 25 to 30 miles (40 to 48 kilometres) per hour. Elevation is controlled by boat speed and, in some kites, by the flyer. Competition on kites includes slalom and trapeze tricks. Some types of kites can cut loose from the tow line and do free flights to a landing. (For lists of water skiing world records and champions, see SPORTING RECORD in the *Ready Reference and Index.*)

Surf riding (surfing). Surf riding, or surfing, was practiced by the Polynesian islanders when the first European explorers visited the Pacific Ocean. Capt. James Cook reported at length on his observations of the sport when he reached Hawaii in 1778. For many years it was regarded as an exotic pastime, practiced only by Pacific islanders, some of whom demonstrated it for tourists, by a few adventurous visitors, and by a few mainlanders in Australia and Southern California who had seen demonstrations. The real growth of the sport came with the growth of other water sports after World War II, and it is now practiced around the world wherever wave action permits. Ideal conditions occur where the open ocean bottom slopes very gradually to depths of not over 10 feet (3 metres) at distances from 100 to 1,000 yards (90 to 900 metres) from the shoreline.

All competitive and most recreational surfing is done in standing position on hollow, light wood or foam-filled fibreglass boards about 2 feet (60 centimetres) wide and 10 to 12 feet (300 to 360 centimetres) long, tapered and rounded on ends and edges with a stabilizing fin on the bottom near the rear end. It is also done in prone position with narrower boards, 3 to 5 feet (1 to 1.5 metres) long; on air mattresses; or with the body itself in a prone position ("body surfing").

The rider first paddles or swims his board out from shore to the area where the ocean rollers peak up, but have not yet broken at the crest. As a suitably shaped wave approaches him, he lies prone on his board and paddles toward shore to gain speed so that when the wave reaches him he will slide down its forward face as it breaks and continue toward the shore. After catching the wave, he climbs to his knees and then to his feet. By banking the board he can steer it; by riding diagonally across the wave as it advances to shore, he can increase his speed and the distance travelled. Breaking waves from 2 to 20 feet (0.6 to 6 metres) in height can be ridden for distances up to 900 metres (1,000 yards).

Competitions are held nationally and internationally both for individual surfers (men and women) and two on a board (mixed). Points are awarded for ability to catch the wave, wave selected, distance travelled, skill in manoeuvring, form, and acrobatic tricks and lifts done.

Surf boards are also ridden in calm water on the large stern wakes of a slowly driven powerboat (not to be confused with aquaplaning, where a board is attached by a rope to the boat and driven at higher speeds). On some inland lakes, surfing is performed on waves produced by special machines.

Skin diving. Skin diving has been practiced throughout history as a means of fishing and salvage of shipwrecks. As a popular sport, it started in the 1930s with the use of glass and rubber face masks that allow a diver to see clearly underwater; foot fins to greatly increase mobility; and a J-shaped breathing tube or "snorkle" to permit breathing while lying face down on the water and watching below. Somewhat later, "wet" suits of cellular rubber material were developed for use in cold water or for prolonged diving.

A great majority of the approximately 3,000,000 modern skin divers do not dive very deep or very often. They prefer to swim with mask, snorkle, and flippers on the surface and enjoy the underwater scenery. The coral reef areas of the tropical zone around the world are generally considered the most rewarding for this purpose, but any shallow sea, pond, spring, or river has interest.

To dive, the body is bent at the waist, the head ducked downward and the feet raised straight up in the air. The weight of the feet plus a breast stroke with the arms will propel the diver down if he is weighted for near neutral buoyancy. On descending, the pressure in the ears must be neutralized by swallowing, yawning, or holding the nostrils while blowing gently. With practice, dives of 35-foot (11-metre) depth and a full minute duration can be achieved by most nonsmokers.

The development and marketing of the scuba (Self Contained Underwater Breathing Apparatus), starting in the 1940s, gave great impetus to sport diving. Using bottles of highly compressed air that gave the diver up to two hours submergence time, the scuba allowed him to live like a fish and explore or work without coming to the surface for air. Valid instruction in the use of scuba is imperative for safety.

For spear fishing, scuba is not used by sportsmen. The spear guns used are powered by rubber tubing, springs, compressed gas, or gun powder. The spear is generally about 4 feet (1.2 metres) long with a hinged barb at its point. Since the range of the gun is 6 feet (1.8 metres) or less, the fish must be closely stalked by the diver to shoot it. For conservation purposes, only fish that will be eaten or stuffed as trophies should be shot.

Spear-fishing, salvage, and photography

Diving to investigate shipwrecks, hunt treasure, and explore archaeological finds has another group of devotees. Salvage involves a great deal of underwater and surface equipment and is more complex and expensive than imagined by most amateur divers.

One of the most popular facets of diving is underwater photography. By means of the great variety of still and movie cameras available to the sport diver, the beauty of the underwater world can be captured on film. Photography is also widely used by naturalists and marine biologists to record in detail the fantastic variety of life in the sea.

Competitions in diving and spear fishing are held at irregular intervals under varying rules specifying hours of diving and species of fish to be taken. Diving for record depths or duration is extremely dangerous and is not encouraged by any sporting organization. (J.A.)

BIBLIOGRAPHY. FRANCOIS OPPENHEIM, *The History of Swimming* (1970), a comprehensive history of swimming from antiquity to the present, tracing its organization as a sport and the evolution of various strokes and conditioning; *Cecil Colwin on Swimming* (1969), an illustrated discussion of backstroke, motivation, and dual resistance exercises; D.L. GAMBRIL, *Swimming* (1969), detailed training schedules of champion swimmers, circuit exercise routines, organization of practice, techniques for the individual, and stroke mechanics; J.E. COUNSILMAN, *The Science of Swimming* (1968), a comprehensive work covering every aspect of competitive swimming, including modern concepts of conditioning, stroke mechanics, instruction techniques, and land exercises; FORBES CARLILE, *On Swimming* (1963), an illustrated work covering all aspects of competitive swimming, with an excellent discussion of early Australian history; G. FORSBURG, *Modern Long Distance Swimming* (1963), a history of long distance, training methods, and famous swimmers and races; A.R. FAIRBANKS, *Teaching Springboard Diving* (1963), an excellent work for the beginner showing methods by which diving may be taught from prediving fundamentals through basic competitive dives; CHARLES BATTERMAN, *Techniques of Springboard Diving* (1968), a profusely illustrated detailed work for the expert, with explanations of difficult movements scientifically presented; A.F. LAMBERT and R. GAUGHRAN, *Technique of Water Polo* (1969), fundamentals, strategy, offensive and defensive play, conditioning, goal keeping, and training with action photographs and schematic drawings.

Publications: *FINA Handbook*, official handbook of the Fédération Internationale de Natation Amateur (quadrennial), constitution, rules and laws for international swimming, diving, water polo, and synchronized swimming; *AAU Guide*, official handbook of the American Athletic Union of the United States (annual), swimming, diving, and water polo rules, age-group and Junior Olympic rules; *NCAA Guide*, official handbook of the National Collegiate Athletic Association (annual), rules for interscholastic and intercollegiate swimming, with All-American selections and sectional, regional, and national championships; *Swimming World* (monthly), news on every aspect of swimming; *Swimming Technique* (quarterly), a more technical journal of swimming, diving, and water polo, featuring mainly scientific articles.

Skin diving, water skiing, and surfing: J. STRYKOWSKI, *Diving for Fun* (1969); JACQUES COUSTEAU, *The Silent World* (1953); H.S. ZIM and H.H. SHOEMAKER, *Fishes: A Guide to Fresh and Salt-water Species* (1956); J.H. ANDRESEN, *Skiing on Water*, 3rd ed. (1960); W.D. CLIFFORD and T.C. HARDMAN, *Let's Go Water Skiing* (1965). The best source of general information on surfing is *Surfing Yearbook* (annual); see also the periodicals *Skin Diver Magazine* (monthly), and *Water Skier* (7/yr.).

(K.B.M./J.A.)

Switzerland

Landlocked amid the mountains of central Europe, Switzerland is a nation the people of which have no unity of ethnic heritage nor of language or religion and the land of which is splintered by barriers of rock and river; but it is a nation that nonetheless is united and prosperous. Officially known as Schweizerische Eidgenossenschaft, or Swiss Confederation, Switzerland is a republican federation of 22 cantons, or states, three of which (Basel, Unterwalden, and Appenzell) are divided into half cantons. It is a *mater fluviorum* ("mother of rivers"), situated at the hydrographic centre of Europe and bordered on the west by France, on the north by West Germany, on the east by Austria and the Principality of Liechtenstein, and on the south by Italy. Two-thirds of its borders follow the natural contours of mountain ridges and crests, lakes, and rivers. Except for minor adjustments, Swiss borders have remained stable since their establishment at the Congress of Vienna in 1815, a stability from which the economy of the nation often has benefitted.

Switzerland's small area of 15,941 square miles (41,288 square kilometres; about one-half the size of Scotland or of Maine) and population of nearly 6,270,000 (1970), give no indication of the country's character or significance on the international scene. Switzerland's geographically imposed role as guardian of Europe's natural trans-Alpine routes has been both a reason for and a basic tenet of its existence—a role expressed in its traditional neutrality in time of war that was sanctioned by the Treaty of Westphalia in 1648 and confirmed in 1815 in Vienna, by the Treaty of Versailles in 1919, and again in 1920, when the Council of the League of Nations acknowledged it as "conditioned by a centuries-old tradition explicitly incorporated in international law." The tradition was upheld even while war raged around Switzerland's borders during 1939–45.

Traditional neutrality

Switzerland owes its existence, however, no less to the will of its inhabitants than to the exigencies of geography. With one-quarter of its area comprising high Alps, lakes, and barren rock and with no seaboard and few natural resources other than waterpower, Switzerland has managed to impose unity on diverse races, religions, and languages: for almost 700 years it has maintained the world's oldest and the European mainland's most virile democracy, achieving an almost unrivalled standard of living. It has become a financial centre and headquarters of more than 300 international institutions, including the International Committee of the Red Cross in Geneva and the Universal Postal Union in Bern. (For information on related topics, see SWITZERLAND, HISTORY OF; special features are covered in ALPINE LAKES; ALPS MOUNTAIN RANGES; JURA MOUNTAINS; GENEVA; and ZURICH.)

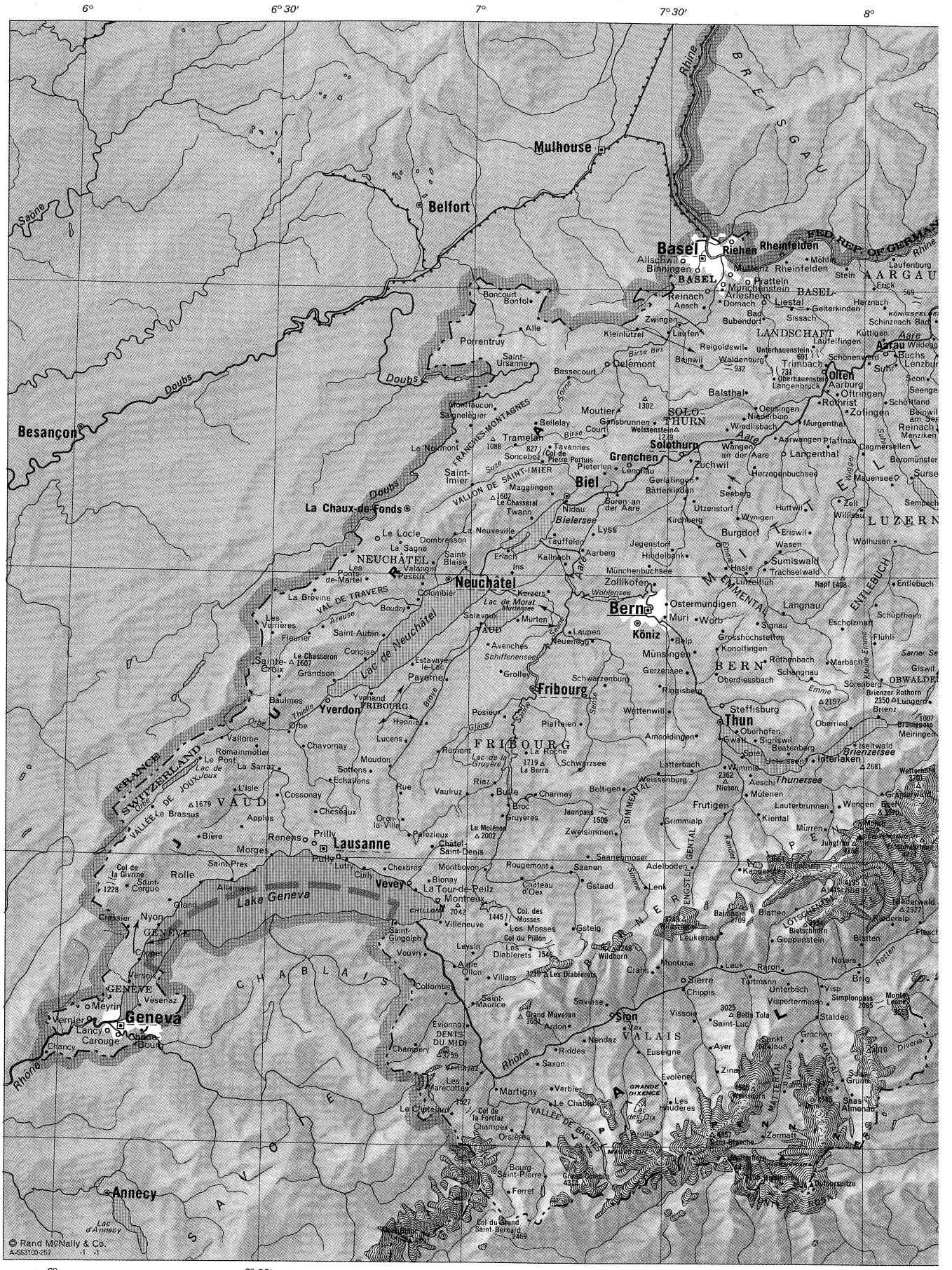
THE NATURAL AND HUMAN LANDSCAPE

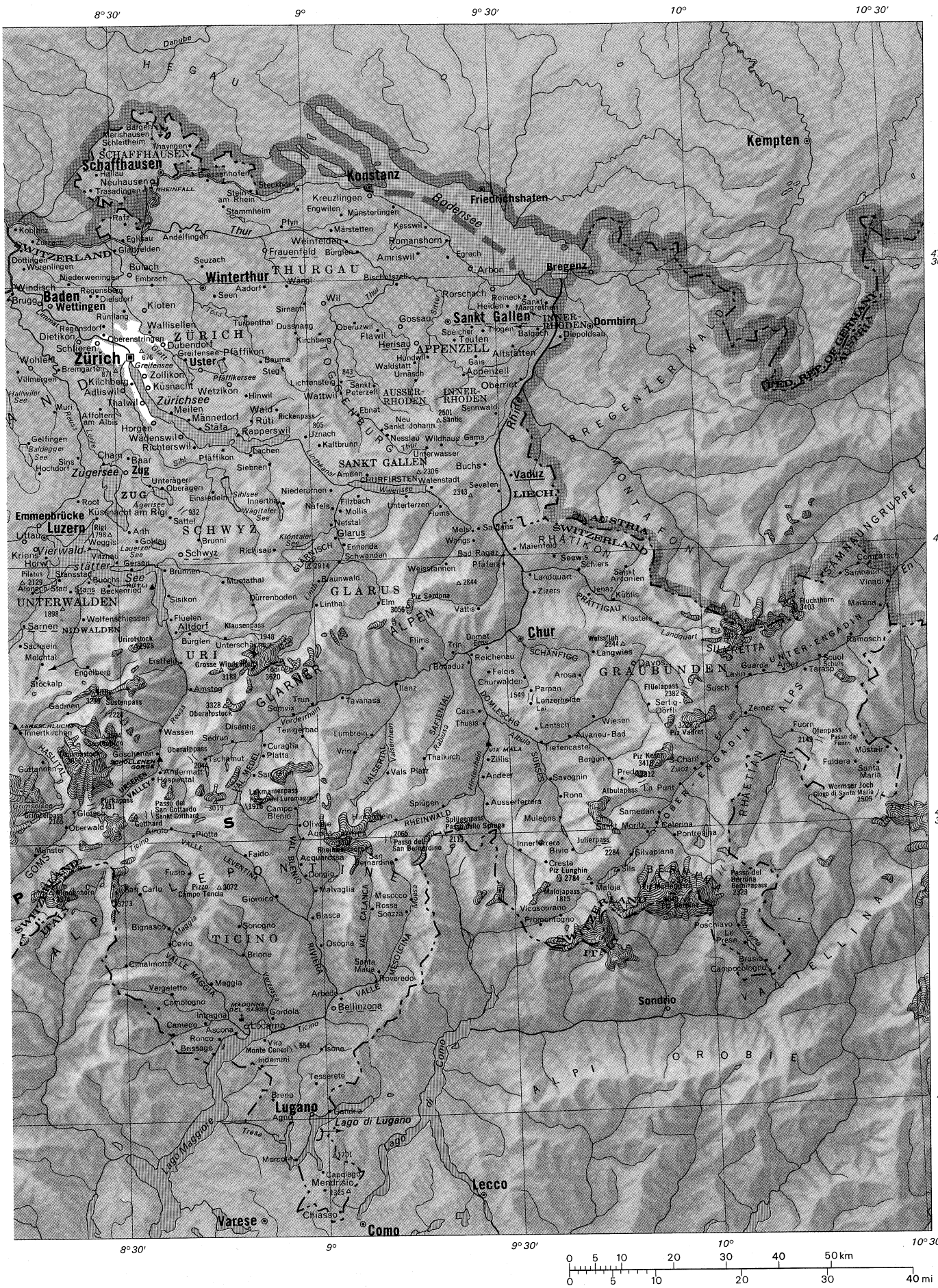
Geographical regions. Switzerland's small area contains an unusual diversity of topographic elements, which are divisible into three distinct regions: the Jura, the Alps, and the plateau area known as the Mittelland.

The Jura (Celtic: "forest"), a bleak and rolling mountain range in the northwest, occupies about 10 percent of the country. The region was formed under the extended impact of the general Alpine folding, particularly of two phases of that folding. Limestone and marl with rich fossil content are the characteristic rocks. The ridges, covered with meadows and only sparsely forested, receive greater amounts of rainfall than do the valleys, the slopes of which are wooded. Between Val St. Imier and the Doubs, a river that forms part of the border with France, mountains have been reduced by denudation to an undulating plateau, or peneplain. This plateau Jura, which extends into France, is known geographically as the Franches-Montagnes (Free Mountains), a name acquired in 1384 when the Bishop of Basel freed the inhabitants from taxation to encourage settlement of the remote area. The unfolded plateaus of the Jura tableland, atop which mixed agriculture is practiced, rise slightly toward the north and northwest, and their steep edges are forested.

Diversity of natural relief

In origin, the Alps were built of large complexes of massed overthrusts of extremely varied rocks. Especially well-known for its scenery is the Valais (a canton),


6°
6° 30′
7°
7° 30′
8°
47° 30′
47°
46° 30′
46°
Basel
Riehen
Rheinfelden
Mulhouse
Belfort
Besançon
Porrentruy
Delémont
Moutier
Solothurn
Grenchen
Biel
La Chaux-de-Fonds
Le Locle
Neuchâtel
Bern
Köniz
Thun
Fribourg
Yverdon
Lausanne
Vevey
Montreux
Geneva
Annecy
Sion
Brig
Zermatt
Interlaken
Olten
Aarau
Lake Geneva
Rhine
Rhône
Doubs
Aare
Sarine
Emme
Thielle
Orbe
Broye
Lac de Neuchâtel
Bielersee
Thunersee
Wohlensee
Schiffenensee
Lac de la Gruyère
Lac de Joux
Lac d'Annecy
Lac de Morat
Murtensee
Lac Léman
FRANCE
SWITZERLAND
VAUD
NEUCHÂTEL
FRIBOURG
VALAIS
GENÈVE
JURA
CHABLAIS
ALPES
BERNER OBERLAND
EMMENTAL
SIMMENTAL
ENGSTLIGENTAL
VALLÉE DE JOUX
VALLÉE DE BAGNES
VAL DE TRAVERS
VALLON DE SAINT-IMIER
FRANCHES-MONTAGNES
BREISGAU
FED. REP. OF GERMANY
AARGAU
SOLOTHURN
LANDSCHAFT BASEL
LUZERN
OBWALDEN
ENTLEBUCH
Mont Blanc
Grand Combin
Dent Blanche
Matterhorn
Weisshorn
Monte Leone
Jungfrau
Eiger
Mönch
Finsteraarhorn
Aletschhorn
Col du Grand Saint Bernard
Col du Pillon
Col de la Givrine
Col de la Forclaz
Col des Mosses
Col de Pierre Pertuis
Col de Jaman
Jaunpass
Gemmipass
Simplonpass
Grimselpass
Grosse Scheidegg
Brünigpass
Oberalppass
Furkapass
Nufenenpass
Grand Muveran
Le Moléson
Les Diablerets
Dents du Midi
Chasseron
Chasseral
Tête de Ran
Mythenstein
Pilatus
Rigi
Titlis
Napf
Säntis
Glärnisch
Tödi
Piz Bernina
Piz Roseg
Piz Palü
Monte Rosa
Dom
Lyskamm
Castor
Pollux
Breithorn
Matterhorn
Gran Paradiso
Mont Pelvoux
Mont Blanc de Cheilon
Grand Cornier
Dent d'Hérens
Aletschgletscher
Rhonegletscher
Grindelwaldgletscher
Unteraargletscher
Fiescherhorn
Finsteraarhorn
Schreckhorn
Wetterhorn
Eiger
Mönch
Jungfrau
Blümlisalphorn
Wildstrubel
Wildhorn
Diablerets
Oldenhorn
Gummfluh
Rochers de Naye
Dent de Morcles
Dent de Jaman
Dent de Lys
Dent de Broc
Dent de Corjon
Dent de Vaulion
Dent de Vaulion
Mont Tendre
Mont Racine
Mont Suchet
Mont Aubert
Mont Vully
Mont Soleil
Mont Crosin
Mont Raimeux
Mont Terri
Mont Repais
Mont Vuache
Mont Salève
Mont Sion
Mont Gros
Mont Gros
Mont Gros


8° 30′
9°
9° 30′
10°
10° 30′
47° 30′
47°
46° 30′
46°
Schaffhausen
Konstanz
Friedrichshafen
Kempten
Bregenz
Winterthur
Sankt Gallen
Dornbirn
Baden
Wettingen
Zürich
Uster
Zürichsee
Zugersee
Zug
Luzern
Emmenbrücke
Vierwaldstätter See
Sarnen
Altdorf
Stans
Schwyz
Glarus
Chur
Vaduz
Lugano
Lago di Lugano
Lago Maggiore
Como
Lecco
Varese
Sondrio
Bellinzona
Locarno
Davos
St. Moritz
Arosa
Klosters
Landquart
Thusis
Andeer
Zernez
Scuol
Samnaun
Ramosch
Martina
Poschiavo
Tirano
Bormio
Livigno
Silvaplana
Pontresina
Samedan
Celerina
Susch
Sils
Maloja
Chiavenna
Splügen
Hinterrhein
Vals Platz
Flims
Ilanz
Disentis
Andermatt
Oberalppass
Gotthard
Lukmanierpass
Passo del San Gottardo
Passo del San Bernardino
Julierpass
Albulapass
Flüelapass
Ofenpass
Wormser Joch
Bernina
Piz Bernina
Piz Palü
Piz Roseg
Piz Kesch
Piz Linard
Piz Buin
Piz Lischana
Piz Languard
Piz Nair
Piz Corvatsch
Piz Bernina
Piz Zupò
Piz Varuna
Piz Cambrena
Piz Scalino
Piz Tschierva
Piz Morteratsch
Piz Roseg
Piz Kesch
Piz Ela
Piz Mitgel
Piz Platta
Piz Curvèr
Piz Beverin
Piz Grisch
Piz Mundaun
Piz Sardona
Piz Vorab
Piz Segnas
Piz Mattmark
Piz Gallegione
Piz Cengalo
Piz Badile
Piz Bernina
Piz Roseg
Piz Kesch
Piz Ela
Piz Mitgel
Piz Platta
Piz Curvèr
Piz Beverin
Piz Grisch
Piz Mundaun
Piz Sardona
Piz Vorab
Piz Segnas
Piz Mattmark
Piz Gallegione
Piz Cengalo
Piz Badile
Danube
Rhine
Inn
Aare
Reuss
Limmat
Thur
Töss
Glatt
Sihl
Linth
Walensee
Zürichsee
Lake Constance
Bodensee
Rheinfall
Rheinwald
Splügenpass
Passo del San Bernardino
Julierpass
Albulapass
Flüelapass
Ofenpass
Wormser Joch
Bernina
Maloja
Chiavenna
Splügen
Hinterrhein
Vals Platz
Flims
Ilanz
Disentis
Andermatt
Oberalppass
Gotthard
Lukmanierpass
Passo del San Gottardo

MAP INDEX

Political subdivisions

Aargau.....47·30n 8·10e
Appenzell Ausser-Rhoden.47·27n 9·30e
Appenzell Inner-Rhoden...47·18n 9·21e
Basel-Land.....47·32n 7·32e
Basel-Stadt.....47·30n 7·50e
Bern.....46·45n 7·40e
Fribourg.....46·45n 7·05e
Genève.....46·15n 6·10e
Glarus.....47·00n 9·03e
Graubünden.....46·45n 9·30e
Luzern.....47·05n 8·05e
Neuchâtel.....47·00n 6·55e
Nidwalden.....46·35n 8·15e
Obwalden.....46·50n 8·25e
Sankt Gallen.....47·10n 9·08e
Schaffhausen....47·40n 8·35e
Schwyz.....47·05n 8·40e
Solothurn.....47·25n 7·35e
Thurgau.....47·35n 9·00e
Ticino.....46·20n 8·45e
Uri.....46·50n 8·40e
Valais.....46·10n 7·30e
Vaud.....46·40n 6·30e
Zug.....47·00n 8·30e
Zürich.....47·25n 8·40e

Cities and towns

Aadorf.....47·30n 8·54e
Aarau.....47·23n 8·03e
Aarberg.....47·03n 7·16e
Aarburg.....47·19n 7·54e
Aarwangen.....47·15n 7·46e
Acquarossa.....46·28n 8·57e
Adelboden.....46·30n 7·33e
Adliswil.....47·19n 8·32e
Aesch.....47·28n 7·36e
Aeschi.....46·40n 7·42e
Affoltern am Albis.....47·17n 8·27e
Agno.....46·00n 8·54e
Aigle.....46·19n 6·58e
Airolo.....46·32n 8·37e
Allaman.....46·28n 6·24e
Alle.....47·26n 7·08e
Allschwil.....47·33n 7·33e
Alpnach-Stad....46·57n 8·17e
Altdorf.....46·53n 8·39e
Altstätten.....47·23n 9·33e
Alvaneu-Bad.....46·40n 9·39e
Amden.....47·09n 9·11e
Amriswil.....47·33n 9·18e
Amsoldingen....46·43n 7·35e
Amsteg.....46·46n 8·41e
Andeer.....46·36n 9·26e
Adelfingen.....47·36n 8·41e
Andermatt.....46·38n 8·36e
Appenzell.....47·20n 9·25e
Apples.....46·34n 6·25e
Áquila.....46·31n 8·57e
Arbedo.....46·12n 9·03e
Arbon.....47·31n 9·26e
Ardez.....46·46n 10·11e
Arlesheim.....47·30n 7·37e
Arolla.....46·02n 7·29e
Arosa.....46·47n 9·41e
Arth.....47·04n 8·31e
Ascona.....46·09n 8·46e
Ausserferrera....46·33n 9·26e
Avenches.....46·53n 7·02e
Ayer.....46·11n 7·36e
Baar.....47·12n 8·32e
Bad Bubendorf..47·27n 7·44e
Baden.....47·29n 8·18e
Bad Pfäfers, see Pfäfers
Bad Ragaz.....47·00n 9·30e
Balgach.....47·25n 9·35e
Balsthal.....47·19n 7·42e
Bargen.....47·48n 8·37e
Basel (Bâle).....47·33n 7·35e
Bassecourt.....47·20n 7·15e
Bätterkinden.....47·08n 7·32e
Baulmes.....46·48n 6·32e
Bauma.....47·23n 8·53e
Beatenberg.....46·42n 7·48e
Beckenried.....46·58n 8·29e
Beinwil.....47·22n 7·35e
Beinwil am See..47·16n 8·13e
Bellelay.....47·16n 7·10e
Bellinzona.....46·11n 9·02e
Belp.....46·53n 7·30e
Bergün.....46·38n 9·45e
Bern (Berne).....46·57n 7·26e
Beromünster....47·12n 8·11e
Biasca.....46·22n 8·58e
Biel (Bienne).....47·10n 7·12e
Bière.....46·33n 6·20e
Bignasco.....46·20n 8·36e
Binningen.....47·32n 7·34e
Bischofszell.....47·29n 9·15e
Bivio.....46·28n 9·38e
Blatten.....46·25n 7·50e
Blonay.....46·28n 6·54e
Boltigen.....46·38n 7·24e
Bonaduz.....46·49n 9·25e
Boncourt.....47·30n 6·56e
Bonfol.....47·29n 7·09e
Boudry.....46·57n 6·50e
Bourg-Saint-Pierre.....45·57n 7·12e
Braunwald.....46·56n 9·00e
Bremgarten.....47·21n 8·21e
Breno.....46·02n 8·53e
Brienz.....46·46n 8·03e
Brig.....46·19n 8·00e
Brione.....46·18n 8·47e
Brissago.....46·07n 8·43e
Broc.....46·36n 7·06e
Brugg.....47·29n 8·12e
Brunnen.....47·00n 8·36e
Brunni.....47·03n 8·42e
Brusio.....46·14n 10·07e
Buchs.....47·10n 9·28e
Buchs.....47·23n 8·04e
Bülach.....47·31n 8·32e
Bulle.....46·37n 7·04e
Buochs.....46·58n 8·22e
Büren an der Aare.....47·08n 7·23e
Burgdorf.....47·04n 7·37e
Bürglen.....46·53n 8·40e
Bürglen.....47·33n 9·09e
Camedo.....46·09n 8·37e
Campo Blenio..46·34n 8·56e
Campocologno ..46·13n 10·08e
Capolago.....45·55n 8·59e
Carouge.....46·11n 6·09e
Cazis.....46·43n 9·25e
Celerina.....46·31n 9·51e
Cevio.....46·19n 8·36e
Cham.....47·11n 8·28e
Champéry.....46·10n 6·52e
Champex.....46·02n 7·07e
Chancy.....46·08n 6·00e
Charmey.....46·38n 7·10e
Château d'Oex...46·28n 7·08e
Châtel Saint Denis.....46·32n 6·54e
Chavornay.....46·43n 6·34e
Chêne-Bourg....46·12n 6·12e
Cheseaux.....46·35n 6·36e
Chexbres.....46·29n 6·47e
Chiasso.....45·50n 9·01e
Chippis.....46·17n 7·33e
Chur.....46·51n 9·32e
Churwalden.....46·47n 9·33e
Cimalmotto.....46·17n 8·29e
Collombey.....46·16n 6·57e
Colombier.....46·58n 6·52e
Comologno.....46·12n 8·34e
Compatsch.....46·58n 10·25e
Concise.....46·51n 6·43e
Coppet.....46·19n 6·12e
Cossonay.....46·37n 6·31e
Court.....47·14n 7·20e
Crans.....46·19n 7·28e
Crassier.....46·22n 6·11e
Cresta.....46·28n 9·31e
Cully.....46·29n 6·44e
Curaglia.....46·41n 8·51e
Dagmersellen....47·13n 7·59e
Davos.....46·48n 9·50e
Delémont.....47·22n 7·21e
Dielsdorf.....47·29n 8·27e
Diessenhofen....47·41n 8·45e
Dietikon.....47·24n 8·24e
Disentis.....46·43n 8·51e
Domat Ems.....46·50n 9·28e
Dombresson.....47·04n 6·58e
Dongio.....46·27n 8·58e
Dornach.....47·29n 7·37e
Döttingen.....47·34n 8·16e
Dübendorf.....47·25n 8·38e
Dürrenboden....46·57n 8·50e
Dussnang.....47·26n 8·58e
Ebnat.....47·15n 9·08e
Echallens.....46·38n 6·38e
Eglisau.....47·34n 8·32e
Egnach.....47·33n 9·23e
Einsiedeln.....47·08n 8·45e
Elm.....46·55n 9·11e
Embrach.....47·30n 8·36e
Emmenbrücke...47·04n 8·17e
Engelberg.....46·49n 8·25e
Engwilen.....47·37n 9·05e
Ennenda.....47·01n 9·05e
Entlebuch.....47·00n 8·04e
Eriswil.....47·05n 7·51e
Erlach.....47·03n 7·06e
Ermatingen.....47·41n 9·06e
Erstfeld.....46·49n 8·39e
Escholzmatt.....46·55n 7·56e
Estavayer le Lac.....46·51n 6·50e
Euseigne.....46·10n 7·25e
Evionnaz.....46·11n 7·01e
Evolène.....46·07n 7·30e
Faido.....46·29n 8·48e
Feldis.....46·48n 9·26e
Ferret.....45·55n 7·06e
Fiesch.....46·20n 8·10e
Filzbach.....47·07n 9·08e
Flawil.....47·24n 9·12e
Fleurier.....46·54n 6·35e
Flims.....46·50n 9·17e
Flüelen.....46·54n 8·38e
Flühli.....46·53n 8·01e
Flums.....47·05n 9·20e
Frauenfeld.....47·34n 8·54e
Fribourg (Freiburg).....46·48n 7·09e
Frick.....47·31n 8·01e
Frutigen.....46·35n 7·39e
Fuldera.....46·37n 10·22e
Fusio.....46·27n 8·40e
Gadmen.....46·44n 8·21e
Gais.....47·22n 9·27e
Gams.....47·12n 9·28e
Gandria.....46·01n 9·00e
Gänsbrunnen....47·16n 7·28e
Gelfingen.....47·13n 8·16e
Gelterkinden....47·28n 7·51e
Geneva (Genève).46·12n 6·09e
Gerlafingen.....47·10n 7·34e
Gersau.....47·00n 8·32e
Gerzensee.....46·51n 7·33e
Giornico.....46·24n 8·52e
Giswil.....46·50n 8·11e
Gland.....46·26n 6·16e
Glarus.....47·02n 9·04e
Glattfelden.....47·33n 8·30e
Gletsch.....46·34n 8·22e
Goldau.....47·03n 8·33e
Goppenstein.....46·22n 7·45e
Gordola.....46·11n 8·52e
Göschenen.....46·40n 8·35e
Gossau.....47·25n 9·15e
Grächen.....46·12n 7·50e
Grandson.....46·49n 6·38e
Greifensee.....47·22n 8·41e
Grenchen.....47·11n 7·24e
Grimmialp.....46·34n 7·29e
Grindelwald.....46·37n 8·02e
Grolley.....46·50n 7·05e
Grosshöchstetten.....46·55n 7·38e
Gruyères.....46·35n 7·05e
Gstaad.....46·28n 7·17e
Gsteig.....46·23n 7·16e
Guarda.....46·47n 10·09e
Guttannen.....46·39n 8·18e
Gwatt.....46·43n 7·38e
Hallau.....47·42n 8·27e
Hasle.....47·01n 7·39e
Heiden.....47·27n 9·33e
Henniez.....46·44n 6·54e
Herisau.....47·23n 9·17e
Herznach.....47·28n 8·03e
Herzogenbuchsee.....47·12n 7·41e
Hindelbank.....47·03n 7·32e
Hinterrhein.....46·32n 9·12e
Hinwil.....47·18n 8·51e
Hochdorf.....47·10n 8·17e
Horgen.....47·15n 8·36e
Horw.....47·01n 8·18e
Hospental.....46·37n 8·34e
Hundwil.....47·22n 9·19e
Huttwil.....47·07n 7·51e
Ilanz.....46·46n 9·12e
Il Fuorn.....46·40n 10·12e
Indemini.....46·06n 8·50e
Innerferrera.....46·31n 9·28e
Innerthal.....47·06n 8·56e
Innertkirchen....46·42n 8·14e
Ins.....47·00n 7·06e
Interlaken.....46·41n 7·51e
Intragna.....46·10n 8·42e
Iseltwald.....46·43n 7·58e
Isone.....46·08n 8·59e
Jegenstorf.....47·03n 7·30e
Jenaz.....46·55n 9·45e
Kallnach.....47·01n 7·14e
Kaltbrunn.....47·12n 9·02e
Kandersteg.....46·30n 7·40e
Kerzers.....46·58n 7·12e
Kesswil.....47·36n 9·20e
Kiental.....46·35n 7·43e
Kilchberg.....47·19n 8·33e
Kirchberg.....47·05n 7·35e
Kirchberg.....47·25n 9·03e
Klausenpass.....46·52n 8·51e
Kleinlützel.....47·26n 7·25e
Klosters.....46·54n 9·53e
Kloten.....47·27n 8·35e
Koblenz.....47·37n 8·14e
Köniz.....46·56n 7·25e
Konolfingen.....46·53n 7·38e
Kreuzlingen.....47·39n 9·11e
Kriens.....47·02n 8·17e
Küblis.....46·55n 9·47e
Küssnacht-am-Rigi.....47·05n 8·27e
Küttigen.....47·25n 8·03e
La Brévine.....46·59n 6·36e
La Chaux de Fonds.....47·06n 6·50e
Lachen.....47·12n 8·51e
Lai, see Lenzerheide
Lancy.....46·11n 6·07e
Landquart.....46·58n 9·33e
La Neuveville....47·04n 7·06e
Langenbruck....47·21n 7·46e
Langenthal.....47·13n 7·47e
Langnau.....46·57n 7·47e
Langwies.....46·49n 9·43e
Lantsch.....46·41n 9·34e
La Punt.....46·35n 9·55e
La Roche.....46·42n 7·08e
La Sagne.....47·03n 6·48e
La Sarraz.....46·40n 6·31e
La Tour de Peilz.46·27n 6·49e
Latterbach.....46·40n 7·35e
Läufelfingen.....47·24n 7·51e
Laufen.....47·25n 7·30e
Laufenburg.....47·33n 8·04e
Laupen.....46·54n 7·14e
Lausanne.....46·31n 6·38e
Lauterbrunnen ..46·36n 7·55e
Lavin.....46·46n 10·06e
Le Brassus.....46·35n 6·13e
Le Châble.....46·05n 7·12e
Le Châtelard.....46·04n 6·58e
Le Locle.....47·03n 6·45e
Lengnau.....47·11n 7·22e
Lenk.....46·28n 7·27e
Le Noirmont.....47·13n 6,58e
Lenzburg.....47·23n 8·11e
Lenzerheide (Lai).....46·44n 9·33e
Le Pont.....46·40n 6·20e
Le Prese.....46·18n 10·04e
Les Diablerets...46·21n 7·10e
Les Haudères....46·05n 7·31e
Les Marécottes ..46·07n 7·00e
Les Mosses.....46·24n 7·07e
Les Ponts de Martel.....46·54n 6·41e
Les Verrières....46·54n 6·30e
Leuk.....40·19n 7·38e
Leukerbad.....46·23n 7·38e
Leysin.....46·21n 7·01e
Lichtensteig.....47·19n 9·05e
Liestal.....47·29n 7·44e
Linthal.....46·55n 9·00e
L'Isle.....46·37n 6·25e
Littau.....47·03n 8·16e
Locarno.....46·10n 8·48e
Lucens.....46·42n 6·50e
Lugano.....46·01n 8·58e
Lungern.....46·47n 8·10e
Lumbrein.....46·41n 9·08e
Lutry.....46·30n 6·41e
Lützelflüh.....47·00n 7·41e
Luzern (Lucerne).47·03n 8·18e
Lyss.....47·04n 7·18e
Maggia.....46·15n 8·42e
Magglingen.....47·09n 7·14e
Maienfeld.....47·00n 9·32e
Maloja.....46·24n 9·41e
Malvaglia.....46·25n 8·59e
Männedorf.....47·15n 8·42e
Marbach.....46·52n 7·55e
Märstetten.....47·36n 9·04e
Martigny.....46·06n 7·04e
Martina.....46·53n 10·30e
Mauensee.....47·10n 8·04e
Meilen.....47·16n 8·38e
Meiringen.....46·43n 8·12e
Melchtal.....46·50n 8·17e
Melide.....45·57n 8·57e
Mels.....47·03n 9·25e
Mendrisio.....45·52n 8·59e
Menziken.....47·14n 8·12e
Merishausen.....47·45n 8·37e
Mesocco.....46·23n 9·14e
Meyrin.....46·14n 6·05e
Möhlin.....47·34n 7·51e
Mollis.....47·05n 9·04e
Montana.....46·18n 7·29e
Montbovon.....46·29n 7·03e
Montfaucon.....47·17n 7·03e
Monthey.....46·15n 6·57e
Montreux.....46·26n 6·55e
Morcote.....45·56n 8·55e
Morges.....46·31n 6·30e
Moudon.....46·40n 6·48e
Moutier.....47·17n 7·23e
Mulegns.....46·33n 9·37e
Mülenen.....46·38n 7·42e
Münchenbuchsee.....47·01n 7·27e
Münchenstein...47·31n 7·37e
Münsingen.....46·53n 7·34e
Münster.....46·29n 8·16e
Münsterlingen...47·38n 9·14e
Muotathal.....46·59n 8·46e
Murgenthal.....47·16n 7·50e
Muri.....46·56n 7·29e
Muri.....47·16n 8·21e
Mürren.....46·34n 7·54e
Murten.....46·56n 7·07e
Müstair.....46·37n 10·27e
Muttenz.....47·32n 7·39e
Näfels.....47·06n 9·04e
Naters.....46·20n 7·59e
Nendaz.....46·11n 7·18e
Nesslau.....47·13n 9·13e
Netstal.....47·03n 9·03e
Neuchâtel.....46·59n 6·56e
Neuenegg.....46·54n 7·18e

MAP INDEX (continued)

Neuhausen.....47·41n 8·37e
Neu Sankt Johann.....47·14n 9·12e
Nidau.....47·07n 7·14e
Niederbipp.....47·16n 7·39e
Niederurnen.....47·07n 9·03e
Niederwald.....46·26n 8·12e
Niederweningen.47·30n 8·23e
Nyon.....46·23n 6·14e
Oberägeri.....47·08n 8·37e
Oberdiessbach..46·51n 7·38e
Oberengstringen 47·25n 8·28e
Oberhofen.....46·44n 7·40e
Oberried.....46·44n 7·58e
Oberriet.....47·19n 9·33e
Oberuzwil.....47·26n 9·08e
Oberwald.....46·32n 8·21e
Oensingen.....47·17n 7·44e
Oftringen.....47·19n 7·56e
Olivone.....46·32n 8·57e
Ollon.....46·18n 7·00e
Olten.....47·21n 7·54e
Orbe.....46·43n 6·32e
Oron-la-Ville.....46·34n 6·50e
Orsières.....46·02n 7·09e
Osogna.....46·18n 9·00e
Ostermundigen..46·58n 7·29e
Palézieux.....46·33n 6·50e
Parpan.....46·46n 9·33e
Payerne.....46·49n 6·56e
Peseux.....46·59n 6·53e
Pfäfers.....46·59n 9·30e
Pfäffikon.....47·12n 8·46e
Pfäffikon.....47·22n 8·47e
Pfaffnau.....47·14n 7·54e
Pfyn.....47·36n 8·57e
Pieterlen.....47·11n 7·20e
Piotta.....46·31n 8·40e
Plaffeien.....46·44n 7·17e
Platta.....46·40n 8·51e
Pontresina.....46·28n 9·53e
Porrentruy.....47·25n 7·05e
Poschiavo.....46·18n 10·04e
Posieux.....46·46n 7·06e
Pratteln.....47·31n 7·42e
Preda.....46·36n 9·46e
Prilly.....46·32n 6·36e
Promontogno....46·21n 9·34e
Pully.....46·31n 6·39e
Rafz.....47·37n 8·32e
Ramosch.....46·50n 10·22e
Randa.....46·07n 7·47e
Rapperswil.....47·14n 8·50e
Raron.....46·19n 7·48e
Regensberg.....47·29n 8·26e
Regensdorf.....47·26n 8·28e
Reichenau.....46·49n 9·24e
Reigoldswil.....47·24n 7·41e
Reinach.....47·15n 8·11e
Reinach.....47·30n 7·35e
Reineck.....47·28n 9·34e
Renens.....46·32n 6·35e
Rheinfelden.....47·33n 7·48e
Riaz.....46·38n 7·04e
Richisau.....47·02n 8·54e
Richterswil.....47·13n 8·42e
Riddes.....46·10n 7·13e
Riederalp.....46·23n 8·01e
Riehen.....47·35n 7·39e
Riggisberg.....46·48n 7·29e
Rolle.....46·28n 6·20e
Romainmôtier....46·42n 6·29e
Romanshorn.....47·34n 9·22e
Romont.....46·42n 6·55e
Rona.....46·34n 9·38e
Ronco.....46·08n 8·44e
Root.....47·07n 8·23e
Rorschach.....47·29n 9·30e
Rossa.....46·22n 9·08e
Röthenbach.....46·51n 7·45e
Rothrist.....47·19n 7·53e
Rougemont.....46·29n 7·12e
Roveredo.....46·14n 9·08e
Rue.....46·37n 6·50e
Rümlang.....47·27n 8·32e
Rüti.....47·16n 8·51e
Saanen.....46·29n 7·16e
Saanenmöser....47·32n 7·18e
Saas Almenau...46·07n 7·58e
Saas Fee.....46·07n 7·55e
Saas Grund.....46·08n 7·56e
Sachseln.....46·52n 8·15e
Saignelégier.....47·15n 7·00e
Saint-Aubin.....46·54n 6·47e
Saint-Blaise.....47·01n 6·59e
Saint-Cergue....46·27n 6·09e
Saint-Croix.....46·49n 6·31e
Saint-Gingolph...46·24n 6·52e
Saint-Imier.....47·09n 7·00e
Saint-Luc.....46·13n 7·36e
Saint-Maurice....46·13n 7·00e
Saint-Prex.....46·29n 6·28e
Saint-Ursanne...47·22n 7·10e
Salavaux.....46·55n 7·02e
Samedan.....46·33n 9·52e
Samnaun.....46·56n 10·22e
San Bernardino..46·28n 9·12e
San Carlo.....46·25n 8·32e
San Gion.....46·38n 8·50e
Sankt Antönien..46·58n 9·49e
Sankt Gallen.....47·25n 9·23e
Sankt Margrethen.....47·27n 9·36e
Sankt Moritz.....46·30n 9·50e
Sankt Niklaus....46·11n 7·48e
Sankt Peterzell...47·19n 9·11e
Santa Maria.....46·36n 10·24e
Sargans.....47·03n 9·26e
Sarnen.....46·54n 8·15e
Sattel.....47·05n 8·42e
Savièse.....46·16n 7·20e
Savognin.....40·36n 9·36e
Saxon.....46·09n 7·11e
Schaffhausen....47·42n 8·38e
S-Chanf.....42·36n 9·59e
Schangnau.....46·50n 7·52e
Schiers.....46·59n 9·41e
Schinznach Bad..47·27n 8·10e
Schleitheim.....47·45n 8·29e
Schlieren.....47·24n 8·27e
Schöftland.....47·18n 8·03e
Schöllenen.....46·40n 8·35e
Schuls, see Scuol
Schönenwerd....47·22n 8·00e
Schüpfheim.....46·57n 8·01e
Schwanden.....47·00n 9·04e
Schwarzenburg..46·49n 7·21e
Schwarzsee.....46·40n 7·20e
Schwyz.....47·02n 8·40e
Scuol (Schuls)...46·48n 10·18e
Sedrun.....46·41n 8·46e
Seeberg.....47·09n 7·40e
Seen.....47·29n 8·46e
Seengen.....47·19n 8·13e
Seewis.....47·00n 9·32e
Sempach.....47·08n 8·11e
Sennwald.....47·16n 9·30e
Seon.....47·21n 8·10e
Sertig-Dörfli.....46·44n 9·51e
Suezach.....47·32n 8·44e
Sevelen.....47·07n 9·29e
Siebnen.....47·11n 8·54e
Sierre.....46·18n 7·32e
Signau.....46·55n 7·43e
Sigriswil.....46·43n 7·42e
Sils.....46·22n 9·46e
Silvaplana.....46·26n 9·47e
Sins.....47·11n 8·23e
Sion.....46·14n 7·21e
Sirnach.....47·28n 9·00e
Sisikon.....46·57n 8·42e
Sissach.....47·28n 7·49e
Soazza.....46·22n 9·13e
Solothurn.....47·13n 7·32e
Somvix.....46·44n 8·56e
Sonceboz.....47·11n 7·11e
Sonogno.....46·21n 8·47e
Sörenberg.....46·50n 8·03e
Sottens.....46·39n 6·44e
Speicher.....47·24n 9·27e
Spiez.....46·41n 7·39e
Splügen.....46·33n 9·20e
Stäfa.....47·15n 8·44e
Stalden.....46·14n 7·52e
Stammheim.....47·38n 8·47e
Stans.....46·57n 8·22e
Stansstad.....46·59n 8·20e
Steckborn.....47·40n 8·55e
Steffisburg.....46·47n 7·39e
Steg.....47·21n 8·56e
Stein.....47·33n 7·58e
Stein am Rhein..47·40n 8·51e
Stöckalp.....46·48n 8·17e
Suhr.....47·22n 8·05e
Sumiswald.....47·02n 7·45e
Sursee.....47·10n 8·06e
Susch.....46·46n 10·04e
Tarasp.....46·38n 10·25e
Tavanasa.....46·45n 9·04e
Tavannes.....47·13n 7·12e
Täuffelen.....47·04n 7·12e
Tenigerbad.....46·42n 8·57e
Tesserete.....46·04n 8·58e
Teufen.....47·23n 9·23e
Thalkirch.....46·38n 9·16e
Thalwil.....47·17n 8·34e
Thayngen.....47·45n 8·42e
Thun.....46·45n 7·37e
Thusis.....46·42n 9·26e
Tiefencastel....46·40n 9·35e
Trachselwald....47·01n 7·45e
Tramelan.....47·13n 7·06e
Trasadingen.....47·40n 8·26e
Trimbach.....47·22n 7·54e
Trin.....46·50n 9·22e
Trogen.....47·24n 9·28e
Trun.....46·45n 8·58e
Tschamut.....46·40n 8·42e
Turbenthal.....47·27n 8·51e
Turtmann.....46·18n 7·41e
Twann.....47·06n 7·10e
Unterägeri.....47·08n 8·35e
Unterbäch.....46·17n 7·48e
Unterschächen..46·52n 8·47e
Unterseen.....46·41n 7·51e
Unterterzen.....47·07n 9·15e
Unterwasser.....47·12n 9·19e
Urnäsch.....47·19n 9·17e
Uster.....47·21n 8·43e
Utzenstorf.....47·08n 7·33e
Uznach.....47·14n 9·00e
Valangin.....47·01n 6·54e
Vallorbe.....46·43n 6·22e
Vals Platz.....46·37n 9·11e
Vättis.....46·55n 9·27e
Vaulruz.....46·37n 6·59e
Verbier.....46·06n 7·13e
Vergeletto.....46·14n 8·36e
Vernayaz.....46·08n 7·02e
Vernier.....46·13n 6·06e
Versoix.....46·16n 6·10e
Vésenaz.....46·14n 6·12e
Vevey.....46·28n 6·51e
Vex.....46·13n 7·24e
Vicosoprano.....46·22n 9·37e
Villars.....46·18n 7·04e
Villeneuve.....46·24n 6·55e
Villmergen.....47·21n 8·15e
Vinadi.....46·55n 10·29e
Vira.....46·08n 8·51e
Visperterminen..46·15n 7·54e
Vissoie.....46·13n 7·36e
Vitznau.....47·01n 8·29e
Vouvry.....46·20n 6·53e
Vrin.....46·39n 9·06e
Wädenswil.....47·14n 8·40e
Wald.....47·17n 8·55e
Waldstatt.....47·21n 9·17e
Walenstadt.....47·07n 9·19e
Wallisellen.....47·25n 8·36e
Wagen an der Aare.....47·14n 7·39e
Wängi.....47·30n 8·57e
Wangs.....47·02n 9·26e
Wasen.....47·03n 7·48e
Wassen.....46·42n 8·36e
Wattenwil.....46·46n 7·30e
Wattwil.....47·18n 9·06e
Weggis.....47·02n 8·26e
Weinfelden.....47·34n 9·06e
Weissenburg....46·39n 7·28e
Weisstannen....46·59n 9·21e
Wengen.....46·36n 7·56e
Wettingen.....47·28n 8·19e
Wetzikon.....47·19n 8·47e
Wiedlisbach.....47·15n 7·39e
Wiesen.....46·43n 9·43e
Wil.....47·27n 9·03e
Wildegg.....47·25n 8·11e
Wildhaus.....47·12n 9·22e
Willisau.....47·07n 8·00e
Windisch.....47·29n 8·13e
Winterthur.....47·30n 8·43e
Wohlen.....47·21n 8·17e
Wolfenschiessen.46·55n 8·24e
Wolhusen.....47·04n 8·04e
Worb.....46·56n 7·34e
Würenlingen.....47·32n 8·16e
Wynigen.....47·06n 7·40e
Yverdon.....46·47n 6·39e
Yvonand.....46·48n 6·45e
Zell.....47·09n 7·55e
Zermatt.....46·01n 7·45e
Zernez.....46·43n 10·05e
Zillis.....46·38n 9·27e
Zizers.....46·56n 9·34e
Zofingen.....47·18n 7·57e
Zollikofen.....47·00n 7·28e
Zollikon.....47·20n 8·35e
Zuchwil.....47·12n 7·33e
Zug.....47·10n 8·31e
Zurzach.....47·35n 8·18e
Zuoz.....46·35n 9·58e
Zürich.....47·23n 8·32e
Zweisimmen.....46·33n 7·22e
Zwingen.....47·26n 7·32e

Physical features and points of interest

Aare, *river*.....47·37n 8·13e
Aareschlucht, *scenic area*.....46·44n 8·12e
Adula (Rhein-waldhorn), *mountain*.....46·30n 9·02e
Ägerisee, *lake*....47·07n 8·38e
Albula, *river*.....46·42n 9·27e
Albulapass, *pass*.....46·35n 9·50e
Aletschhorn, *mountain*.....46·28n 8·00e
Alps, *mountains*..46·30n 8·30e
Bagnes, Vallée de, *valley*.....46·03n 7·18e
Baldegger See, *lake*.....47·12n 8·16e
Balmhorn, *mountain*.....46·26n 7·41e
Bella Tola, *mountain*.....46·15n 7·39e
Berner Alpen, *mountains*.....46·30n 7·30e
Bernina, *mountains*.....46·22n 9·50e
Bernina, Passo del (Bernina-pass), *pass*.....46·24n 10·03e
Bernina, Piz, *peak*.....46·21n 9·51e
Berninapass (Passo del Bernina), *pass*..46·24n 10·03e
Bieler See, *lake*..47·05n 7·10e
Bietschhorn, *mountain*.....46·24n 7·51e
Birse Birs, *river*..47·22n 7·22e
Blanche, Dent, *mountain*.....46·03n 7·36e
Blenio, Val, *valley*.....46·27n 8·58e
Blinnenhorn, *mountain*.....46·26n 8·19e
Blümlisalp, *mountain*.....46·30n 7·47e
Bodensee (Lake Constance; Lake Konstanz).....47·35n 9·25e
Breithorn, *mountain*.....45·56n 7·45e
Brienzer Rothorn, *mountain*.....46·48n 8·04e
Brienzer See, *lake*.....46·43n 7·57e
Broye, *river*.....46·55n 7·02e
Brünigpass, *pass*.46·46n 8·09e
Buin, Piz, *peak*...46·50n 10·08e
Calanca, Val, *valley*.....46·22n 9·07e
Campo Tencia, Pizzo, *peak*.....46·26n 8·43e
Ceneri, Monte, *pass*.....46·08n 8·55e
Chillon, *castle*....46·25n 6·56e
Churfirsten, *mountain*.....47·08n 9·17e
Constance, Lake, see Bodensee
Dammastock, *mountain*.....46,39n 8·25e
Dents du Midi, *mountains*.....46·10n 6·56e
Diveria, *river*.....46·12n 8·09e
Dix, Lac des, *lake*.....46·03n 7·24e
Dom, *mountain*...46·06n 7·50e
Domleschg, *valley*.....46·44n 9·28e
Doubs, *river*.....47·20n 7·03e
Dufourspitze, *peak*.....45·55n 7·52e
Eiger, *mountain*..46·35n 8·00e
Emme, *river*.....47·13n 7·34e
Emmental, *valley*.46·56n 7·45e
En (Inn), *river*....46·55n 10·28e
Engadin, Ober-, *valley*.....46·37n 9·58e
Engadin, Unter-, *valley*.....46·50n 10·20e
Engstli Gental, *valley*.....46·30n 7·47e
Entlebuch, *valley*.....46·58n 8·00e
Finsteraarhorn, *mountain*.....46·32n 8·08e
Fluchthorn, *mountain*.....46·53n 10·15e
Flüelapass, *pass*..46·45n 9·57e
Forclaz, Col de la, *pass*.....46·04n 7·00e
Franches-Montagnes, *mountains*.....47·15n 7·00e
Fuorn, Passo del (Ofenpass), *pass*.....46·37n 10·15e
Furkapass, *pass*..46·34n 8·25e
Geneva, Lake (Lac Léman)....46·25n 6·30e
Givrine, Col de la, *pass*.....46·27n 6·05e
Glâne, *river*.....46·47n 7·08e
Glarner Alpen, *mountains*.....46·55n 9·00e
Glärnisch, *mountain*.....47·00n 9·00e
Glatt, *river*.....47·34n 8·28e
Goms, *valley*.....46·27n 8·15e
Gornergrat, *scenic area*.....45·49n 7·47e
Gotthard, Massif, *mountains*.....46·31n 8·31e
Grand Combin, *mountain*.....45·56n 7·18e
Grande Dixence, *dam*.....46·05n 7·28e
Grand Muveran, *mountain*.....46·14n 7·08e
Grand Saint-Bernard, Col du, *pass*.....45·50n 7·10e
Greifensee, *lake*..47·21n 8·41e
Grimselpass, *pass*.....46·34n 8·21e

MAP INDEX (continued)

Grimselsee, *lake*..46·34n 8·18e
Gross Windgällen, *mountain*.....46·49n 8·44e
Gruyère, Lac de la, *lake*.....46·38n 7·06e
Hallwiler See, *lake*.....47·18n 8·13e
Haslital, *valley*...46·42n 8·10e
Hinterrhein, *river*.....46·49n 9·25e
Inn, see En
Jaunpass, *pass*...46·36n 7·20e
Joux, Lac de, *lake*.....46·38n 6·18e
Julierpass, *pass*..46·28n 9·44e
Jungfrau, *mountain*.....46·32n 7·58e
Jungfraujoch, *rail terminal*.....46·33n 7·58e
Jura, *mountains*..46·45n 6·30e
Kander, *river*.....46·43n 7·38e
Kesch, Piz, *peak*.46·38n 9·52e
Kleine Emme, *river*.....47·04n 8·17e
Klöntaler See, *lake*.....47·02n 8·58e
Königsfelden, *church*.....47·29n 8·14e
Konstanz, Lake, see Bodensee
La Berra, *mountain*.....46·41n 7·11e
Landquart, *river*..46·58n 9·32e
Lauerzer See, *lake*.....47·02n 8·36e
Le Chasseral, *mountain*.....47·08n 7·03e
Le Chasseron, *mountain*.....46·51n 6·33e
Léman, Lac, see Geneva, Lake
Le Moléson, *mountain*.....46·33n 7·01e
Leone, Monte, *mountain*.....46·15n 8·06e
Lepontine, Alpi, *mountains*.....46·25n 8·50e
Les Diablerets, *mountain*.....46·19n 7·12e
Leventina, Valle, *valley*.....46·25n 9·52e
Limmat, *river*.....47·25n 8·25e
Linth, *river*.....47·07n 9·07e
Linthkanal, *canal*.....47·13n 8·57e
Lorze, *river*.....47·15n 8·25e
Lötschental, *valley*.....46·25n 7·50e
Lucerne, Lake, see Vierwald Stätter See
Lucomagno, Passo del (Lukmanïer-pass), *pass*.....46·34n 8·48e
Lugano, Lago di, *lake*.....45·58n 8·56e
Lukmanïerpass (Passo del Lucomagno), *pass*.....46·34n 8·48e
Madonna del Sasso, *monastery*.....46·11n 8·33e
Maggia, *river*.....46·09n 8·48e
Maggia, Valle, *valley*.....46·17n 8·40e
Maggiore, Lago, *lake*.....45·59n 8·47e
Malojapass, *pass*.46·23n 9·41e
Matterhorn, *mountain*.....45·59n 7·43e
Mattertal, *valley*..46·10n 7·49e
Mauvoisin, *dam*..46·00n 7·25e
Medel, Val, *valley*.....46·37n 8·50e
Mesolcina, Valle, *valley*.....46·20n 9·10e
Mittelland, *physical region*..47·15n 8·00e
Moesa, *river*.....46·13n 9·03e
Mönchsgrat, *mountain*.....46·34n 7·59e
Morat, Lac de (Murtensee), *lake*.....46·55n 7·05e
Morteratsch, Piz, *peak*.....46·22n 9·51e
Mosses, Col des, *pass*.....46·24n 7·06e
Murtensee, see Mora, Lac de
Napf, *mountain*..47·00n 7·56e
Neuchâtel, Lac de, *lake*.....46·52n 6·50e
Niesen, *mountain*.....46·39n 7·39e
Oberalppass, *pass*.....46·39n 8·40e
Oberalpstock, *mountain*.....46·40n 8·42e
Oberhauenstein, *pass*.....47·21n 7·56e
Obre, *river*.....46·33n 6·09e
Ofenpass (Passo del Fuorn), *pass*.....46·37n 10·15e
Pennine, Alpi, *mountains*.....46·05n 7·50e
Pfäffikersee, *lake*.....47·21n 8·48e
Pierre Pertuis, de, *pass*.....47·12n 7·11e
Pilatus, *mountain*.....46·59n 8·15e
Pillon, Col du, *pass*.....46·22n 7·13e
Piz Lunghin, *peak*.....46·28n 9·35e
Poschiavino, *pass*.....46·14n 10·07e
Poschiavino, *river*.....46·12n 10·10e
Prättigau, *valley*..46·55n 9·45e
Rabiusa, *river*....46·48n 9·20e
Reuss, *river*.....47·28n 8·14e
Rhaetian Alps, *mountains*.....46·30n 10·00e
Rhätikon, *mountains*.....47·03n 9·40e
Rheinfall, *waterfall*.....47·41n 8·38e
Rheinwald, *valley*.....46·32n 9·17e
Rheinwaldhorn (Adula), *mountain*.....46·30n 9·02e
Rhine, *river*.....47·37n 7·35e
Rhône, *river*.....46·09n 5·57e
Rickenpass, *pass*.....47·14n 9·02e
Rigi, *mountain*.....47·05n 8·30e
Riviera, *valley*....46·15n 8·58e
Rosa, Monte, *mountain*.....45·55n 7·53e
Rotten, *river*.....46·17n 7·33e
Saane, *river*.....46·59n 7·16e
Saastal, *valley*....46·10n 7·56e
Safiental, *valley*..46·40n 9·18e
Saint-Imier, Vallon de, *valley*.....47·10n 7·00e
Samnaungruppe, *mountains*.....47·00n 10·25e
San Bernardino, Passo del, *pass*.46·30n 9·11e
Sankt Gotthard (Passo del San Gottardo), *pass*..46·33n 8·34e
Santa Maria, Giogo di (Wormser Joch), *pass*.....46·34n 10·25e
Säntis, *mountain*.47·15n 9·21e
Sardona, Piz, *peak*.....46·55n 9·15e
Sarine, *river*.....46·54n 7·14e
Sarner See, *lake*..46·52n 8·13e
Schanfigg, *valley*.....46·51n 9·38e
Schiffenensee, *lake*.....46·50n 7·10e
Schöllenen Gorge.....46·36n 8·31e
Sempacher See, *lake*.....47·09n 8·09e
Sense, *river*.....46·54n 7·14e
Sihl, *river*.....47·23n 8·32e
Sihlsee, *lake*.....47·07n 8·47e
Siluretta, *mountain*.....46·50n 10·15e
Simme, *river*.....46·41n 7·38e
Simmental, *valley*.....46·37n 7·25e
Simplonpass, *pass*.....46·15n 8·02e
Sitter, *river*.....47·29n 9·14e
Sorne, *river*.....47·22n 7·22e
Spluga, Passo dello, *pass*.....46·30n 9·20e
Suhr, *river*.....47·25n 8·04e
Surses, *valley*....46·34n 9·38e
Sustenhorn, *mountain*.....46·42n 8·28e
Sustenpass, *pass*.....46·44n 8·27e
Suze, *river*.....47·08n 7·14e
Thielle, *river*.....47·03n 7·05e
Thunersee, *lake*..46·40n 7·45e
Thur, *river*.....47·36n 8·35e
Ticino, *river*.....46·11n 8·51e
Titlis, *mountain*..46·47n 8·25e
Tödi, *mountain*...46·49n 8·56e
Toggenburg, *valley*.....47·45n 9·10e
Töss, *river*.....47·33n 8·33e
Travers, Val de, *valley*.....46·57n 6·38e
Tresa, *river*.....45·59n 8·47e
Unterhauenstein, *pass*.....47·23n 7·52e
Urirotstock, *mountain*.....46·52n 8·33e
Urseren Valley...46·35n 8·32e
Vadret, Piz, *peak*.....46·41n 9·57e
Valserrhein, *river*.....46·42n 9·10e
Valsertal, *valley*..46·37n 9·10e
Verzasca, *river*...46·09n 8·52e
Via Mala, *gorge*..46·40n 9·26e
Vierwald Stätter See (Lake Lucerne).....47·00n 8·28e
Vorderrhein, *river*.....46·49n 9·25e
Wägitaler See, *lake*.....47·06n 8·55e
Walensee, *lake*...47·07n 9·12e
Weissenstein, *mountain*.....47·15n 7·31e
Weissfluch, *mountain*.....46·50n 9·48e
Weisshorn, *mountain*.....46·06n 7·42e
Wetterhorn, *mountain*.....46·39n 8·08e
Wigger, *river*.....47·18n 7·53e
Wildhorn, *mountain*.....46·21n 7·22e
Wildstrubel, *mountain*.....46·24n 7·32e
Wohlensee, *lake*..46·58n 7·20e
Wormser Joch (Giogo di Santa Maria), *pass*.....46·34n 10·25e
Zuger See, *lake*..47·08n 8·30e
Zürichsee, *lake*...47·13n 8·45e

site of the Dufourspitze (15,200 feet [4,634 metres]), the highest point in Switzerland, and the Matterhorn (14,691 feet [4,478 metres]). The northern and southern Swiss Alps are separated by the trough formed by the Rhône and Upper Rhine valleys. The narrowest portion of the trough, the Urseren Valley, lies between two crystalline central massifs, the Gotthard and the Aarmassif.

The Alps' role as the European watershed is most apparent in the central Alpine region of Switzerland, where the different chains meet; from there, the Rhône River flows west, the Rhine River east, the Ticino River south, and the Reuss River north. The fundamental Alpine source point, however, is located in the Upper Engadine at Piz Lunghin, from which streams flow toward the North, the Adriatic, and the Black seas.

In the central Alpine region lies the Gotthard route, the first and shortest north–south passage through the mountains and an important European connection; it was opened around 1225 with the construction of a bridge in the Schöllenen Gorge. The gorge traverses the northern chain, while the southern range is crossed by the St. Gotthard Pass at an altitude of 6,916 feet. The 9.3-mile-long St. Gotthard Tunnel through the mass was opened in 1882.

Between the Jura and the main Alpine ranges lies the hilly Mittelland accounting for 30 percent of the country and enclosed by the two mountain ranges and the two largest lakes, Lake Geneva (Lac Léman) in the west and Lake Constance (Bodensee) in the east. The Mittelland, where the greatest part of the population is concentrated, supports Swiss economic life and guarantees the country's industrial and international importance. As a result, it is becoming highly urbanized, although even in urban Switzerland the feeling of security and rootedness remains alive, symbolized by the mountain peaks that can be seen on the horizon from every point on the plateau.

Climate. The waters may flow away from the Alps in all directions, but the climates of Europe converge upon the Swiss mountains. The Alps are essentially the dividing line between the different climates of central Europe, but Switzerland as a whole combines the four main European climatic types to form a single transitional climate, influenced on the west by the Atlantic climate, which is affected by the Gulf Stream; on the east by that of the eastern European continent; on the south by the Mediterranean climate; and on the north by the middle European cyclonal climate.

Wind systems. Prevailing winds are from the west, but, in the valleys, air currents are channelled into particularly frequent or violent local winds such as the Bise, the northeast wind characteristic of Geneva. The Föhn, a brisk south wind that occurs in the large transverse valleys of the northern Alps, makes it possible to grow sweet

chestnut, laurel, and tobacco on the northern slopes of the Alps.

Precipitation. Since rainfall tends to increase in direct proportion to altitude, the precipitation map corresponds very closely to the relief map; and, because of the marked variation in reliefs, differences in precipitation within short linear distances are often very great. For example, Sankt Gallen, at 2,180 feet, has an average annual precipitation of 52 inches (1,318 millimetres), while that of the Säntis, 8,200 feet but only 32 miles away, is 114 inches (2,904 millimetres). The average annual precipitation of three-quarters of the country exceeds 40 inches, varying amounts of which fall as snow, which, in turn, characterizes and explains the dense variety of weather phenomena. In the southern town of Lugano (905 feet), which has a Mediterranean climate, only 5 percent of the precipitation is in the form of snow; in Zürich (1,620 feet) 11 percent is snow; and on the Säntis 72 percent is snow. At altitudes above 11,500 feet, all precipitation consists of snow, which becomes impacted into perpetual snowfields and glaciers; the snow line is at about 9,200 feet in the northern Alps and at 10,800 feet in the southern Alps of the Valais.

Dry areas

In the mountains of the interior, distinct dry pockets occur, the most famous being the Rhône Valley in the Valais, closely encircled by the highest (13,000-foot) mountain groups. Precipitation is also slight on the slopes with southern exposure near the cantonal capital of Sion (1,800 feet), but extensive irrigation is possible since the area is surrounded by large snowfields and by glaciers that extend far down into the valleys. The rarefied, germ-free, and unexpectedly dry air of such high-altitude towns as Davos (5,121 feet) and Arosa (5,987 feet) permits a more intense, broader spectrum solar irradiation, thus producing a radiant climate long famous in tuberculosis cures. At altitudes of 13,000 feet, precipitation levels rise to 160 inches, and the Mönchsgrat (13,460 feet) in the Jungfrau group has the highest average annual precipitation (163 inches, or 4,141 millimetres) in Switzerland, while Stalden in the encircled Valais has the minimum (21 inches, or 530 millimetres).

Skies and temperatures. The stable high-pressure weather conditions prevailing over central Europe and the Alps during the winter create cold air masses that often are covered by thick, gray layers of fog, a climatic phenomenon with widely varying consequences. The mouths of the northern Alpine valleys, the basins of the Jura Mountains, and the villages and cities of the low areas of the Mittelland are blanketed with fog for days and, often, for weeks on end, while towns located at high altitudes enjoy warm, brilliant, high-pressure conditions and the view of the glistening sea of fog below them. Temperature inversions between mountain and valley locations in close proximity can be quite pronounced, with the higher point having the higher temperature reading. Frequent temperature inversion is a primary reason why high-altitude Swiss sanatoriums remain occupied even during the winter and why, in the present century, the Alpine winter season has gained popularity in Europe for sports as well as for health cures.

Avalanches

With the increase in winter tourism, the study of avalanches has developed as a branch of Alpine climatology, and during threatening weather the research station near Davos publishes daily avalanche bulletins, mainly as a warning for tourists. The Alpine cantons have about 10,000 avalanches annually, more than 8,000 of them in February, March, and April. For centuries, village communes have used *Bannwälder* ("protective forests") on the mountain slopes as protection from these slides, since a 20- to 30-year-old forest can usually stop avalanches. Costly artificial structures protect the villages, highways, and Alpine paths.

Vegetation and animal life. The vegetation is derived from that of the four European climatic regions converging in Switzerland and from the varied relief. It includes the beeches and oaks of the Atlantic west; hornbeam and larch trees in the more continental east, predominantly in the Engadine and the dry Valais; extensive spruce forests in the northern sub-Alpine region; and chestnut groves in the south. The vegetation boundaries are several hundred metres higher in the south than in the north because of the generally southern exposure.

Crop cultivation at high altitudes

The Alpine pastures of the Monte Rosa region reach an altitude of 10,800 feet, and until recently the highest Alpine grainfields, which have fallen victim to the decline in Alpine agriculture, lay above Zermatt at an altitude of 6,900 feet. The highest vineyard of central Europe, at an altitude of 3,900 feet, is still located near Visperterminen in the Valais, although the largest vineyards are on the southern-exposed shore of Lake Geneva, on the sunny slope of the Rhône valley, along lakes Neuchâtel and Biel at the foot of the Jura, and in the northern Alpine valley of the Rhine, which is affected by the Föhn. On the western Mittelland, a considerable grain-producing area has developed on the sheltered side of the Jura Mountains, an area of scanty rainfall, while in the more humid eastern region, mainly in the cantons of Thurgau and Sankt Gallen, grass cultivation is combined with fruit growing.

Swiss animal life is mainly Alpine, but a mixture of species familiar to southern and north central Europe is also found. Animal life is protected, except during a brief annual hunting season. The Alpine tourist may be pleasantly surprised by the sight of the marmot, which lives near the snow line, and the chamois; large herds of the proud ibex, which had died out in the Swiss Alps and has since been carefully re-introduced, have again developed in several areas, especially in the Bernina region of Graubünden (canton). In the forests there are rabbits, foxes, badgers, squirrels, and numerous varieties of birds, while the lake and river trout are especially famous. Snakes and lizards are concentrated in the south, but insects, in profuse variety, are diffused throughout the country.

Patterns of human settlement. The diversity of geomorphology, climate, and plant distribution in Switzerland provides a wide variety of bases for the formation of settlements, a variety further enhanced by the country's central European location. Switzerland supplies more varied material for the study of the geography of settlements than does any other European country. Even the types of houses vary from one region to another, especially well-known examples being those of the Engadine, the Ticino, the Bernese region, the Jura, and northeastern Switzerland. In atmosphere, too, there are distinct variations: the German-speaking cantons and large cities, such as Bern, Zürich, and Basel, express an earnest and stern efficiency, whereas in the west, especially in Geneva and the canton of Vaud, the Latin influence prevails and life becomes less thrifty and more graceful. In the south, the Ticinese enjoy a more luxuriant climate, and this, too, is perhaps reflected in an easier life-style.

Rural communities. The valleys are the residential areas of Switzerland. Alpine villages extending along the base of slopes are a characteristic sight. Since the extensive river diversions made chiefly during the second half of the 19th century, villages often have expanded into the valley plains, where intensive farming is carried on. The isolation of portions of many valleys, such as those of the Rheinwald, Poschiavino, and Urseren, by barriers of resistant rock or by prehistoric landslides was an impetus to the formation of communes and of the early *Alpgenossenschaften* ("Alpine cooperatives").

Self-sufficiency of the rural communes

The modern network of communes, small but politically important units, stands out in the parallel relief of the Jura and the Alps, since the communes are generally laid out with their boundaries drawn from one mountain crest to the next. Every commune has all the basic living requirements, including pasture, forest, fertile valley bottom, and water. Terraced slopes characterize the sites of villages that serve as bases for "Alpine nomadism," the seasonal moving of livestock to or from the mountains.

Some villages, such as Guarda in the lower Engadine, have a reputation for picturesque beauty, while others, such as Montana in the Valais or Wengen in the Bernese Oberland, have developed into famous resorts and spas. Valley forks, where the traffic from two valleys combines, are attractive sites for settlement, as are places in the Alps such as Amsteg (canton of Uri), Silvaplana (the

Switzerland, Area and Population

	area*		population†	
	sq mi	sq km	1960 census	1970 census
Cantons (*kantone, cantons, cantoni*)‡				
Aargau	538	1,393	361,000	433,000
	542	1,404		
Appenzell Ausser-Rhoden§	94	243	49,000	49,000
Appenzell Inner-Rhoden§	66	172	13,000	13,000
	67	172		
Basel-Landschaft§	165	428	148,000	205,000
Basel-Stadt§	14	37	226,000	235,000
Bern	2,608	6,756	890,000	983,000
	2,659	6,887		
Fribourg	615	1,592	159,000	180,000
	645	1,670		
Genève	95	246	259,000	332,000
	109	282		
Glarus	261	675	40,000	38,000
	264	684		
Graubünden	2,739	7,095	147,000	162,000
	2,745	7,109		
Luzern	551	1,428	253,000	290,000
	577	1,494		
Neuchâtel	276	715	148,000	169,000
	308	797		
Nidwalden§	92	240	22,000	26,000
	106	274		
Obwalden§	185	480	23,000	25,000
	190	492		
Sankt Gallen	754	1,953	339,000	384,000
	778	2,016		
Schaffhausen	115	298	66,000	73,000
Schwyz	326	845	78,000	92,000
	350	908		
Solothurn	305	791	201,000	224,000
	305	791		
Thurgau	333	862	166,000	183,000
	389	1,006		
Ticino	1,056	2,736	196,000	245,000
	1,085	2,811		
Uri	407	1,055	32,000	34,000
	415	1,075		
Valais	2,012	5,212	178,000	207,000
	2,020	5,231		
Vaud	1,088	2,818	430,000	512,000
	1,240	3,211		
Zug	80	207	52,000	68,000
	92	239		
Zürich	639	1,656	952,000	1,108,000
	668	1,729		
Total Switzerland	15,418‖	39,932	5,429,000¶	6,270,000
	15,941‖	41,288		

*Where two figures are given, the first is the land area, the second the total area. †De jure. ‡In parentheses the term is given in German, French, and Italian, in the order named. §Demicanton; functions as a full canton and has the same legal prerogatives as a full canton. ‖Converted area figures do not add to total given because of rounding. ¶Figures do not add to total given because of rounding.
Source: Official government figures.

upper Engadine), and Gordola (near Locarno), where the entrance into an adjacent valley of higher altitude is made gradually. In the Mittelland, with its abundant lakes, delta villages are especially closely related to the environment; in recent decades villages have expanded toward each other and merged, creating population belts all along the lakes. Unfortunately, uncontrolled property speculation has permitted many characteristic, substantial village settlements to spread into the surrounding areas with very little architectural planning. The goals of local environmental protection are clashing with large-scale financial interests, a conflict that is becoming a matter of concern throughout Europe in the important case of Sankt Moritz (Graubünden).

Urban Switzerland. In the early 1970s, about 100 Swiss communities had more than 10,000 inhabitants and thus were classed as cities. Some had developed around monasteries, as did Sankt Gallen, or around Roman settlements, as in the case of Zürich; many developed from villages into market cities. Some, such as Fribourg (1157) and Bern (1191), were founded on a plan; their old sections, located within the protected areas formed by looped river bends, present a fascinating vision with their unified urban architecture. Each city is geographically unique, particularly those lying at the head of a lake, such as Zürich, Geneva, and Luzern, which were essentially harbour towns until the opening of the railroads (a status to some extent regained with the lakes' new importance to the tourist industry). A lake inlet, or the mouth of a river such as the Limmat, Rhône, or Reuss, and moraines and rings of mountains combine with the cities' houses and scenically situated cathedrals and minsters to form composite landscapes of nature and art. Many towns have sprung up around the site of a bridge; as a rule, these have two sections separated by the narrows of a river. Hill towns such as Regensberg and Gruyères, which were medieval fortified settlements with castles and the architecture of which developed through the late Gothic period, have a natural dominance that rather gives them the appearance of scenic ornaments, of crowns gleaming in the distance.

In Switzerland, isolated from Europe by mountainous barriers, location in relation to the traffic pattern has played an important role in urban development. Chur, Martigny, and Bellinzona are keys to a series of Alpine passes. Lugano, which, along with Locarno and Ascona, forms an extensive resort area benefitting from the abundant sunshine of the Mediterranean climate, is a gateway from the south to the Alpine region. Basel's location at the head of Rhine navigation has been of particular significance, since this was the basis of its early prominence as a city of scholars and of its present international importance as an industrial centre. Switzerland's eight largest cities in order of size in the early 1970s were Zürich, Basel, Geneva, Bern, Lausanne, Winterthur, Sankt Gallen, and Luzern.

THE PEOPLE OF SWITZERLAND

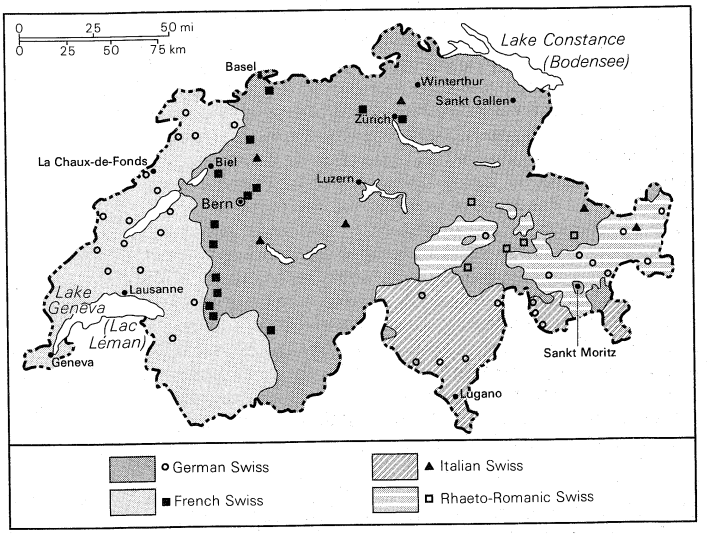
Etruscans, Raetians, Celts, Romans, and Germans have left their mark on Switzerland in the course of its historical evolution. The present population mainly shows traces of the Alpine, "Nordic," and southern Slav, or Dinaric, peoples. To survive as a cohesive unit, the disparate elements of the Swiss people have had to learn a mutual cooperation to protect the neutrality that has been their safeguard. Their outlook has been shaped largely by economic and political necessity, which has made them realistic, cautious, and prudent in accepting innovation and ingenious in the use of what resources they have. Their lives exhibit acumen, discipline, and thrift and a somewhat Germanic love of order and thoroughness allied with a Latin independence of spirit and an admiration for quality and craftsmanship that makes their products highly valued throughout the world.

Mixed character of the Swiss

Louis Renault—Photo Researchers, Inc.



Bern. The prison tower (centre) is a remnant of the medieval city walls.



Ethnic composition of Switzerland.

Language and religion. One of the distinctive features of the Swiss Confederation is the variety of its languages. About two-thirds of the total population speak German, one-fifth French, one-tenth Italian, and 1 percent Rhaeto-Romanic. Switzerland encompasses the area in which the major cultural regions of western continental Europe—the German, the French, and the Italian—come into contact. Its fourth language, Rhaeto-Romanic, is a linguistic relic preserved amid the mountainous regions between the Gotthard and the East Alps; it is subdivided into even smaller dialect regions. In 1938, a federal popular vote indicative of Swiss sentiment recognized Romansh as a fourth official national language, which was tantamount to a political declaration in favour of federalist diversity and in opposition to centralized power.

Switzerland also exhibits diversity in the area of religion, although conflicts over faith are rare and there are even churches in which Protestants and Catholics worship together. The great internal population shift caused by industrialization has also resulted in a greater mixture of the religions. The Swiss are divided almost equally into Protestants and Catholics, with a tiny Jewish community. The constitution of 1874 guaranteed full religious liberty but repeated the 1848 constitution's prohibition of settlement by Jesuits (members of the Roman Catholic Society of Jesus) and their affiliated societies in Switzerland. Such discrimination had become a point of controversy by the early 1970s and seemed likely to be abandoned in any revision of the constitution.

Demography. The doubling of the Swiss population over 100 years, to 5,000,000 in 1955, contributed to considerable emigration, especially from the economically unproductive Alpine regions, to places such as the United States. Of the more than 300,000 Swiss residing in foreign countries in 1971, nearly one-half possessed dual citizenship. Switzerland has 393 inhabitants per square mile, a high density in view of the large proportion of the country occupied by high mountains; if density is calculated for only the permanently settled territory, it reaches 628 inhabitants per square mile. The population has increased by about 15 percent during the 1960s, and future planning projects the population at 7,500,000 in the year 2000, with some projections rising as high as 10,000,000. Between 1965 and 1970 the Swiss birth rate dropped from 18.8 to 15.8 per 1,000 inhabitants, while the death rate remained almost constant at 9.1, reducing the excess of births over deaths, and consequently the rate of natural increase.

Population shifts in Switzerland are highly complex. In the less mountainous cantons the population has increased greatly, rising in the half canton of Basel-Landschaft, for example, by 38 percent since 1960; some communities near Zürich have increased to an unwholesome extent, as much as 535 percent. In contrast, the figures reported by the mountain cantons remain unchanged, with Glarus reporting a 5 percent decrease. Hundreds of mountain communes reported population losses, which result from the decline of agriculture and the shift to an industrial economy. The business boom since 1950 has created a labour shortage in spite of the presence of almost 700,000 foreign workers in the country.

THE NATIONAL ECONOMY

The Swiss economy has been molded by several specific factors: a paucity of raw materials; the pressure of urbanization on the limited amount of cultivable land; precipitation, which to a large extent determines the type and size of cultivation; the advantage of being centrally located on international trade routes; scenic beauty; and the inability of the domestic market to absorb the total output of a skilled and efficient population. Factors shaping the economy

These factors have given rise to industries such as shipping, freighting, banking, and insurance and to export industries requiring skill rather than large imports of raw materials, resulting in high-value, small-bulk industries such as precision-instrument manufacture. An important consequence of the traditional Swiss neutrality has been economic stability even in times of European war.

Between 1888 and 1970 the percentage of the population involved in basic raw-materials production dropped from 37 percent to 8 percent; that involved in manufacturing rose from 41 to 48 percent; and that involved in the services sector rose from 22 to 44 percent.

Components of the economy. *Agriculture.* Agriculture, which employs perhaps 4 percent of the total population, is characterized by fodder crops and grazing, which occupy nearly one-half of the total area. Grain, occupying only 6 percent of productive land, satisfies barely one-half of the domestic demand. The variation in soil quality within small areas produced by geological conditions and by the relief makes large-scale, single-crop farming impossible; instead, a particularly varied assortment of crops is grown in a limited space. Two-thirds of all farms combine grass and grain cultivation. The Valais, with its abundant sunshine and artificial irrigation, specializes in cultivating fruits, vegetables, and berries. In the southernmost canton of Ticino, a mixed Mediterranean agriculture has been attained, although it is endangered by the disruptions of oversettlement. Viticulture characterizes slopes along many lakes.

The primary agricultural pursuit, however, is cattle raising, for which Switzerland has become important throughout Europe. The income from dairying and cattle raising amounts to about one-half of all agricultural income.

Since the importance of forest to the ecology of large areas was recognized early, an exemplary forestation law forbids reduction of woodlands, which amount to nearly one-quarter of the total area of the country. Forests have become a main feature of the recreational areas located near cities.

Industry. Switzerland's transformation into an industrial state began during the second half of the 19th century; now, some 90 percent of exports consist of manufactured goods and chemicals, with some companies exporting 97 percent of their production. Lacking its own mineral and agricultural raw materials, however, Switzerland is forced to import raw materials, foodstuffs, and manufactured goods by costly land routes; raw materials, food, vegetable oils, and fuel account for nearly 25 percent of the imports. Only the outstanding quality of its products, ranging from motors, generators, and turbines to the atomic clock, enables Switzerland to remain competitive. The operations of the first European factory to produce aluminum by electronic methods have remained a model for Europe. The chemical–pharmaceutical industry, centred in Basel, is also a prime competitor in world trade. The chemical industry alone spends great sums on research, and many other industries also collaborate with the country's seven universities and with the Federal Institutes of Technology at Zürich and Lausanne. Many large firms are merging under the pressure of international competition, although the typical Swiss traits of diver-

sity and division into small units still make it possible for firms in small towns and villages to continue to supply products of worldwide importance.

Hydroelectric power

Although Switzerland lacks mineral power sources to run its industries, the extensive Alpine precipitation, glacial water storage, and the great range of altitudes make hydroelectric power widely available. The electrical industry has become an essential branch of the Swiss economy as a whole, with 400 large hydroelectric power plants currently in operation and numerous low-pressure plants situated on the lower courses of the rivers in the Mittelland. Major electrotechnical progress has occurred in the Alps, where large systems of tunnels and subterranean powerhouses have been constructed in suitable valleys. The highest dams in Europe, Mauvoisin (777 feet) and Grande Dixence (932 feet), have been erected in the Valais, where, on the Rhône, the highest rate of hydroelectric power consumption in Europe has been attained. By 1971, 95 percent of the hydroelectric energy worth harnessing for power plants was utilized, and Switzerland had begun to operate thermal power plants. In addition, several nuclear power plants were in the planning stage, under construction, or already in operation.

International trade. Switzerland has trade relations with 80 countries and is outranked only by Belgium–Luxembourg in total per capita imports and exports among the trading nations of the world. Its most important trading partners are the Federal Republic of Germany, France, Italy, the United States, and Great Britain.

Banking and insurance. Its central location has been a vital factor in making Switzerland one of the most important of the world's financial centres, with 475 banks, in 1969, on its territory financing the enormous freight stream that flows through the country. In the same context a large industry has grown up, specializing in reinsurance.

Tourism. The deficit in the balance of trade is to a large extent covered by income from tourism. During the Middle Ages, healing spas such as Baden, Bad Pfäfers, Leukerbad, and Rheinfelden flourished, while mountain-pass hospices were the predecessors of Alpine hotels. Since World War II, travel has increased at an explosive rate: hotels, guesthouses, and high-altitude sanatoriums counted 16,800,000 overnight stays in 1951, a figure that more than doubled by 1970, with 23,000,000 additional overnight stays in vacation apartments, youth hostels, and on campgrounds. Of the tourists, about 80 percent come from European countries, one-half of these from Switzerland itself. In 1971 the tourist industry as a whole employed almost as many as were engaged in farming.

Management of the economy. The long-standing tradition of direct democracy and federalism in Switzerland and the country's heavy dependence on foreign trade have given rise to an equally traditional suspicion of state intervention and to strong and constant support for worldwide free trade. Apart from the post office, telegraph, telephone, and main railway network, all utilities are privately owned or municipal enterprises, in some cases subsidized by cantonal governments. Even the hydroelectric plants are owned by private companies under federal or cantonal grants.

Areas of governmental involvement and restraint

Taxation. In matters of taxation, federal regulations extend mainly to customs duties and sales and defense taxes. Income taxes are cantonal responsibilities, and rates are fixed by decision of the voters of communal or cantonal parliaments.

Trade unions. Employer–employee relations have generally been good in recent years. The more than 30 trade unions have a combined membership of over 900,000. The Swiss Federation of Trade Unions, with almost 450,000 members, has Social Democratic tendencies. Since the Depression of the early 1930s, the unions have gradually denounced the use of strikes as economic and political weapons, and disputes are usually settled by arbitration.

Contemporary policies and problems. Although the Swiss have tended to reject state intervention in their free market economy, since World War II such intervention has been unavoidable, particularly in relation to the labour market. The sustained boom of the 1950s and later, with its very low rate of inflation, was only possible because the Swiss labour market was augmented by large numbers of foreign workers. By the 1970s, every third worker in industry was a foreigner. Social tensions became unavoidable, particularly where foreigners threatened to displace the Swiss, and the government was forced to restrict the inflow, thus for the first time breaching the liberal doctrines of Swiss economic policy.

This restriction of an important factor in an otherwise fast-expanding economy intensified inflationary pressures. Contemporary policy thus was directed to the difficult job of damping down the expansionary forces. Just as centralized bureaucracy was distrusted at home, so, in relation to European integration, the Swiss, though seeking a special arrangement with the European Economic Community, preferred the first type of integration offered by the European Free Trade Association, of which Switzerland was already a member.

TRANSPORTATION

Control of the most important Alpine passes and the ancient routeway through the Mittelland between the Rhône, Rhine, and Danube waterways has given Switzerland a key position in European transit traffic. The main artery of European trans-Alpine traffic, the St. Gotthard route, runs through Swiss territory.

Roads. In 1960 construction of a national highway network 1,143 miles in length was begun at Lake Geneva. In 1964 the 3.6-mile motor tunnel under the Great St. Bernard Pass was opened and, in 1967, the 4.1-mile San Bernardino Tunnel, thus ensuring important north–south Alpine connections to traffic throughout the year. Work was also started on a St. Gotthard highway tunnel.

Railways. The large-scale technical undertakings of highway construction were preceded by those of the railway system, which has a total track length of 5,833 miles and includes 608 tunnels, among them the 12.3-mile Simplon Tunnel, the longest in the world, and the famous winding tunnels of the Gotthard railway, by means of which altitude differences between valley levels are overcome. The entire Swiss railway system has been electrified, thus saving the country from dependence on coal imports. The Swiss Federal Railways, which constitute over half of the system, are operated by the federal government, while the remainder, including the numerous mountain railways, is distributed among over 75 private railroads, partially owned by the cantons and municipalities. The Vitznau-Rigi-Bahn, the first cogwheel railway in the world, was built in 1871 by Nikolaus Riggenbach of Basel and achieved early fame. Automobiles have not displaced rail traffic: in 1970 the Swiss Federal Railways carried 104 percent more passengers than in 1938, as well as 235 percent more freight.

Shipping. In 1970 the Rhine harbours of Basel handled a record tonnage; 1969 figures were lower. Heating and motor fuels were the major items. Since World War II, Switzerland has also maintained its own small "oceangoing fleet" of 31 ships, with a total tonnage of 200,000 metric tons. The popularity of the steamers on Lake Geneva has never flagged since they were inaugurated in 1823. Regular summer services are provided on several lakes by more than 120 vessels, which include a few paddle-wheelers.

Airlines. Swissair ranks 20th in number of passengers (1969) among the world's commercial airlines. The largest airports are in Zürich–Kloten and Geneva. The total number of motor-powered planes in Switzerland is over 1,000—including helicopters, which play an increasingly important role in mountain rescue operations.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. The Swiss Confederation consists of almost 3,000 communes that, like the 22 cantons, are largely autonomous and that range in size from Bagnes, with an area of 110 square miles (282 square kilometres), to Rivaz, with an area of 0.1 square mile (0.3 square kilometre). From the multiplicity of small communal republics stems a special quality and,

paradoxically, a basis of the inner strength of the entire nation, for each citizen treasures and supports the freedom of the commune, a shared conviction that unites him with the rest of the population in a way that transcends differences of language and of party.

Constitutional bases of the confederation

The Swiss Confederation's main body of laws rests upon the Federal Constitution, adopted in 1848 and revised in 1874; its 123 articles established the rights and duties of the citizen and of the governing bodies. Although in the 1960s there were discussions throughout the population, in numerous professional associations and societies, and in the press concerning a proposed overall revision of the Federal Constitution, no future version was expected to deviate from the basic political concepts derived from the Bundesbrief (document of confederation) signed in 1291 by the three original cantons on Lake Lucerne.

The federal government supervises external and internal security, all transportation affairs, forestry, water conservation, and the like and deals with foreign policy and customs, the postal, telephone, and telegraph systems, the monetary system, and the military organization, as well as with the extensive social insurance programs. It may take steps to adjust the course of the economy and provides for uniform administration of justice in the areas of criminal and civil law.

The organizational units of the federation are the people and the states, or cantons, the Federal Assembly, the Federal Council, and the Federal Supreme Court. Legislative power resides in the bicameral Federal Assembly, comprising the National Council, with 200 deputies elected by a system of proportional representation, and the Council of States, in which each canton is represented by two deputies. The administrative Federal Council is a seven-member collegial board, an organization of equals with a rotating presidency, with each councillor presiding over a federal department. Supreme judicial power is vested in the Federal Supreme Court, the seat of which is in Lausanne (Vaud).

The Swiss Civil Code of 1912 has furnished a model for the administration of justice in many countries, and parts have been adopted verbatim. The difficult task of creating and preserving a uniform judicial system within so diverse a national structure has produced a number of great jurists and international law experts in Switzerland. Each canton elects and maintains its own magistracy for ordinary civil and criminal trials. Capital punishment was abolished by the unified federal penal code of 1937, except under circumstances of martial law, general mobilization, or war.

Sovereign power rests with the people, however, who vote on proposed legislation several times a year and elect representatives every four years. In 1971 a popular vote for the first time gave women the right to elect officials and to vote on issues of national concern.

The principal political parties are the Swiss Social Democratic Party, the Liberals, and the Conservative Christian Socialist Catholics, who in the 1967 election to the National Council won 51, 49, and 45 seats, respectively.

Security. *Police.* Police authority is generally exercised by the cantons, but larger communes also maintain police forces. A small federal police corps enforces special federal laws concerning such crimes as treason and forgery (mainly in collaboration with cantonal police).

The citizen army

Defense. In accordance with confederation neutrality, which dates back to the 16th century and in 1815 became international law, the army serves solely to preserve the independence of the country. Defense is based on a system of universal conscription under which every Swiss male is liable for military duty between the ages of 20 to 50 years and, for officers, to 55 years. The training of recruits is followed by eight annual three-week refresher courses and later by shorter supplementary courses. Swiss women may serve as volunteers in the Women's Auxiliary Force. The Swiss is the only soldier in the world who keeps his equipment, including arms and ammunition, at home and who performs his obligatory gunnery duty each year in civilian clothes, a manifestation of the extraordinary degree of trust between citizen and government.

The social milieu. *Education.* The 1874 constitution stipulates that each canton or half canton is sovereign in education, that elementary education is compulsory and free, and that all public schools must be state directed. The confederation financially assists public elementary schools, vocational training, and the cantonal universities, and, by regulating examinations for the professions, it also influences the curriculum of the secondary schools. The only institutions of higher education maintained by the confederation itself are the Federal Institutes of Technology at Zürich (founded 1855) and Lausanne (federalized 1969). The interior department in Bern administers education, and there is an education department in every canton.

The combination of Alpine scenery and the pedagogic reputation of the country of educational theorists Jean-Jacques Rousseau and J.H. Pestalozzi forms a common attraction, and approximately 10,000 students annually receive their schooling during a sojourn in healthful surroundings, thus combining the useful with the beautiful. In addition, more than 8,000 foreign students per semester attend the ten cantonal and federal institutes of higher education, where they comprise approximately 25 percent of the total number of students.

Social insurance. The law stipulates that assistance to the poor is primarily the concern of the family. Public welfare services were developed in a typically federalistic way: first in the communes, then in the cantons, and later in the confederation. Social insurances that were introduced by amendments to the federal constitution already existed in some communities. The most important was compulsory old-age insurance (introduced in 1948, extended in 1960 to cover chronic illness), which provides annuities and rent allowances to all men over 65, all women over 62, widows, orphans, and invalids; it is financed by contributions from every worker under 65 and his employer and by smaller contributions from the cantons and the confederation. The old age subsidy does not cover the cost of living: communes, cantons, and the confederation provide additional support for the destitute out of general revenues. Compensation for loss of earnings due to military service is similarly financed by contributions.

Federal-local interaction in public welfare

The national income. The national income has been increasing constantly since 1945. As extremely high individual incomes are rare, the repartition of the national income appears tolerably balanced. In spite of a growing inflation, the average income of a Swiss worker showed an increase in buying power of more than 30 percent in the 15 years after 1949.

Leisure. Surrounded as they are by scenic splendour and always within one-half-hour's journey of mountains, pastures, lakes, and rivers, the Swiss have become extremely sports conscious and have encouraged the growth of skiing, tobogganing, and mountaineering as tourist attractions. Other sports include Swiss-style wrestling, gymnastics, regular Sunday-morning shooting, Hornussen (a kind of Alpine baseball), tennis, golf, ice hockey, soccer, sailing, and swimming. The fishing is good in the lakes and rivers, and when the lakes freeze over, they are used for curling and even horse racing.

CULTURAL LIFE

Although Switzerland may not rank foremost among the centres of European culture, the country can yet boast an impressive list of contributors to the arts and sciences. Perhaps because their country has presented too fine a view with too little challenge, some of Switzerland's creative minds have chosen to live elsewhere. This has been especially true of architects. On the other hand, Switzerland's traditional neutrality and its laws of political asylum have made the country a magnet for many creative persons during times of unrest or war in Europe. The mid-19th century was such a period, as were the 1930s and 1940s, when the rise of Fascism caused a number of German and Austrian and Italian writers such as Thomas Mann, Stefan George, and Ignazio Silone to seek harbour in Switzerland.

If a "Swiss culture" can be spoken of in its broader

implications beyond the arts, distinctive French, Italian, and German cultural circles must be recognized, as well as a robust Rhaeto-Romanic culture. In a real sense the four parts of the country represent a definite resistance to levelling or homogenizing influences, political or cultural. In addition, Switzerland's geographical centrality in Europe can be seen reflected in the notion of *Helvetia mediatrix*—Switzerland the mediatress. The Swiss Henri Dunant, founder of the International Red Cross, lives on in the continued sense of many Swiss in a distinct mission of cultural union, a mission revealed not only in the foreign assistance programs toward self-help being carried out in several small and developing nations but also in the unusual number of literary translators within the country.

Architecture. The 12th-century Romanesque style found particularly rich expression in the cathedrals of Geneva, Basel, Lausanne, Sion, and Chur and in the many castles and fortresses; and the Gothic style was expressed in the cathedrals of Zug, Zürich, and Schaffhausen. But in the Renaissance period many architectural masters, especially from the Ticino, gave their talents to Italy: Antonio da Ponte built the Prisons near the Doge's Palace and the Rialto Bridge in Venice; Antonio Contino built Venice's Bridge of Sighs; Domenico Fontana (1543–1607) designed the whole of Lateran Palace, and the facade of St. John Lateran Church, and the Royal Palace in Naples; his nephew Carlo Maderno became architect to Pope Paul V; Francesco Borromini built San Carlino, the gallery of the Spada Palace, and the Filippini Monastery and modified the Falconieri and Barberini palaces; Carlo Fontana built the facade of San Marcello al Corso and the Montecitorio Palace; Baldassare Lenghena, from Maroggia, built the Church of Santa Maria de la Salute, the Rezzonico, Pesaro, and Widmann palaces, the Ospedaletto, and the interior of Scalzi. In a later period, G.B. Gilardi rebuilt the Kremlin in Moscow and his son Domenico the Moscow University and a number of the Soviet capital's palaces. In the 20th century Le Corbusier (Charles-Édouard Jeanneret) was one of the major creative forces behind the International school of architecture that dominated most building trends throughout the West.

Visual arts. The Protestantism of 16th-century Switzerland had a strongly inhibiting effect on Swiss painting and sculpture in general. The peculiar lack of either French or Italian Renaissance influence left Swiss artists in a limbo from which they showed signs of recovery only in modern times. Artists of international renown have included Alberto Giacometti and Paul Klee. The arts are championed by renowned museums not only in cities such as Basel, Zürich, and Genf but also in small towns such as Winterthur and Schaffausen, which pride themselves as cultural bastions far beyond the usual provincial standards.

Literature. In the field of literature, Switzerland has produced Jean-Jacques Rousseau, the historian and critic Jacob Burckhardt, Madame de Staël, whose home at Coppet became a centre of European literary life during the 18th century, Gottfried Keller, and Nobel Prize winners Hermann Hesse and Carl Spitteler. The contemporary dramatists Friedrich Dürrenmatt and Max Frisch have been widely produced outside Switzerland. By the great number of dialects in all the four languages, a living dialect literature is comprehensible. It is an expressive cultural characteristic of nearly each valley.

Music. Although Switzerland is not a leading musical nation, in the 20th century such composers as Arthur Honegger, Othmar Schoeck, and Frank Martin have gained international renown. An annual international music festival is held in Lucerne.

Science. Swiss men of science have included Paracelsus (Theophrastus Bombastus von Hohenheim), who in the 16th century brought chemistry into the field of medicine; the Bernoulli family of Basel, who made significant contributions to mathematics over three generations; the innovative mathematician Leonhard Euler; the naturalist and pioneer Alpine scholar Horace Bénédict de Saussure; and the Nobel Prize-winning chemist Alfred Werner. Zürich's Federal Institute of Technology has produced more Nobel Prize winners than any other scientific school in the world.

Folk arts. Since Switzerland remains largely a nation of small rural communities, folk arts have continued to flourish, although sometimes for the sake of the tourist industry. Survivals include yodelling, alphorn blowing, weaving, wood carving, and ceramics and metalwork.

Communications. Radio and television in Switzerland are largely noncommercial, but, because of the separate language channels, efforts are somewhat dissipated. Journalism is of a high standard, and the *Neue Züricher Zeitung* is regarded as one of the finest of the world's newspapers.

PROBLEMS AND PROSPECTS

Relations with the European community of nations

Switzerland in the early 1970s continued to enjoy an economic boom. It was a member of the European Free Trade Association, the General Agreement on Tariffs and Trade, the Organization for Economic Cooperation and Development, and the United Nations Conference on Trade and Development. In addition, it participated in several other United Nations organizations, such as the Economic Commission for Europe and the UN Industrial Development Organization. With all this and a policy that advocated ultimate participation in a European common market, Switzerland still remained outside the European Economic Community (EEC). Stumbling blocks to entry included the country's traditional neutrality, the federalist and directly democratic political structure, the sensitivity of this small nation's specialized economy, and difficulties in maintaining its agriculture under the community's agricultural policy. Although negotiations were under way for the establishment of special relations with the EEC in 1972, it appeared that Switzerland might be entering a period of political and economic transition to fit it for closer association with its more powerful neighbours, who themselves hoped to emulate Swiss success in forging unity out of diversity. In 1972 a special arrangement with the EEC was concluded and signed in Brussels.

BIBLIOGRAPHY

General and topographical: JOHN and MOIRA HAWKES, *Switzerland and Liechtenstein* (1970), provides a short introduction, to be supplemented by JOHN GURDON, *Switzerland* (1967). See also the comprehensive ALFRED WALDIS, *Switzerland: The Traveller's Illustrated Guide* (1966); other introductions are provided by THEO CHOPARD (ed.), *Switzerland Present and Future: A Small Country Re-Examines Itself*, 4th rev. ed. (1963); and JOHN RUSSELL, *Switzerland*, rev. ed. (1962). Specific topography is covered in R.G. COLLOMB, *Bernese Alps West, from the Col du Pillon to the Lötschen Pass* (1970). On the architecture, see ALFRED ALTHERR (ed.), *New Swiss Architecture* (1965).

Economy: For industrial figures, see the FEDERAL BUREAU OF STATISTICS, *Statistiches Jahrbuch der Schweiz* and the SWISS CREDIT BANK, *The Swiss Economy* (both published annually); the ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT, *The Capital Market, International Capital Movements, Restrictions on Capital Operations in Switzerland* (1971); T.R. FEHRENBACH, *The Gnomes of Zurich* (U.S. title, *The Swiss Banks;* 1966), both on banking aspects; MAX WEIBEL, *A Guide to the Minerals of Switzerland* (1966), dealing with a section of the country's raw materials; and the ELECTRONICS INTELLIGENCE DIGEST, *The Electronics Industry and Market in Switzerland* (1965), are illuminating in their various fields.

Administration and social conditions: The INTERNATIONAL HOSPITAL FEDERATION, *Hospitals in Switzerland: Report on a Study Tour, September 15–29, 1968* (1970), covers some of the health administration; while G.A. CODDING, *The Federal Government of Switzerland* (1965); E.A. GUTKIND, *Urban Development in the Alpine and Scandinavian Countries* (1965); and CHRISTOPHER HUGHES (ed. and trans.), *The Parliament of Switzerland* (1962) and *The Federal Constitution of Switzerland* (1954), deal with other social conditions.

Transportation: An up-to-date work is C.W. SEX and B.J. PRIGMORE, *Swiss Motive Power Survey* (1972). See also C.J. ALLEN, *Switzerland: Its Railways and Cableways, Mountain Roads and Lake Steamers* (1967) and *Switzerland's Amazing Railways*, rev. ed. (1965).

(E.Eg.)

Switzerland, History of

The frontiers of Switzerland have remained almost unchanged since 1815. This stability is relatively modern. The history of this area was for centuries that of a small part of the Holy Roman Empire; changes were frequent, and the emergence of a coherent community in the absence of a common race, language, or religion was long in doubt.

SWITZERLAND TO C. 1300

Before the Christian Era there were sporadic settlements of hunters, fishermen, agriculturalists, and armed warriors in caves along the rivers and occasionally on piles beside the lakes, much as elsewhere in the West. A Tène civilization was followed by growing numbers of Celts. By the time of Julius Caesar, Helvetia (as the region was then called) was known as strategically important to Rome, if only because of the Alpine passes. A Roman invading army asserted its superiority at Bibracte in 58 BC, and Roman colonies or garrisons followed at Nyon (Julia Equestris), Augst (Augusta Raurica), Avenches (Aventicum), and near Windisch (Vindonissa), bringing Roman ideas and civilization to the inhabitants.

North of the Rhine were the Germanic "barbarians." In AD 259 the Alemanni began to penetrate the Roman defensive barrier, and during the 5th and 6th centuries they had become permanent residents east of the Aar, with the less numerous Burgundians to the west of it.

These were followed by Ostrogoths, who were in transit only, and by Franks, whose rulers by 639 had secured recognition as kings, thus ensuring permanent connection with the France of the future. All these invaders were pagans; survivors of Roman converts to Christianity had practically disappeared by the 6th century, when missionaries like St. Columban and St. Gall arrived. Five rather inconvenient dioceses were set up—Chur, Sion, Basel, Constance, Lausanne—and Benedictine monasteries, forerunners of many more, were settled and endowed at St. Gall (modern Sankt Gallen), Zürich, Disentis, Romainmôtier, and elsewhere.

Arrival of missionaries

Such political authority as remained was exercised by the counts, who accepted the Frankish monarchs after Clovis as their overlords. Charlemagne crossed this remote corner of his empire more than once, but such unity as his reign implied was broken at the partition of Verdun (843). This left the land west of the Aar in the hands of Lothair, while that to the east was assigned to Louis the German. Thus, these separate influences, along with local devolution, were fused with the Latin traditions of the Roman Church to shape the future.

In the 10th century both Muslim and Magyar marauders reached Swiss territory, but in each case they were too far from their base to have anything more than nuisance value. By 1033 dynastic accident and deliberate policy had brought the whole of what is now Switzerland into the orbit of the Holy Roman emperor Conrad I, and it remained part of the empire until the end of the Middle Ages. With the re-establishment of imperial authority in the 11th century came conflict with the papacy. Switzerland was divided between the supporters of Emperor Henry IV and those of Pope Gregory VII. The Emperor was supported by the counts of Lenzburg, the bishops of Basel and Lausanne, and the abbot of St. Gall, while the Pope could rely on help from the counts of Rheinfelden, Toggenburg, Kyburg, and Zähringen. It was the counts who gained most. Switzerland was by now thoroughly feudalized. The dukes and counts exercised local jurisdiction practically unchecked. From their castles they dominated and also protected the peasants of the countryside, as did the great abbots and bishops also. Only in walled cities such as Basel, Zürich, Solothurn, Bern, and Schaffhausen was there hope of independence.

Its assertion came, however, not from thence but from the peasant communities of inaccessible mountain valleys, the members of which had learned economic cooperation to avert starvation and were determined to secure freedom from forced labour services or feudal payments in cash or kind.

Towns and dynasts. Many towns were deliberately set up by the ruling feudal families as centres of defense, trade, and administration, Bern most obviously so. Fribourg, Thun, Burgdorf, Laupen, and Morat illustrated the power and extent of the Zähringen dynasty, which died out in 1218. The Zähringen authority reverted in part to the emperor, from whom it had derived. The surviving families—Kyburg, Habsburg, and Savoy—pursued much the same policy of territorial aggrandizement. The Kyburgers, with great possessions in Thurgau and Aargau, came to an end in 1264; and after a struggle with the House of Savoy, Rudolf IV (later Rudolf I of Germany) of Habsburg obtained most of the Kyburger territory, thereby becoming the most powerful ruler with whom the Swiss were in contact. Rudolf was already in control of a sizable part of Alsace and of much of the route from the Rhine to Schwyz and Nidwalden, either as protector of the abbeys of Muri and Murbach or as representative of the emperor.

Foundation of towns

At about the same time, the opening of the St. Gotthard Pass greatly increased the value of the north–south route to Italy and therefore of the communities of Uri, Schwyz, and Unterwalden (Obwalden and Nidwalden). There the peasants had long enjoyed a large measure of freedom from the interference of any feudal overlord and had limited rights to choose their own magistrates, although "high justice" rested with the representative of the feudal lord. In 1231 the men of Uri became directly subject only to the distant Holy Roman emperor, with the right of appointing their own magistrate; and in 1240 Schwyz and Nidwalden obtained a similar privilege from the emperor Frederick II, the validity of which was promptly challenged by the Habsburg overlords. Such uncertainty as there was, however, was largely removed when Rudolf of Habsburg became German king in 1273. The crown meant that he could assert imperial rights in Uri and hereditary family claims in Schwyz and Unterwalden. The heart of Switzerland was at his mercy when he died on July 15, 1291.

The league of 1291. In 1291 Schwyz, Uri, and Nidwalden (Obwalden followed that December) formed the Everlasting League, a perpetual alliance for the settlement of disagreements among themselves by arbitration, for the punishment of crime, for resort to law rather than to violence, and for mutual defense. They would accept only natives as judges, and these must not have paid for their office. All this was confirmed by oath and by a sealed document, which marks the traditional inauguration of the Swiss Confederation (an even earlier union is clearly implied, but its exact nature is unknown).

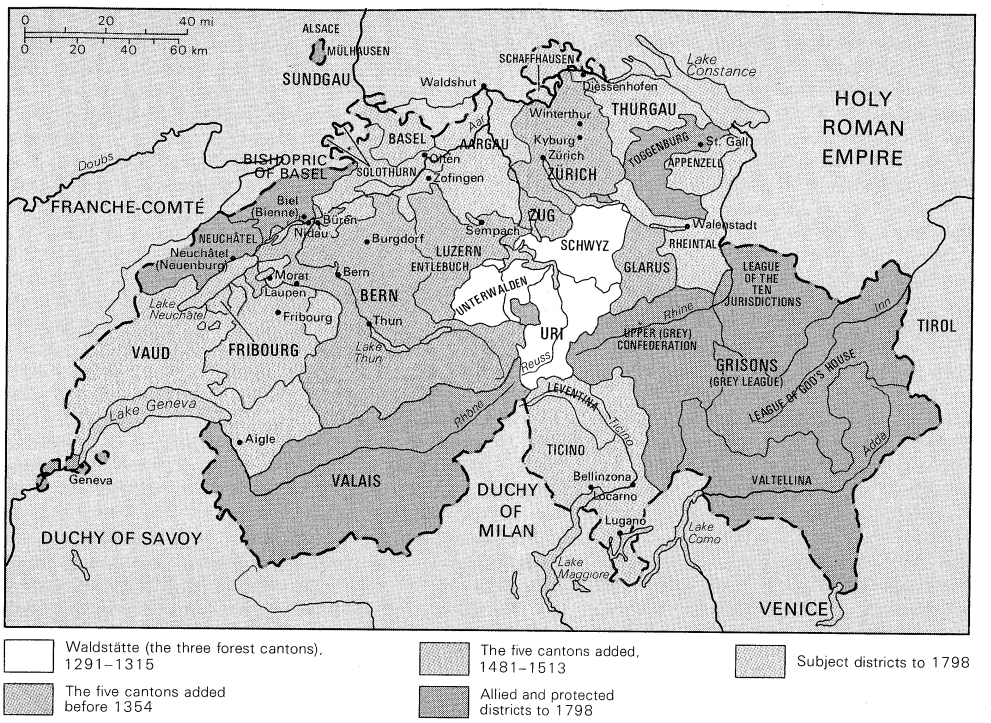
The traditional inauguration of the Swiss Confederation

Rudolf's son, Albert I of Austria, who obtained the German crown in 1298, forced Zürich to abandon its attempt to be comprised within the union. The successive German rulers—Adolf of Nassau (reigned 1292–98), Albert of Austria (reigned 1298–1308), Henry VII of Luxembourg (reigned 1308–13, Holy Roman emperor 1312–13), and Louis IV the Bavarian (reigned 1314–47)—were far too busy elsewhere to intervene in Switzerland, although the Habsburgs relinquished none of their hereditary claims.

THE DEVELOPMENT OF THE SWISS CONFEDERATION FROM C. 1300 TO C. 1530

The struggle for independence and subsequent expansion. In 1314 the men of Schwyz attacked the abbey of Einsiedeln, which enjoyed the protection of Frederick III of Austria, who was challenging the claims of Louis the Bavarian to the empire. The Waldstätte (the three forest cantons, Uri, Schwyz, and Unterwalden) supported the latter claimant, and this led to an armed invasion by Leopold, Frederick's brother, in 1315.

Morgarten and the league of 1315. Ships from Luzern attacked Schwyz by water, and Unterwalden was threatened from Entlebuch. Meanwhile a force of mail-clad horsemen advanced toward Lake Egeri (Ägerisee). At Morgarten on November 15, 1315, their passage was blocked at a point where they were unable to charge and were vulnerable to boulders hurled from the mountain slopes, and the attackers were completely defeated. To



The Swiss Confederation to 1798.
Adapted from W. Shepherd, *Historical Atlas*; Barnes & Noble Books, New York

consolidate their gain, the three allies reaffirmed in a German document (December 9, 1315) their defensive alliance, rejected Habsburg rights, and agreed to make no separate peace. The Habsburgs agreed to a long truce, which lasted, with occasional intermissions, until the end of the century, neither side accepting any permanent renunciation of its claims.

Alliances with Luzern and Zürich

A notable advance was made on November 7, 1332, when Luzern entered into a perpetual alliance with Uri, Schwyz, and Unterwalden, while at the same time maintaining the Habsburg link. It was an alliance and not absorption, but it was advantageous to both sides.

Zürich and Bern. More important was a similar arrangement made with Zürich in 1351. Zürich had secured "immediate" status as a free imperial city in 1218, politically within the Habsburg orbit. It was a flourishing market and manufacturing centre with good defenses, where the nobles were depressed and the artisans were organized into craft guilds. In 1336 an internal revolution enabled the artisans to secure parity with the aristocrats and wealthy merchants; this, in practice, enabled the artisans to exercise supreme power. The dispossessed nobles returned to their castles or to neighbouring Rapperswil; they failed to regain power by a coup d'etat in 1350.

The Habsburgs failed to reduce Zürich to submission, and the city's alliance with Luzern and the three interior communities followed as a natural defensive measure (May 1, 1351). Zürich, besieged or threatened by Austria in 1351, 1352, and 1355, resisted at great expense.

Bern escaped Habsburg domination and Savoyard protection, and after 1218 it asserted its freedom as an imperial city and administrative centre. The predominating influence was military and patrician, not artisan, and accident and design enabled its governors to expand steadily over the neighbouring countryside. During the first half of the 14th century, Bern had sometimes been anti-Habsburg without any long-term commitments and had received help from the Waldstätte in successfully defending Laupen against a coalition including Fribourg and Austria in 1339. At other times—in 1342, 1351, 1352, and later—Bern enjoyed Habsburg friendship. For greater security, and to ensure that Unterwalden did not support hostile peasant movements, Bern agreed to a perpetual union with Uri, Schwyz, and Unterwalden in 1353, although this did not prevent Bernese–Habsburg cooperation when it was advantageous to both sides.

Expansion of the confederation. An unsuccessful attack on Solothurn by Rudolf II, count of Kyburg, in 1382 enabled Bern to secure Thun and Burgdorf. This meant direct collision with Leopold III of Tirol, who was opposed by a Swabian League (founded 1377) and by Zürich, Bern, Solothurn, and Zug. At the end of 1385 Luzern suddenly occupied the Habsburg-controlled areas of Rothenburg, Sempach, and the Entlebuch valley. Leopold replied by leading an army of some 4,000 feudal nobles to Sempach; there, on July 9, 1386, he was defeated and killed by a similar force from central Switzerland.

The Habsburgs still remained the lawful overlords of the small district of Glarus, where the inhabitants were anxious to imitate the successful moves toward independence of their immediate neighbours. They had supported Schwyz in 1323 and in 1351 had willingly allowed their land to be used as a base for confederate operations against the Austrian besiegers of Zürich. In June 1352 Glarus was admitted as a somewhat subordinate partner in the Swiss alliance, followed almost immediately by Zug. Both became subject to Austria in 1355; Zug was occupied by Schwyz in 1364 and remained under its protection after the Peace of Torberg in 1368. The men of Glarus occupied Windegg in July 1386; and in April 1388, unsupported by their neighbours, they met the Austrians at Näfels and won their freedom. Thereafter an independent Glarus was a member of the union. Bern characteristically took advantage of the Austrian embarrassment to secure Büren, Nidau, and Olten. After Näfels, the Swiss Confederation was a fact of history; and the eight states—Uri, Schwyz, Unterwalden, Zug, Luzern, Glarus, Bern, and Zürich—had to learn the lessons of cooperation.

The eight-state confederation

In 1370 all these except Glarus and Bern faced the demand from their ecclesiastics for exemption from taxation and military service. This, when war threatened, was unacceptable, and the states agreed by the Priests' Charter (Pfaffenbrief, October 1370) that clergy and laity alike were subject to the same laws; unauthorized commando expeditions were also prohibited, and, in certain cases, the contracting powers agreed to abide by a majority vote. Military success led to an agreement about laws of war in 1393 by eight signatories, including Bern and Solothurn—looting was to be controlled; women, children, and clerical property were protected; and deserters were repudiated by all.

Switzerland in the 15th and early 16th centuries. With the 15th century came an interval of peace, consolidation, and expansion. Zürich, stronger and wealthier, came to terms with its Habsburg neighbour while adding to its own territory, and Bern expanded to the east and west. The central states recognized that security meant expansion southward as the St. Gotthard route became increasingly used. Glarus opened talks with its southern neighbour, the Upper, or Grey, Confederation. There the bishop of Chur, the abbot of Disentis, and the local barons, Misox and Räzüns, accepted the existence of peasant leagues; thus the Grisons (valley area that included the city of Chur) became almost a Switzerland in miniature on the southern and eastern wings of the confederation.

Around the monastery of St. Gall a flourishing trading city had grown; and as the economic importance of the monastery declined in spite of its wide possessions, tension grew between city and abbot. The citizens found allies in the communities of Appenzell, over which the monastery claimed jurisdiction. The abbot sought help from the Austrians in the Rheintal, to which the Appenzellers replied by an alliance with Schwyz. In 1403 a monastic army was defeated, whereupon the abbot sought and obtained the support of Duke Frederick IV of Austria. Schwyz was bound by treaty to keep clear of such a conflict, but the citizen army of St. Gall beat off one attack and the Appenzellers another in June 1405. After this Appenzell was accepted as an ally by seven confederate states, excluding only Bern.

Alliance with Appenzell

First Italian conquests. In the south the men of Uri were becoming anxious about the security of their communications with northern Italy. The inhabitants of the Leventina (the narrow valley of the upper Ticino River) appealed for help and protection and so, in 1403, became subject to Uri and Obwalden (western Unterwalden); Bellinzona was purchased by Uri but was retaken by Milan in 1422. Uri and Unterwalden appealed for help from their partners. Zürich and Bern took no action, but Luzern and Zug cooperated and Schwyz sent a force toward Domodossola. All these forces joined at Arbedo, where on June 30, 1422, they were surprised and defeated.

The southern expeditions also made contact with Valais, part of which was controlled from Savoy and part ruled by the bishop of Sion. The latter in 1403 made an alliance with Uri, Unterwalden, and Luzern, which only increased internal dissensions.

Conquest of the Aargau and the Old Zürich War. A dispute in 1415 between Duke Frederick IV of Austria and the emperor Sigismund provided the confederates with a pretext to attack the Habsburg possessions south of the Rhine. A Bernese force occupied the Aargau, Zofingen falling in April 1415; a Luzern contingent occupied Sursee and St. Urban; and Zürich made similar gains. Sigismund was forced to pledge part of Aargau to Zürich and Bern—the former taking the portion of the Freiamt east of the Reuss, Bern securing the western half, with pickings left for Luzern. The remainder of the land became a "common lordship" (*gemeine Herrschaft*) of the confederation. This was to cause infinite trouble later, but it was also to prove part of the cement that held the confederation together.

In 1436 Frederick VII, count of Toggenburg and owner of Uznach, Sargans, Gaster, Rheineck, Altstätten, and Feldkirch, died childless. Schwyz and Zürich quarrelled over the partition of the inheritance, and war followed. Unsuccessful, in 1442 Zürich made an alliance with the Habsburg emperor Frederick III and in 1443 was at war with the whole confederation. The city was besieged; its *Bürgermeister* fell in an unsuccessful sortie at St. Jakob an der Sihl (1443) after negotiations for an armistice had broken down and the garrison of Greifensee had been liquidated. In 1444 Frederick III secured support from Charles VII of France, who sent a large force of irregulars to the neighbourhood of Basel. There, at St. Jakob an der Birs, this great army was halted by a Swiss detachment one-tenth its size, which fought until only a few wounded remained (August 26, 1444). The French mercenaries were unprepared for further such encounters, and they left for Alsace (thereafter France sought Swiss friendship rather than face hostility, and the already high reputation of Swiss fighters was enhanced). France made peace at Ensisheim (October 1444) with the confederation (minus Zürich, but with Basel and Solothurn added). This implied recognition, and it was followed in June 1446 by the Peace of Constance between Zürich and the confederates on the basis of a return to prewar frontiers, with Schwyz slightly advantaged and Austria excluded south of the Rhine.

Recognition by France

St. Gall and Thurgau. In August 1451 the abbot of St. Gall entered into a protective agreement with Zürich, Luzern, Schwyz, and Glarus; the city of St. Gall did likewise but added Zug and Bern. Appenzell and Schaffhausen obtained similar support in 1454. Fribourg secured complete independence from Austria in 1452 and assurance of support from Bern. The Swiss seized Thurgau from Sigismund of Tirol, cousin of the Habsburg emperor, in 1460–61. Like Aargau, it became a common lordship. This successful aggression was accompanied by economic advance and a growing degree of comfort and good living, to which the profits of mercenary military service contributed.

Bern gained the leadership that Zürich had sacrificed in its vain attempt to secure the Toggenburg inheritance, and Bernese hegemony was enhanced by the successful absorption of its feudal nobility into the government.

In 1466, with Bernese influence preponderant in the Aargau, Mülhausen (Alsace) secured its alliance. Support for Mülhausen involved hostility toward the Habsburgs and a general Swiss attack on the Sundgau and Waldshut. By the peace terms of 1468, Mülhausen remained independent and Waldshut was pledged as a guarantee of the payment of a war indemnity of 10,000 gulden by Sigismund of Austria. This was repaid when Charles the Bold of Burgundy lent Sigismund 50,000 gulden and took over Austrian rights in Alsace and the Black Forest as security.

The Burgundian War. Out of these events developed the Burgundian War. Bernese policy was largely shaped by the soldier Adrian von Bubenberg and by his rival, the capitalist politician Niklaus von Diesbach, who was also influential at the court of Louis XI of France. Charles of Burgundy used his rights under the Truce of Saint-Omer to overrun Alsace and threaten Mülhausen in 1470; as a result Louis XI secured a guarantee of Swiss neutrality in the event of a Burgundian attack on France. In 1474 the Swiss were persuaded to extend this neutrality to Sigismund of Austria by the Treaty of Constance, which was followed by Sigismund's denunciation of his concessions to Burgundy.

The execution at Breisach of Charles's representative made a Burgundian threat to the whole of Sigismund's possessions a certainty; the Swiss obtained a promise of financial support from Louis XI and in October 1474 opened hostilities against the Burgundian duke and his ally, Savoy. In March 1476 a Burgundian army was defeated by a Swiss force at Grandson, and rich booty was taken; the Burgundians answered by advancing against Morat. A large Swiss force from every state of the confederation accepted this challenge, relieved Morat, and defeated the enemy in the open field (June 22, 1476). This was the high-water mark of Swiss power; any army in Europe could now be met on at least equal terms. The final demonstration came at Nancy, where Charles the Bold was defeated and slain in January 1477. Had the confederates been able to agree on a common foreign policy, the whole of Franche-Comté could easily have been secured and much of Alsace as well.

High point of Swiss power

Divisions and reconciliation. Instead there was disunion: Zürich refused to support a Bernese "western" expansionist policy; the farmers of Bern, Zürich, and Luzern opposed their own governments; the inner Swiss states resisted any centralization or expansion of the federal union; and everywhere young men willingly sold themselves into service as mercenary soldiers. Civil war was averted by the hermit-saint Niklaus von Flüe (Bruder Klaus); and by the Agreement of Stans (December

1481), Fribourg and Solothurn were admitted to the confederation, now enlarged to ten states. All would help one another to repress any internal opposition.

Diesbach of Bern died in 1475, and his place as the leading Swiss notable was taken unworthily by Hans Waldmann (1436–89), a Zürich millionaire adventurer and dictator, with a policy of civic protection and domination over the agriculturalists, including the central states. Waldmann was deposed and executed in 1489, but the Swiss advance to great power status had been checked at a critical juncture, and the harmony of Bern and Zürich, on which so much depended, had been broken.

In 1499 the emperor Maximilian I, in pursuit of impractical ideals of an efficient and solvent empire, attempted to force imperial jurisdiction and taxation on his Swiss subjects. The result was resistance, leading to the Swabian War; the outcome was a double defeat for the Emperor—at Calven (May 1499) and at Dornach (July 1499). By the Peace of Basel that followed on September 22, the practical independence of the Swiss was assured and Basel (1501) and Schaffhausen (1501) were admitted to the confederation (further evidence that all south of the Rhine repudiated Habsburg overlordship).

The Italian campaigns. With the opening of the 16th century the "Old Confederation," completed by the admission of Appenzell in 1513, faced external conditions that transformed its international position. France had been united under an almost absolute hereditary monarch; the emperor after 1519 was ruler of Spain and much of Italy as well as of the German-Austrian Habsburg territories; the Italian states, which had given intellectual leadership to the West, were no longer able to keep out the "foreigner"; and the papacy, conscious of the danger from the infidel Muslims, could not even eliminate heresy within the Christian Church.

Changes in the Swiss international position

French invasions of Italy, and particularly claims of the French kings Louis XII and Francis I to the Duchy of Milan, affected the Swiss directly. Some 5,000 paid Swiss soldiers helped Louis enter Milan in 1499. Swiss armed men were found both with and against the French at Novara in 1500; their refusal to fight one another led to French control of Milan. Three years later, at Arona, the French reluctantly recognized the fact of Swiss occupation of Bellinzona; but already there was growing opposition to mercenary service and its implications.

The next important paymaster of Swiss troops was the warrior-pope Julius II. In 1510 he brought the confederation, together with Valais, into an agreement to support the Pope (and Maximilian) against Venice. The murder of two Swiss envoys at Locarno was the pretext for sending a Swiss force against the French in northern Italy in 1511, in a raid that reached Milan. In 1512, in conjunction with Venice, a large Swiss army occupied Verona, Pavia, and Milan; the French were again forced to leave, and the Swiss could have been masters of the duchy had political conditions made this feasible. Accepted as allies by Savoy and as "Protector of the freedom of the church" by the Pope, the Swiss had to face yet another French force, again at Novara (June 1513), of which they regained possession.

In 1515 Francis I purchased much Swiss goodwill, and on his invasion of Piedmont in 1515 many of the Swiss soldiers occupying northern Italy agreed to return home. The others, perhaps 20,000 strong, met the invaders at Marignano (September 13–14, 1515) and were driven back decisively after heavy fighting with serious losses. French artillery, and the arrival of Venetian reinforcements, may have been the effective instruments in this (but it meant that the day of Swiss invincible infantry was over, although this was not yet apparent). France also gained a treaty of Perpetual Peace in November 1516, whereby in return for money and economic advantages the Swiss renounced any claims on Milan but retained Bellinzona, Lugano, Locarno, and the Valtellina. Included in the expeditions to Pavia and to Marignano were men from Glarus accompanied by their chaplain, the Catholic priest Huldrych Zwingli (1484–1531).

The Reformation. The condition of the church in Switzerland was no better than elsewhere in western Europe. A poor population supported many privileged clergy, who were often inadequately educated and sometimes negligent in their duties. The many monasteries were wealthy but not popular, and the bishops saw no need for change in the parishes. Religion consisted largely of routine duties, attendance at mass, confession, rituals, and the performance of many external acts of piety. Clerical concubinage was widespread and tolerated by the authorities.

Condition of the church in Switzerland

The popes were paymasters of Swiss soldiers, who formed their bodyguard and served in the papal armies; but enthusiasm waned when payment was lacking or when the service demanded was too strenuous.

Huldrych Zwingli. The impact of the Reformation on Switzerland was felt more in the cities than in the countryside, and it was associated with the names of Erasmus, Zwingli, and Luther. Erasmus had been critical of worldly popes, ignorant friars, complacent monks, and petty abuses, while being careful not to cross the bounds of orthodoxy; he felt that, by the study of the Fathers and of the Bible, the Christian Church could return to the simplicity and excellence of its earliest days. Luther publicly attacked the sale of indulgences in 1517.

Zwingli likewise had turned to Bible study, including Hebrew, and had anticipated Luther in rejection of traditional doctrine where it seemed to conflict with Holy Writ. At the same time, Zwingli was actively concerned, as a patriotic confederate, with the ill effects of mercenary service and the consequential pensions paid to leading politicians by foreign powers. In 1518 he was elected by the canons of the Zürich Great Minster as their stipendiary priest (*Leutpriester*). From 1522 he developed his independent thought into a challenge to the church authorities, first on the question of compulsory abstention from meat eating during Lent, then about clerical marriage—his own marriage was not publicly admitted until 1524—and then to an attack upon the mass as a sacrifice. He defended his views in public debates in Zürich in January and October 1523, which led to making evangelical preaching compulsory and to the rejection of pilgrimages, relics, images, pictures, church organs, and the sale of indulgences.

Zwingli's reforms

The appeal to the Bible as the sole guide to doctrine and observance involved a clash with the Anabaptists, who in 1524 insisted that infant baptism was unbiblical and who called for an independent, locally based church instead of the all-embracing state church of the Zwinglians. They regarded tithes, interest, oaths, military service, and capital punishment as unacceptable and, in turn, were themselves unreasonably feared as antisocial revolutionaries. After a third public debate in Zürich in 1525, they were outlawed.

Other states followed the example of Zürich, the decisive action in every case being the substitution of a commemorative communion service for the Roman mass. Attendance at the Catholic mass was made illegal in Zürich in April 1525, an example followed later by the city of St. Gall and by Schaffhausen, Glarus, Appenzell, Basel, and Bern. Reform was accompanied by the dissolution, often by agreement, of the monasteries, their revenues being used for education and poor relief in some cases, for normal public expenditure in others. The acceptance of the Reformation in Basel was the result of a popular demonstration, itself a protest against the conservatism of a patrician oligarchy, and of the effective preaching, discussion, and writing of John Oecolampadius.

Zwingli's greatest triumph was to secure the adhesion of Bern to his teaching at a well-managed public discussion in January 1528. A Catholic attempt to demonstrate the errors of their opponents at the Baden disputation (1526), at which Zwingli was not present, was followed by efforts to eliminate "heresy" from the confederation. Bern refused to yield to this pressure; there was much sympathy for Zwingli's attack on pensions and mercenary service, and the moral and intellectual condition of the Bernese clergy was deplorable. In 1527 Bern's government accepted married clergy, freedom of evangelical preaching, and control of monastic property, and it finally agreed that the case for and against the older and

newer beliefs should be settled by public discussion. At this debate the Zwinglian triumph was both apparent and far-reaching; images and the mass were rejected, and Bern remained exclusively Protestant until the end of the 18th century. There was a minor revolt of the conservative peasants of the Bernese Oberland, which served only to stiffen the determination of the government to maintain religious uniformity.

Problem of jurisdiction over the common lordships

The spread of the Reformation in the common lordships of Aargau, Thurgau, and neighbouring areas provided a problem of jurisdiction that could only be settled by force. The inner Catholic states—Luzern, Uri, Schwyz, Unterwalden, and Zug—obtained an alliance with Ferdinand II of Austria, brother of the emperor Charles V, which was countered by a Protestant Civic Union that included Constance and Strassburg. When the Catholics made it clear that there would be no freedom for preaching the gospel in lands in which they shared authority, Zürich prepared for war. Zwingli tried unsuccessfully to form a Protestant union, of which Philip of Hesse would be the leading member. Although Philip brought Luther and Zwingli face to face at Marburg (September–October 1529), no full agreement was possible about the presence of Christ in the bread and wine of the Holy Communion.

An armed conflict between Zürich and the five inner Catholic states was soon terminated. The Peace of Kappel that followed in June 1529 suggested some toleration in the common lordships, but this was acceptable to neither side in practice. The alliance with Austria was torn up, but this, too, had little practical effect. The prince abbot of St. Gall was refused recognition, and his subjects were encouraged to revolt as well as to apostasy. A threat by a robber baron from the castle of Musso was represented as part of a renewed Catholic plot against the evangelicals; and the failure of Unterwalden to keep promises to Bern enabled Zürich to prevent indispensable supplies of grain and salt from reaching the inner Catholics and then to threaten coercion by force of arms. The second Kappel War brought the defeat and death of Zwingli (October 11, 1531) at Kappel and another defeat, also involving Bern, at Gubel (October 22, 1531).

By the second Peace of Kappel (November 1531), Zürich, Bern, Basel, Schaffhausen, and some other states remained Protestant but unable to advance their cause further; the abbot of St. Gall was restored and imposed his faith on his dominions; and the communities of the common lordships lost their right of future local religious option. Heinrich Bullinger (1504–75) replaced Zwingli as the chief minister of Zürich, but he had none of Zwingli's political influence. Bern turned to defend its southern borders from Savoy and Catholic Valais, which led to its occupation of Vaud (now a canton in southwestern Switzerland) and direct contact with Geneva.

Direct contact between Bern and Geneva

Geneva (Genève), which did not become part of the confederation until 1815, could escape domination by Savoy only by reliance upon Bern, which in turn needed for its security to control the territory, including Lausanne, as far as Lake Geneva. Guillaume Farel (1489–1565) after 1527 evangelized much of Vaud as far as Aigle but met with obstruction in Geneva, which, along with Lausanne, had secured an alliance with Bern and Fribourg. To protect their interests and the achievement of the Reformers, Bernese troops appeared outside Geneva in 1530, forcing the Duke of Savoy to pledge Vaud for security and to allow the gospel to be preached. In 1535 the mass was abolished in Geneva; and in 1536, to prevent a French occupation, a Bernese force easily overran Vaud. The territory on both sides of Lake Geneva was incorporated into the Bernese possessions, and Zwinglian worship was accepted everywhere.

John Calvin. John Calvin (1509–64; *q.v.*), already a distinguished French Humanist, had abandoned the Catholic faith and in 1536 issued from Basel his *Institutes of the Christian Religion*. As Calvin passed through Geneva in 1536, Farel, who was finding his pastoral duties there beyond his powers, called on him to stay. Calvin became, even more than Zwingli at Zürich, the dominant influence in Genevan affairs. His political opponents were driven underground; and the freethinker Jacques Gruet and the Unitarian Michael Servetus were put to death. Calvin's system and doctrine were similar, but not identical, to those of Zwingli; but in 1566 they were reconciled by the Second Helvetic Confession. A Presbyterian church with complete local autonomy for the congregations, with ministers, doctors, elders, and deacons, became a model for others to follow. Geneva's application for admission to the confederation was rejected; the danger from Savoy was continued, and as late as 1602 an attack (the "Escalade") was beaten off with difficulty. Thus Geneva was in many ways a besieged fortress, and it submitted to the requirements of a mixed lay and clerical consistory the more willingly for this reason. Its alliance with Bern and the determination of the French that Geneva should not be Savoyard enabled the city to become an intellectual centre and a haven for refugees until the French Revolution.

THE CONFEDERATION SINCE C. 1530

The Counter-Reformation and the Valtellina. While Calvin was preaching in Geneva, the Catholic Church belatedly set its house in order at the Council of Trent (1545–63). The reforms were welcomed by the Catholic cantons of Luzern, Uri, Schwyz, Unterwalden, Zug, Fribourg, and Solothurn, while Glarus and Appenzell remained with divided allegiance. The inner cantons insisted on the expulsion of the Protestant families from Locarno (1555); many of these found refuge in Zürich, where they contributed valuably to its industrial development, particularly in silk. The Jesuits established a school in Luzern in 1577, and a permanent resident papal nuncio was received there from 1596.

Reinforcement of the Catholic cause

Preachers from Geneva encouraged the French Huguenots to resist, and Catholic Swiss mercenaries fought against them. Numerous French Protestants found refuge in Switzerland after the massacre of St. Bartholomew (August 23–24, 1572). Ludwig Pfyffer of Luzern not only supported the Catholic "Holy League" against Henry of Navarre (later Henry IV) but also united the five inner states with Fribourg in a "Golden League" with Spain (1586), one result of which was the permanent division of Appenzell into two separate half cantons. The accession of Henry IV as king of France (1589) prevented civil war in Switzerland, where 12 cantons (Zürich not included) renewed the league with France in 1602, again allowing French recruitment of soldiers.

A Switzerland of about 1,250,000 inhabitants, divided by religion, laws, language, and economic interests, could no longer play an independent role in 17th-century Europe. Geneva, however, was saved from either Savoyard or French domination; and, even more significantly, the Grisons maintained their position as a valued ally of the confederation. There were three leagues—the League of God's House and the League of the Ten Jurisdictions (both Reformed) and the Grey League (Catholic)—which were threatened by Milan and Venice from the south and by the Habsburgs of Vorarlberg and the Tirol from the east. Internally, much power lay with the great local families, such as the Planta, traditionally pro-Spanish, and the Salis, supported by France and Venice.

An attempt by Venice in 1603 to gain economic advantage and control of the passes leading to the Austrian territories led to a Spanish–Milanese counterattack. The result was internal discord in the Grisons, which continued for a quarter of a century. The leader of the Protestant party in the Grisons, supported by Venice and France, was Georg Jenatsch (1596–1639), who was confronted in 1620 with a massacre of some 500 leading Protestants in the Valtellina. Bern and Zürich offered military aid to Jenatsch, but the inner cantons prevented it from arriving, and the Valtellina was left open to Spanish domination.

Jenatsch, after some preliminary successes, was forced into exile. He returned in 1624 with a French army sent by Cardinal de Richelieu to regain control of the invaded territory but not to restore the Reformed faith there. Jenatsch consequently changed sides and was murdered in January 1639 by conspirators.

Swiss neutrality

The Thirty Years' War. Switzerland's neutrality during the Thirty Years' War was the outcome of its religious and economic divisions rather than of conscious policy; although the value of a Swiss infantry army was diminished by 1618, it was still not negligible. At first the Habsburgs seemed likely to win, which might have had serious implications for eastern Switzerland; but by good management official neutrality was observed, although a number of individuals volunteered for military service with either side, and war needs brought profit to Swiss traders. The threat of invasion made agreement between the 13 Swiss cantons highly desirable, and all the more so because of the need to provide for the administration of the common lordships. The independence of each of the 13, with their allies—the cities of St. Gall, Mülhausen, and Biel (Bienne) and the abbot of St. Gall—was, however, fully recognized.

In 1631–32 Gustavus II Adolphus of Sweden narrowly failed to secure an alliance with the Swiss and insisted upon a stricter observance of neutrality; in September 1632 a Bernese force on its way to help allied Mülhausen was attacked in Solothurn, but further hostilities were avoided by mediation. In September 1633, perhaps because of deliberate negligence by Zürich, Swedish forces entered Stein am Rhein intending to attack the imperialists at Constance but retreated; and again, after much trouble, the situation was adjusted without bloodshed. A parallel episode involving Basel led in 1647 to an agreement for common Swiss defensive measures—the Defensionale of Wyl, which marked a further stage in common Swiss action.

The long negotiations that resulted in the Peace of Westphalia (1648), ending the Thirty Years' War, owed a good deal to the efforts of the representative of Basel—*Bürgermeister* Johann Rudolf Wettstein. In October 1647 he succeeded in obtaining a declaration from the Emperor that the city of Basel and the 13 cantons, with their allies and common lordships, were not subject to imperial jurisdiction. This was written into the final treaty and became part of the public law of Europe.

The Villmergen wars. Difference of religion after 1648 helped the relative stability that characterized 18th-century Switzerland. The Protestant cantons were wealthier and more thickly populated than their Catholic neighbours; but in the Diet they were in a minority, and no agreed alteration in the constitution of the confederation was possible. Internally, power rested in the hands of a relatively small number of patrician families; in Bern, Fribourg, Solothurn, Luzern, Basel, and even Zürich the aristocrats dominated the ruling councils and retained for their members power and the emoluments of office.

The gulf between nobles and peasants

The common people, in general, meekly accepted the religion and the legislation of their superiors. There was, however, a gulf between nobles and peasants; and after 1648 an economic recession helped to intensify the objections of the farmers to paying taxes and dues for the defense and support of the walled towns. Each canton had its own coinage, and in some of them tension was increased by attempts to replace a depreciated currency.

Open trouble started in 1653 with a demand from the Entlebuch farmers for better treatment by Luzern, closer consultation, more legal autonomy, a more satisfactory currency, and relief from a hated salt tax. The government obtained promises of help from neighbouring cantons and made a few concessions, and the trouble died down. But the example, once set, was more widely followed; and a general claim from the peasantry for greater attention led to a gathering at Huttwil (May 1653) and an uprising against Bern, which was easily suppressed, its leaders being punished.

In 1656, after the forcible expulsion of a harmless Protestant community from Arth, fighting took place with the five Catholic cantons against Zürich and Bern. In a battle at Villmergen (January 24, 1656) Bernese forces were defeated; meanwhile, forces from Zürich failed to capture Rapperswil. At the Third Land Peace of Baden, which followed, the right of each canton to enforce religious unity was reasserted. After incidents at Wigoltingen in 1664, Wartau in 1694, and St. Gall in 1697, peace was openly broken in 1712 because the aggressive abbot of St. Gall, Leodegar Bürgisser, had promoted the construction of a highway from Wattwil to Uznach that could be to the disadvantage of the Protestants in the Toggenburg. Bern and Zürich used this as a pretext to occupy most of Aargau, Thurgau, Rheintal, and Toggenburg. Attempts at a compromise failed; there was a serious fight at Villmergen (July 25, 1712), won by Bernese discipline, after which the Peace of Aargau (August 11, 1712) agreed that Protestants and Catholics should enjoy equal rights in the common lordships, repealed the clauses of the second Peace of Kappel (1531), and left the balance of power in the confederation in Protestant hands.

Louis XIV of France did not intervene directly. There had long been a standing agreement allowing Swiss soldiers to serve for pay with the French. This Franco-Swiss alliance had been renewed in September 1663, complete with pensions, arrangements for payments of recruits, and other advantages. But the French occupation of Franche-Comté (1668) and the revocation (1685) of the Edict of Nantes, which had granted religious liberty to the Huguenots, were resented. The Defensionale of Wyl was refurbished in 1668, and in 1674 the Swiss Diet publicly proclaimed its neutrality in continental disputes. French occupation of Strassburg was accepted as inevitable, and Basel itself was constantly watched by a French garrison in Hüningen. Following diplomatic moves by Bern and Zürich at the French court in 1687, French pressure on Geneva was removed and Swiss soldiers were permitted to serve in British and Dutch regiments if needed (a secret agreement by the Catholics in 1715 to serve France only—the "Trücklibund"—had no binding validity).

Relations with France

The death of the Duchess of Nemours, Princess of Neuchâtel (Neuenburg), in 1707 enabled Bern to arrange for the inheritance to pass to the Protestant King Frederick I of Prussia, but not without objections then and repercussions later.

The 18th century. From 1712 to 1798 Swiss neutrality was fully respected—the 18th century wars passed Switzerland by. The country remained predominantly pastoral and agricultural, poor, and little visited by foreigners for pleasure; it was around the cities of Basel, Zürich, and St. Gall that home industries flourished. Adequate capital and relatively cheap labour were available; weaving in linen, cotton, and silk occupied many hands; and watchmaking and clockmaking began to spread.

There were few exceptions to the general practice of obsequious deference to higher rank and obedience to authority. An attempt to promote a mutiny in the Bernese army in 1723 to secure the freedom of the inhabitants of the Vaud from Bernese oppression ended in failure. The small number of ruling families that controlled everything in Bern took care to stamp out any threat to this monopoly.

In Geneva the aristocratic ruling clique was less successful; fighting in the streets in 1738 ended with a compromise arranged by Bern and Zürich allowing the "citizens," themselves a minority among the inhabitants, a larger voice in affairs. But in 1768 and 1782 attempts to widen the franchise still further were forcibly smothered.

Cantonal separatism and religious exclusiveness were in some measure countered by the formation in 1761–62 of the "Helvetian Society," for all Swiss without distinction of birthplace or confession—and thus a better reflection of the ideas of the Enlightenment than its Catholic parallel, the Helvetische Konkordia Gesellschaft. This did not prevent trouble in Luzern, the most Catholic of Swiss cities, where a struggle between church and state grew out of an attempt to regulate or forbid public dances.

The French Revolution and Napoleon. Close to the surface there was a good deal of vague popular discontent: Fribourg, for example, in 1781 had to deal quite heavily with the peasants of Greyerz (Gruyères) before they submitted to dictation from the centre. The situation was altered by the American Declaration of Independence and by the French Revolution. Swiss residents in Paris formed themselves into a club of patriots in 1790

Popular discontent

and were in contact with radicals in Basel and Zürich and with well-wishers in Bern and Brugg. The chief exponents of revolutionary ideas, however, were Peter Ochs, a vain and ambitious artisan leader from Basel, and an idealist from Vaud, Frédéric-César de La Harpe, who acted as tutor to the future tsar Alexander at St. Petersburg. La Harpe returned, determined to secure the independence of Vaud from Bern. There was a rising of inhabitants of Lower Valais against their masters in Upper Valais; the last of the feudal servitudes were removed from the Basel peasantry in 1790; also in 1790 a Swiss battalion in the service of the French king mutinied in Nancy. Even Bern was prepared to offer some carefully guarded concessions. The inhabitants of Vaud, with a new-found watchword "Liberté et Patrie," wanted more.

The 13 cantons had renewed their military alliance with France in 1777 at Solothurn, but circumstances had altered greatly by the time the King's Swiss Royal Guard was annihilated in Paris by a mob in August 1792. Other Swiss regiments were unreliable and ill-equipped, and a Swiss foreign policy hardly existed. To prevent a Swiss revolution in which equality would end the traditional aristocratic regime, Schultheiss Niklaus Friedrich von Steiger of Bern took the lead; he became head of a party prepared for a possible war with France.

In the autumn of 1792 French military forces entered Savoy and threatened Geneva. Bern and Zürich sent 1,600 men to support the city, but they withdrew on receiving a French assurance that an attack would not take place. A revolution inside Geneva, protected by French arms, followed without incident.

The formation of the First Coalition Against France in 1793 (Great Britain, the Netherlands, Spain, Naples, Austria, and Prussia) was answered by French resistance. The French government demanded the expulsion from Switzerland of dangerous and hostile émigrés, and other opportunities for French intervention in Swiss affairs came easily. The French supported liberal movements in Stäfa and St. Gall; French and Austrian hostile forces crossed Swiss territory near Basel; French agents spread "democratic" propaganda; French generals coveted control of the Swiss passes; the French demanded the expulsion of the British ambassador and called for the annexation by France of part of Valais and the creation of an "independent" republic of Vaud. In 1797 French forces occupied Geneva and Biel; several cantons hurriedly liberalized their constitutions to give wider rights to hitherto disfranchised subjects.

French intervention in Swiss affairs

In January 1798 the French occupied Lausanne and proceeded to liberate Vaud from Bernese oppression; Bern itself was entered on March 5 after little more than token resistance. The remaining cantons, except Schwyz and Nidwalden, hurriedly surrendered, and the French proclaimed on March 28 that they had assumed governmental authority over the "Helvetian Republic."

The Helvetian Republic. Geneva, Neuchâtel, Biel, the territory of the bishop of Basel, Mülhausen, and the Valtellina were excluded from this new Swiss republic of 1798. The republic was to be "one and indivisible," with an elected legislative assembly and a directory. There were 23 cantons, a great council, a senate, and a cabinet, or directory of five, all indirectly elected. There was to be freedom of worship and conscience. Class distinctions were abolished, and all citizens were equal before the law. Justice was centralized; the directors chose the prefects, who, in turn, appointed the subordinate administrative officers, with the commune (*Gemeinde*) as the basic local unit. Large sums of money were paid to France by way of occupation costs, and attempts in central Switzerland to resist the new regime were crushed.

The renewal of war in Europe in 1799 was accompanied by local revolts, a breakaway movement in the Grisons, and an Austrian invasion there. Behind the Austrians came a Russian army, which occupied Zürich in June 1799 and was defeated at the second Battle of Zürich (September 25–26, 1799). For all practical purposes, at the end of 1799 Switzerland was a French-occupied country.

Within the new Swiss state the Republicans found the new democracy meaningless and called for a qualified return to earlier conditions, a suggestion emphatically rejected by the Patriots, defenders of the status quo. A little later the Unionists, who emphasized the coherence and indivisibility of the republic, were opposed by the Federalists, who insisted upon local self-determination and regional differentiation. The Directory was displaced by a committee of the Helvetic Council in January 1800 and this in turn by a controlling assembly of 50 supporting the Republicans. All parties appealed to Napoleon.

The Treaty of Lunéville (February 1801) enabled Napoleon to impose the Constitution of Malmaison in May 1801. This intensified the rivalries of Federalists and Unionists, and the latter produced in 1802 a second highly centralized Helvetic constitution, which was rejected by popular vote. Valais was accepted as an independent republic; there were serious opposition riots demanding a return to the old arrangements. French soldiers reappeared, and in February 1803 Napoleon produced the final text of his Mediation Constitution. A loose unitary state was formed, with 19 cantons, each with its own constitution, a legislature, a president (*Landammann*), and a chancery. The country now formed a free-trade area, with the peasants freed from feudal obligations, torture abolished, education encouraged, many monasteries dissolved, church and state separated, and marriage made a civil contract. Citizenship was widened, there was a federal army, foreign policy, and code of laws. France was permitted to recruit 16,000 Swiss soldiers, many of whom perished in the ill-fated Moscow campaign of 1812. The Napoleonic "Continental system" pressed hard upon a country that lived by exports and international services; there was a certain amount of smuggling and much unemployment and poverty. There was also, as elsewhere on the continent, a marked upsurge of nationalist feeling; cultural and scientific societies embraced the whole country, for which Johann Rudolf Wyss composed a national anthem, "Rufst Du, mein Vaterland." The Linth Canal (Linthkanal) was a federal enterprise, the first of its kind; Johannes von Müller wrote the first comprehensive history, *Geschichten schweizerischer Eidgenossenschaft* ("The History of the Swiss Confederation").

The loose unitary state of 1803

The restoration. There was no Swiss nationalist rising against Napoleonic domination after 1813 as there was in Prussia or Spain. Instead a renewed insistence upon Swiss neutrality, which might protect Napoleon's flank from an invasion, was accepted as public policy but ignored by the allies in their determination to occupy Paris and end the war. In December 1813 the Mediation Constitution was denounced and an attempt made to return to the pre-Revolutionary 13 sovereign cantons. Aargau and Vaud at once protested; the tsar Alexander I supported the protest, and Bern was forced to agree to an enlarged federation. The details were worked out in the "Long Diet" of 1814.

From the Congress of Vienna (1814–15) the Swiss hoped to obtain a recognized frontier, with the provinces annexed by France restored, compensation for damages inflicted by the war, and a recognition of neutrality and independence. The congress decided that the Valtellina, Chiasso, and Bormio should not return to the confederation and that the Pays de Gex should not come under the control of Geneva. Aargau and Vaud secured recognition, the former territories of the bishop of Basel were divided between Basel and Bern, Neuchâtel remained a Prussian principality within the Swiss confederation, the French annexation of Valais was denounced, and Geneva was admitted with narrowly defined frontiers, north Savoy and the Pays de Gex being accepted as custom-free zones. Swiss neutrality was guaranteed at Paris on November 20, 1815, by France, Austria, Great Britain, Portugal, Prussia, Sweden, and Russia.

Guarantee of Swiss neutrality

A new conservative Swiss constitution (*Der Bundesvertrag*) was signed on August 7, 1815. There were 22 cantons, each with its own constitution and each contributing to a federal army; those with external frontiers were allowed to levy customs duties on imports. Each canton had one vote in the Diet, which alone could (by a two-thirds majority) declare war, conclude peace, and enter

into alliances with foreign powers. On all other matters a simple majority was sufficient. Bern, Luzern, Fribourg, and Solothurn retained, with minor modifications, their patrician-ruled status. Zürich, Basel, and Schaffhausen widened their guild organizations. The remaining cantons adopted semidemocratic forms of government.

Switzerland to 1914. Between 1815 and 1830, despite economic difficulties and complaints that political refugees (including Louis Napoleon) were being given asylum, peace and stability reigned. Communications were greatly improved; commerce, including banking, advanced. New demands were heard for freedom of the press, religious toleration, rights of association, and greater free trade. After the July Revolution in Paris (1830), pressure for a more liberal Swiss constitution grew steadily; objections to the domination of the countryside by the towns were frequently heard, together with demands that popular sovereignty should become a reality. In Neuchâtel voices were raised against Prussian conservatism. In Schwyz the peasants threatened to form a breakaway state; and in Basel this actually took place, Basel-Landschaft separating from Basel-Stadt in 1833. Revisionists called for a representative parliament, a federal postal system, coinage, and a foreign office.

The Sonderbund War. The secularization in January 1841 of the monasteries of Aargau was condemned by the federal diet, which in turn led to numerous protests. When the Jesuits came to Luzern in 1844 there were demands for their exclusion from the confederation. Radical "volunteers" attempted to expel the Jesuits by force. The Catholic and conservative cantons refused to be intimidated and stood by their rights as guaranteed in 1815. In March 1845 volunteers from Bern marched against Luzern to expel the Jesuits and were easily repulsed. The result was that Luzern, Uri, Schwyz, Unterwalden, Zug, Fribourg, and Valais formed a protective union for mutual support—another breach of the Federal Pact. This soon became a *Sonderbund* (a separate federation), which was declared illegal in July 1847, when the diet resolved to inaugurate a new constitution and to drive out the Jesuits. The *Sonderbund* confidently hoped for support from the rulers of Austria, Prussia, and France. Civil war was imminent. In October a federal army under the command of Henri Dufour took the field against the rebels, whose expectation of outside assistance was disappointed. Dufour, with far superior forces, showed his strategic skill. Fribourg capitulated on November 14; Zug on November 21; and, after some fighting at Meierskappel, Hanau, and Gisikon, Luzern was entered on November 24. Schwyz, Uri, and Valais submitted in quick succession. The revolt was over within a month: the seven insurgent cantons had to meet the cost of the campaign; southern Appenzell and Neuchâtel were fined for failure to contribute to the army; Luzern and Fribourg were obliged to exclude Jesuits; and a revision of the constitution was worked out.

Formation of a separate federation

The new federal state, 1848–60. A government by two chambers was formed—a national council based on numerical representation and a senate, or council of states (Ständerat), consisting of two members from each canton. Together they formed a federal assembly (Bundesversammlung), with a federal council, or cabinet (Bundesrat), chancery (Bundeskanzlei), and supreme court of justice. The president, elected annually, was simply first among equals. Relations with foreign powers were thenceforth the exclusive affair of the federal government; mercenary service and the acceptance of pensions, gifts, tithes, or orders from foreign powers was forbidden. Equality before the law and freedom of the press, of movement, of association, and of religious worship were guaranteed to all. The Jesuits remained excluded, and clergymen could not hold elected offices. Bern was named as the capital, and the Swiss Federal Institute of Technology was opened in Zürich in 1855. Customs ceased to be levied between the cantons, and Switzerland became a unified economic area. The postal system became federal, as did currency, weights, and measures. In August 1847 the opening of the first Swiss railway, between Zürich and Baden, brought new problems—the country lacked coal; there were fierce disputes about capital, contracts, and ownership; and a struggle for control developed. Later the railway system was nationalized.

External affairs. Help to Charles Albert of Sardinia to eliminate the Habsburg regime was refused; even so, the troubles of 1848 in the Grand Duchy of Baden, in northern Italy, and in Neuchâtel nearly involved the confederation in war in defense of political refugees. In Neuchâtel —a canton of the confederation and a principality ruled by the king of Prussia—a republic was proclaimed; but in 1856 a "royalist" reaction gained temporary control and appealed to Frederick William IV of Prussia. The situation was soon reversed, and the royalist leaders were imprisoned, but it required the personal intervention of Napoleon III and the threat of a German-Russian-Austrian invasion to settle the matter. All Prussian rights in Neuchâtel were abandoned (for a consideration), although the monarch retained the empty title of prince.

Proclamation of a republic in Neuchâtel

Swiss neutrality had contributed to the Austrian defeats at Magenta and Solferino (1859). There were Swiss protests at the annexation of Savoy by France in 1860, but again a compromise had to be accepted not altogether to their advantage or satisfaction. Swiss neutrality narrowly escaped infringement during the Franco-Prussian War of 1870–71, when a French army entered Swiss territory, was disarmed and interned, but was allowed to return after peace was restored.

This was followed by renewed railway construction and continued controversy over private ownership, although it was not until 1882 that technical and financial difficulties accompanying the opening of the essential St. Gotthard route were overcome (an indirect consequence was the final abolition of separate cantonal currencies in 1874).

Religion and politics, 1860–1914. The final incorporation of the States of the Church into the kingdom of Italy and the Vatican Council of 1870, followed by the *Kulturkampf*—the conflict between state and church in Germany—had repercussions in Switzerland, where the "Old Catholics" were numerous. Obligatory civil marriages were introduced, in spite of opposition from the Catholic bishop of Basel and the papal representative at Geneva. In 1873 diplomatic relations with the Vatican were broken off (and not resumed until 1920). "Popular Initiative" had come in 1867, and the principle of the referendum was now accepted, complete internal free trade was assured, federal control over the army was extended, legal procedure was simplified, and greater economic and educational uniformity were furthered by constitutional revision in 1874.

As elsewhere, 19th-century Switzerland saw the rise of working class influences on the political scene. Home industry was being displaced by factories, where long hours were worked, often by small children. This trend brought about outbreaks of machine smashing by those who feared the loss of both livelihood and independence. Attempts at cooperative unions such as the Grütliverein, formed in 1838, were followed by the formation of trade unions proper. Successful demands for factory legislation were made by the Social Democrats, who emerged as a party of reform between 1870 and 1888.

Rise of working class influences

There were abiding difficulties over education, a sphere in which the cantons, determined to retain controlling influence, resisted federal coordination. In the Ticino local political rivalries almost led to a revolution in 1890 which was prevented by federal intervention. At the same time, the prosecution of German agents or spies sent by Bismarck caused some trouble with Germany, and the traditional Swiss acceptance of political refugees was resented by tsarist Russia. In each case the Swiss government successfully maintained its position.

Switzerland since 1914. Beginning with World War I, Switzerland faced a number of problems and changes during the 20th century.

The War of 1914–18. The advent of World War I in 1914 brought great problems: foreign visitors and workers left the country, which was obligated to defend its neutrality and independence. Tension between the French-speaking minority and the pro-Germans was in-

tensified when evidence was produced of secret information being sent by high-ranking Swiss officers to Germany. Difficulties were created by the considerable number of political refugees (including Lenin) living in Switzerland and by well-meant but ineffective attempts at peace mediation between the combatants. Receipts from Swiss customs dwindled, and an income tax became a necessity as defense costs mounted. Price fixing and rationing were also introduced. Discontent culminated in a general strike in November 1918, inflamed by Marxist elements and accompanied by an influenza epidemic that swept across Europe. Firm government action and the loyalty of the army quickly broke the strike, but political concessions, such as proportional representation, followed. The Soviet embassy was closed.

Peace in Europe brought its own special problems. France demanded, and obtained, full sovereignty over Savoy, reluctantly conceding a free zone. An attempt by the inhabitants of the Vorarlberg to become a 23rd canton of the confederation failed, but Liechtenstein was accepted as an ally and brought within the Swiss postal, customs, and monetary union.

The League of Nations. The League of Nations was formed in May 1920 with its headquarters at Geneva. The Declaration of London (February 1920) recognized the special nature of permanent Swiss neutrality and exempted the Swiss from participation in any future military "sanctions." Switzerland, however, could not be entirely indifferent to or unaffected by Italian Fascism and German National Socialism; there were deep divisions between the French- and German-speaking cantons and between the conservative "bourgeois" elements and the radical socialists. Something of this was reflected in the party disputes between the Catholic-Conservatives, Liberal-Democrats, and Social Democrats. Elections to the Federal Council were hotly disputed, while youthful adherents to a "National Front" rejected traditional democratic procedures in favour of obedience to a leader. Continuous demands for female suffrage from 1884 onward were rejected (until finally met in 1971).

Divisions within Switzerland

The world depression of the 1930s affected Switzerland less severely than it did the great industrial countries, but the same results on a reduced scale were apparent—bankruptcies, overproduction, and poverty. German claims to rule over all German speakers and German anti-Semitic legislation caused much alarm, as did Communist demonstrations—at one of these in Geneva (November 1932), 16 persons were killed and 60 wounded. Suggestions of a total revision of the constitution were decisively rejected, but agreements favouring industrial peace were readily entered into.

The admission of the Soviet Union to the League of Nations (1934); the Italian invasion of Ethiopia (1935), involving Switzerland in economic sanctions; and the Spanish Civil War (1936), with individual Swiss volunteers serving in the International Brigade, caused disillusionment with the League of Nations. Fascist and Nazi propaganda was countered and military service was made more acceptable. The German occupation of Czechoslovakia in March 1939 intensified defense measures so that the outbreak of war, following Hitler's invasion of Poland at the end of August, saw the Swiss fully prepared.

World War II and after. About 430,000 men were mobilized for the defense of the Swiss frontiers, rationing was introduced, and arrangements were made for adequate payment for soldiers on active service. Disapproval of the Soviet invasion of Finland was evidenced by the collection of 4,000,000 Swiss francs for war sufferers there. The German occupation of Denmark and Norway strengthened determination to resist if attacked, and more men were called to the colours. After the fall of France (when 42,000 men were interned), followed by intensified pro-German propaganda in Switzerland, the army command decided to defend, in case of invasion, only the "*réduit*," the heartland, including the St. Gotthard and Simplon tunnels.

Determination to resist if attacked

The expected invasion did not take place; instead, amid great economic difficulties and by supplying its neighbours with some equipment (chiefly precision instruments) and by assisting them indirectly in air warfare (and by suffering from mistaken air attacks, as when the Americans bombed Schaffhausen in April 1944), the Swiss rode out the war. At its end there were more than 100,000 refugees or internees within the country. The determination of the Allies that "war criminals" and high Nazi officials should not take refuge with their property in Switzerland caused difficulties, as did the reopening of diplomatic relations with the Soviet Union.

After 1945 Switzerland adopted by plebiscite many of the activities of the welfare state—old age pensions and social insurance of various kinds—and every effort was made to sell Swiss manufactured goods abroad, to attract visitors, and to forward "invisible exports" by banking, insurance, and other services. Without becoming a member of the United Nations, Switzerland participated fully in the Organization for European Economic Co-operation (OEEC) and in the General Agreement on Tariffs and Trade (GATT), and Basel became the centre for the World Bank.

By the 1970s more and more of the population had moved to or near the five great cities. Women were given the right to vote for the first time in 1971, and the country's general prosperity made the immediate future seem secure.

BIBLIOGRAPHY. General histories include: J. DIERAUER, *Geschichte der schweizerischen Eidgenossenschaft*, 4th ed., 6 vol. (1924–31), standard but antiquated; E. GAGLIARDI, *Geschichte der Schweiz von den Anfängen bis zur Gegenwart*, 3 vol. (1939), detailed; H. NABHOLZ *et al.*, *Geschichte der Schweiz*, 2 vol. (1932–38), scholarly, written by specialists; J.G. GUGGENBUHL, *Geschichte der schweizerischen Eidgenossenschaft*, 2 vol. (1947–48), a reliable narrative; P. DURRENMATT, *Schweizer Geschichte* (1957); W. DRACK (ed.), *Illustrierte Geschichte der Schweiz*, 3 vol. (1958–61), admirable illustrations; E. BONJOUR, *Geschichte der schweizerischen Neutralität*, 2nd ed., 6 vol. (1965–71), an important study; E. BONJOUR, *Geschichte der Schweiz im 19. und 20. Jahrhundert* (1938); E. BONJOUR, H.S. OFFLER, and G.R. POTTER, *A Short History of Switzerland* (1952); M. GODET, H. TURLER, and V. ATTINGER (eds.), *Dictionnaire historique et biographique de la Suisse*, 7 vol. and suppl. (1921–34).

The following English-language works are recommended: W. MARTIN, *Histoire de la Suisse, essai sur la formation d'une confédération d'états* (1926; Eng. trans., *A History of Switzerland: An Essay on the Formation of a Confederation of States*, 1931); EDGAR BONJOUR, *Geschichte der schweizerischen Neutralität* (1946; Eng. trans., *Swiss Neutrality: Its History and Meaning*, 1946); and W. OECHSLI, *History of Switzerland, 1499–1914* (1922), rather antiquated.

(G.R.P.)

Sydney

The capital of New South Wales, Sydney is the oldest and, with a population of over 2,700,000 by the early 1970s, largest city in Australia. Because of its magnificent harbour and strategic position on the southeast coast of Australia, it is also one of the most important ports in the South Pacific. At the beginning of the 19th century, while it was still a small convict settlement and, indeed, even before the first settlers had penetrated 50 miles into the interior, it had already established trade with the Pacific Islands, India, China, South Africa, and the Americas.

The character of the city

The first sight of Sydney, whether from the sea or the air, is always impressive. Built on low hills surrounding a huge harbour with innumerable bays and inlets, the city is dominated by the bulk of the Sydney Harbour Bridge, one of the biggest single-span bridges in the world. The glittering confusion of water and buildings makes a striking impression either by day—under Sydney's brilliant light and blue skies—or by night.

Situated on latitude 34 degrees south and with an average mean temperature ranging from 72.1° F (22.3° C) in February to 54.1° F (12.3° C) in July, Sydney enjoys a warm, sunny, but temperate climate that has encouraged its citizens to develop a pleasure-loving and easy-going attitude to life and to make full use of the opportunities for sailing, swimming, and surfing at their doors. Although the average annual rainfall is 44.8 inches (1,138 millimetres), much of this falls in short tropical deluges in summer. The humid heat of summer, when



Sydney Harbour and commercial centre. At left is the Sydney Harbour Bridge; the sail-like structure on a point of land at left-centre is the city's new Opera House.
Helmut Gritscher—DPI

the prevailing wind is from the northeast, is tempered from time to time by the arrival of a cold front from the Tasman Sea heralded by a stiff wind from the south known locally as the "Southerly Buster." Sydney is unbearably hot only on a few days each year when a westerly wind brings the hot, dry air from the inland. In winter, however, the westerly wind is cool.

Because of its history as a great port and its more recent status as the site of the country's main international air terminal, Sydney is perhaps the only city in Australia with a genuinely international atmosphere. Most visitors to Australia begin and end their stay in Sydney. Yet it remains a very Australian city, hectic but casual, rough but good-natured, with a nice compromise between the Anglo-Saxon efficiency of its British heritage and the South Seas attractions of its climate and environment. (For related information see NEW SOUTH WALES.)

The history of Sydney. When the English explorer Capt. Arthur Phillip arrived off the coast of southeast Australia with the First Fleet in 1788, he sailed first to Botany Bay, which had been discovered by Capt. James Cook in 1770 and to which he had been directed by the British government. Finding the bay too exposed for safe anchorage and the surrounding country unsuitable for settlement, he looked further north and soon discovered the entrance to Port Jackson (now Sydney Harbour) only a few miles away. Phillip's first impressions of Port Jackson, which had been named but not explored by Cook, are recorded in a famous dispatch to Lord Thomas Townshend Sydney, then Home Secretary, dated May 15, 1788.

Discovery of the harbour

> We got into Port Jackson early in the afternoon, and had the satisfaction of finding the finest harbour in the world, in which a thousand sail of the line may ride in the most perfect security. . . .

Captain Phillip immediately decided to move the whole fleet to Port Jackson and to establish the first settlement on a cove, which had a good freshwater spring and in which his ships could anchor close to the shore in deep water. He called it Sydney Cove, after the Home Secretary. Today, Sydney Cove, where passenger liners tie up at the international terminal, is still the city's heart, though it is more commonly known as Circular Quay. The main commercial port, however, has been moved to Darling Harbour, another bay a little to the west.

The early history of Sydney was grimly dominated by its existence as a British penal colony. Convicts, dumped on this alien shore, found the environment a harsh one. The soil was poor, and the land was rough and had to be cleared by hand. The little settlement was often short of food until the settlers were able to cross the Blue Mountains and find the richer land to the west of the Eastern Highlands. There were also constant troubles between the governors, the free settlers, the convicts, and the marines and soldiers sent to guard them.

With the exploration and settlement of New South Wales, Sydney grew quickly; the British government provided free land, free convict labour, free capital works, and guaranteed markets for the produce of the new colony. Trading links with the rest of the world were quickly established. Under the enlightened governorship of Lachlan Macquarie (1810 to 1821), Sydney developed from a precarious penal settlement into a thriving, respectable town. Macquarie also began a program of public works, including the building of churches, hospitals, barracks, schools, and courthouses, and laid out several parks in and around the city. In this work he was aided by a convict-architect, Francis Greenway, who had been banished for forgery in England. Greenway built several fine buildings in the Regency style, notably the Hyde Park Barracks on Macquarie Street, and St. James's Church, which survive today.

Nineteenth century growth

The really rapid growth of Sydney—from 60,000 to 400,000 population—took place between 1850 and 1890, as suburbs of tightly packed terrace houses were built. These houses, with their balconies and cast iron railings in floral and geometrical patterns, are today Sydney's most attractive heritage from the past. The first railway, from Sydney to Parramatta, was opened as early as 1855.

The financial collapse of the 1890s acted as a slight check to Sydney's growth, but population doubled again by 1914 and reached the million mark soon after. Yet during this period Melbourne (*q.v.*), the capital of Victoria, was growing still faster, partly as a result of the gold rush in that colony. Temporarily overtaking Sydney both in size and importance, it became the financial centre of

Australia and the seat of the first Commonwealth Parliament from 1902 until the Federal Capital of Canberra was built halfway between the two cities. It was not until after World War II that Sydney again took the lead, partly as a result of a shift in Australia's trade toward North America and Asia and away from Britain. The Sydney of the early 1970s is slightly bigger than Melbourne and is equally important as a centre of finance, commerce, and manufacturing.

Other landmarks in the history of Sydney were the opening of the city underground railway in 1926, the Harbour Bridge in 1933, and the Opera House, a magnificent arts centre designed by a Danish architect, Jøern Utzon, in 1973.

The contemporary city. Greater, or metropolitan, Sydney is spread over an area of 670 square miles (1,735 square kilometres) and stretches from the Blue Mountains in the west to the Pacific Ocean in the east and from the Hawkesbury River in the north to south of Botany Bay. This suburban sprawl, caused partly by the cheapness of land in earlier days and by the ordinary Australian's determination to own his own house and garden, has caused great problems to the authorities responsible for sewerage and transport. (Several suburbs in Sydney were still unsewered by the early 1970s.) The sprawl is also in marked contrast to the comparatively small and compact central business district, which is crammed into a narrow rocky peninsula between two arms of the harbour.

Layout and boundaries

The same contrast can be seen in the style of buildings. Whereas most of the houses in the suburbs are one-story bungalows built of brick or wood, the high cost of land in the central business district has forced occupiers to build high. The rapid development initiated in the 1960s has transformed the city centre, which now looks like a smaller Manhattan rising from the shores of the Harbour, though in fact the buildings are not high by North American standards. A rival business centre is now rising in North Sydney, which is linked to the city by the Harbour Bridge.

The standard of architecture is generally mediocre, though there are a few handsome buildings surviving from the 19th century and a few contemporary buildings of distinction. Yet the general effect of Sydney is attractive, and the innumerable bays and arms of the harbour, stretching into the land and providing unexpected views from almost every street, make Sydney one of the most pleasing cities in the world.

Sydney is also fortunate in having two magnificent national parks within 25 miles of the city centre to the north and south. In these parks the natural Hawkesbury Sandstone bush can be seen at its best and forms a refuge for the wild life of the area. The mouth of the Hawkesbury River, with its many finger-like inlets, provides a superb expanse of sheltered water for yachtsmen, while both to the north and south of the city there stretches a string of magnificent sandy beaches, on which the Pacific surf breaks throughout the summer.

The people. By the early 1970s the population of greater Sydney was over 2,700,000. As in the rest of New South Wales, about 80 percent of these are of British stock and the remainder immigrants from Europe. These immigrants, however, have tended to crowd into the inner suburbs, some of which now have a distinctly Greek or Italian character.

Ethnic origins

The city is still growing rapidly, partly as a result of Australia's immigration policy and partly as a result of the decline of the rural population. In 1968 the State Planning Authority estimated that the city of Sydney would have grown to 5,500,000 by the year 2000. Some of this increase, however, would have to be housed in a wider Sydney region extending beyond the present boundaries of the metropolitan area. Some planners fear that this will lead to a vast conurbation stretching from Newcastle in the north to Wollongong in the south. These three cities already contain over two-thirds of the population of New South Wales.

Economic life. About 43 percent of the total Sydney population is in the work force, and of these about a third are employed in manufacturing (mostly in secondary industry) and the rest in tertiary industry or services. There is no major industry in the city itself, though oil refining is becoming increasingly important. Sydney has always been essentially a city of government; of local, national, and international commerce; and a shopping and entertainment centre for the State of New South Wales. More recently it has rivalled Melbourne as a financial centre, and the Sydney Stock Exchange is now the largest in Australia.

Sydney is still the largest port in Australia handling about 21,000,000 tons of cargo a year. A second port is being developed in Botany Bay, which is becoming the main industrial area of Sydney.

Political and governmental institutions. Sydney is the capital of New South Wales and the seat of its government and parliament. The state government has delegated matters of local interest to elected local councils, including the City Council, but has retained direct control of some important matters, such as police and transport. In addition, a number of functions normally exercised by local councils have been transferred to boards, commissions, trusts, and departments established by the state government. The most important of these responsibilities concern water, sewerage and storm drainage, electricity, major roads, fruit and vegetable markets, and many aspects of health and welfare. The result is a somewhat confused and complicated system of government. For historical and political reasons, moreover, the boundaries of the City Council have been changed several times. In 1968 the City Council was reduced to approximately its original boundaries so that today it is responsible only for the central business district and the inner suburbs. The outer suburbs elect their own councils.

Relations with the state government

Even within the area controlled by the City Council, large areas of land are controlled either by the state government or by the Commonwealth Government. This fact has greatly added to the difficulties of planning.

Services and transportation. Sydney has had a railway linking the city to some of its suburbs since 1855. Between World Wars I and II, this radial suburban rail system was electrified, and construction of the city underground railway was begun (the first section was opened in 1926). In 1933, after the completion of the Harbour Bridge, the north shore was linked directly by rail to the city. By the 1970s the underground railway was being extended to the eastern suburbs.

In spite of development and the existence of an extensive metropolitan bus service, public transport has failed to keep pace with demand. Since World War II, Sydney, like so many other cities, has fallen a victim to the motor car. The number of motor vehicles registered in New South Wales increased from almost 330,000 in 1939 to nearly 2,000,000 in 1970, and a high proportion are based in the Sydney metropolitan area. By the late 1960s over 580,000 vehicles entered or left the central business district on a typical day.

Traffic congestion

The resulting traffic problem is made worse by the geographical layout of Sydney. A high proportion of those who work in the central business district live in the outer suburbs and must find their way to and from the city by a limited number of routes. All those living on the North Shore, for instance, must cross the harbour by the Harbour Bridge, the only bridge crossing the harbour from north to south. (At the head of the harbour there is a second bridge, at Gladesville, which takes some of the traffic.) It is also difficult for traffic travelling from north to south through Sydney to bypass the city. The inevitable consequence is severe traffic congestion aggravated by the narrowness of the streets. Attempts to relieve this by building expressways and thruways have been only partially successful, and Sydney, like other cities, has begun to think of banning all traffic from certain sections of the inner city: to this end, a pedestrian plaza was created in Martin Place, in the centre of the city, in 1971.

A delightful alternative means of transport exists for those who live on the shores of the harbour in the public ferry system, but, for reasons that have always baffled the authorities, this has never been as popular or as profitable

as might be expected. The ferry boats crisscrossing the harbour nevertheless add a good deal to the attraction and amenity of the city.

The main airport, both for international and internal traffic, is at Mascot, on the shores of Botany Bay, to the south of the city, but this, too, is proving inadequate to the growing demand, and a second airport is planned for the late 1970s. Access to and from the present airport is poor. On the other hand, travellers by sea are fortunate in that passenger liners generally tie up at Circular Quay in the very heart of the city, from which they can stroll to the city centre.

Sydney has only recently become a centre for international tourists. The number of first-class hotels is still inadequate, and they are not of the highest standard, but Sydney has many excellent motels offering accommodations at a moderate price. There are a great many restaurants, particularly in the eastern suburbs, many of them quite good and a few excellent. It is possible to find restaurants offering almost every kind of international cooking, though all of them serve good Australian wine and Sydney's famous oysters. The entertainment quarter of Sydney and the centre of its night life is King's Cross. This is a small, crowded, and somewhat sleazy area that has nevertheless had a certain irresistible gaiety and vitality for the two generations of American soldiers stationed in Australia during World War II or sent there from Vietnam on rest and recreation leave.

The role of tourism

Cultural life and recreation. Sydney is, above all, a city for physical enjoyment. Few great cities in the world offer such unrivalled opportunities for swimming, surfing, sailing, and other outdoor sports. It is still possible to swim from several beaches in the harbour itself—though pollution is slowly reducing the number of these—and the surf beaches to the north and south of the city are world famous. One of these, Bondi, is a 15-minute drive from the city centre. Some of the beaches to the north of the city, especially those on the Palm Beach peninsula, are very beautiful. The surf is good, though variable, and made comparatively safe by the voluntary lifesaving association, which provides guards and rescue equipment on all beaches throughout the summer months. Sydney is also a city of yachtsmen. On any weekend Sydney Harbour is covered with sails of every shape and size, while the Pittwater, 25 miles to the north, is almost equally popular. There are also many golf courses and excellent facilities for football, cricket, and tennis, and the racecourse at Randwick is first class.

Sporting traditions

Australians have been accused of being sports-mad, and Sydney-siders, as they call themselves, are certainly no less devoted to sport than their compatriots in other cities. Yet Sydney also has a vigorous cultural and intellectual life. There are three universities in the metropolitan area; the University of Sydney, the University of New South Wales, and Macquarie University. A fine public library, two museums, and an art gallery (to which an extension was added in 1971), and a large number of private galleries devoted to contemporary Australian art add to the cultural attractions. A serious shortage of theatres and concert halls existed in the city, but this was remedied when the controversial Opera House—in fact a major arts centre—was opened in 1973. This beautiful and costly building was designed by a Danish architect, Jøern Utzon, who won an international competition held for the purpose in 1955. Built on a magnificent site on the harbour, surrounded by water on three sides, its white shells have already become almost as much a symbol of Sydney as the Harbour Bridge. It contains a concert hall for the Sydney Symphony Orchestra, a large theatre for opera and ballet, a smaller theatre for plays, a cinema, and rehearsal and recording studios. Until the centre's completion, the Sydney Symphony Orchestra had to play in the Town Hall, which was acoustically inadequate, but the Orchestra has long maintained a high standard. For opera and ballet Sydney depends on the Australian national companies, which visit the city for a season each year. Sydney has several small theatre groups, but none has yet reached the highest standard.

Four morning papers, two evening papers, and four Sunday papers are published in the city. Television and radio are competitive. In addition to the Australian Broadcasting Company, which has its headquarters and main studios in Sydney, there are three commercial television stations and six commercial radio stations.

Thus, though Sydney today is a large, sophisticated, international city offering services and cultural amenities comparable with those offered by any great city in the world, there is still no doubt that its most lasting impression and greatest attraction is its physical setting, in which a very high proportion of Sydney's residents can glimpse at least a strip of blue water from their windows and nearly all can live within half an hour or an hour of a beach.

BIBLIOGRAPHY. The best sources for facts and figures about Sydney are *The Sydney Region Outline Plan*, published by the State Planning Authority of New South Wales (1968); and the *City of Sydney Strategic Plan*, published by the Council of the City of Sydney (1971). A charming portrait of contemporary Sydney, unencumbered by statistics, is GAVIN SOUTER, *Sydney* (1965), with photographs by QUINTON DAVIS. The early history of the convict settlement is graphically told from contemporary documents in JAMES JOHNSTON AUCHMUTY, *The Voyage of Governor Phillip to Botany Bay* (1970); and JOHN COBLEY, *Sydney Cove, 1788* (1962). All early histories of Australia necessarily deal with Sydney in part. There are good biographies of *Lachlan Macquarie* (1942) and *Francis Greenway* (1949), both by M.H. ELLIS.

(J.D.Pr.)

Syllogistic

Syllogistic is the study of the logical structures that make it possible to infer a proposition (a conclusion) from other given propositions (the premises). Its foundations were laid by Aristotle in his book *Analytica priora*. Though Aristotle defined the syllogism broadly as "an argument, in which, certain facts having been laid down, something different follows with necessity simply from their being so," he himself investigated almost exclusively the so-called categorical syllogism; *i.e.*, a type confined to simple declarative statements and its variations involving expressions of necessity and possibility, known as modalities. Later, other forms of syllogistic reasoning and especially those of what is called propositional logic (*i.e.*, a logic using unanalyzed whole propositions as its units) were dealt with by Stoic logicians. Though the European tradition usually followed Aristotle in trying to transpose all other forms of syllogistic reasoning into categorical syllogistics, important differences developed between the true Aristotelian syllogism and the mainstream of later traditional logic. In this article, the main emphasis will be placed upon the accepted results of modern research in this field.

NATURE AND ELEMENTS OF SYLLOGISTIC

Since the propositions that compose all syllogisms are built up from terms and operators (signs expressing logical connections), traditional logic begins with an analysis of terms. Terms may be singular (such as "Plato" or "the author of *Hamlet*") or general (such as "man," "triangle," or "healthy"). General terms may be taken either in extension (also called denotation), which indicates the set of individuals to which a term applies, or in intension (also called connotation), which describes the set of attributes that something must have to qualify for the predication of the term. The extension, for example, of "bachelor" is the set including all men that are bachelors, and its connotation is the set of such attributes as being human, being male, being above the legal age of marriage, and being single. If the connotation of a term *a* includes that of *b*, the extension of *b* will include that of *a*. Whereas the class of bachelors is included in the class of men, the attributes of "man" are included in the class of attributes of "bachelor." The definition of a term is an analysis of its connotation.

Basic definitions

To form a proposition of the most primitive sort, a singular term (the subject) may be joined to a general term (the predicate) by "____ is (a) ____" or "____ is not (a) ____," signs called copulae. "Plato is a philosopher" is called an affirmative proposition, and "Plato is not

stupid" is called a negative proposition. General terms cannot form propositions merely by being joined by copulae; in order to construct a complete proposition, a sign of quantification must be added—traditionally "every" or "some," as in "Every philosopher is a man." Instead of writing "Every man is not perfect," it is customary to write "No man is perfect," the "no" doing double duty as a sign of negation *and* quantification. A proposition of the type "Every ____ is (a) ____" is called universal affirmative; "No ____ is (a) ____" is called universal negative; "Some ____ is (a) ____" is called particular affirmative; and "Some ____ is not (a) ____" is called particular negative. These forms of propositions are also referred to, in that order, as *A*-, *E*-, *I*-, and *O*-propositions, the vowels being taken from the Latin words *affirmo* ("I affirm") and *nego* ("I deny"). Negative and affirmative propositions are said to differ in *quality;* universal, particular (and singular) propositions are said to differ in *quantity*.

Truth and validity of inference

A proposition is said to be *inferred* from other propositions if it is validly drawn as a conclusion from them. A conclusion is validly drawn if it is a logical consequence of the premises, and this is the case whenever it is impossible that the premises be true and the conclusion false. The following cases are therefore possible:

1. A true conclusion invalidly drawn from true (or false) premises,
2. A false conclusion invalidly drawn from true (or false) premises,
3. A true conclusion validly inferred from true (or false) premises, and
4. A false conclusion validly inferred from false premises.

The only case that cannot arise is:

5. A false conclusion validly inferred from true premises.

Therefore, to know the truth or falsity of the premises and the conclusion does not enable one to determine the validity of an inference. To judge the validity of an argument, one must examine its logical form, which is what syllogistic teaches him to do in the proper way.

PROCEDURES OF IMMEDIATE INFERENCE

In expressing syllogistic propositions, logicians find it convenient to employ a notation introduced by Jan Łukasiewicz, an important early-20th-century logician: lowercase Latin letters *a*, *b*, *c*, . . . are employed to stand for general terms (*i.e.*, for term variables); capital letters *A*, *E*, *I*, and *O*, are then used as symbols for the syllogistic operators, which, together with the terms, make up the four kinds of propositions. Accordingly, "Every *b* is an *a*" is written "*Aba*"; "Some *b* is an *a*" is written "*Iba*"; "No *b* is an *a*" is written "*Eba*"; and, finally, "Some *b* is not an *a*" is written "*Oba*."

Square of opposition

Relations among propositions. In order to display the logical relations obtaining between different pairs of propositions of these four forms, they are customarily arranged in a so-called square of opposition (Figure 1). It is important to note that, in considering these different forms of opposition between propositions, the logical tradition conventionally distinguishes empty terms (such as "round square," "centaur," or "*perpetuum mobile*"), which cannot be truly applied to any individual, and excludes them from being introduced in the place of term-variables in syllogistic propositions. It is also necessary to be aware that the operators *I* and *O* ("Some ____ is (a) ____" and "Some ____ is not (a) ____") are taken in the sense of "At least one and possibly all ____ is (or is not) (a) ____." With these specifications made, it may be seen that "*Aba*" and "*Eba*" are contraries—*i.e.*, they cannot be jointly true; that "*Iba*" and "*Oba*" are subcontraries—*i.e.*, they cannot be jointly false; that "*Aba*" and "*Oba*" and also "*Eba*" and "*Iba*," taken pairwise, are contradictories—*i.e.*, if "*Aba*" is true, "*Oba*" must be false and vice versa; and, finally, that if "*Aba*" is true, "*Iba*" must also be true, and if "*Eba*" is true, so is "*Oba*." In the last cases, however, since the implication does not hold analogously in the opposite direction, it is said that "*Aba*" "subalternates" "*Iba*," and "*Eba*" subalternates "*Oba*," "*Aba*" being the subalternant and "*Iba*" the subalternate (and, similarly, for "*Eba*" and "*Oba*"). For any given pair of terms *a* and *b*, exactly two of these four forms of proposition are true: for every pair of terms "*a*" and "*b*," either "*Aba*" and "*Iba*" will be true or "*Eba*" and "*Oba*" or, finally, "*Iba*" and "*Oba*."

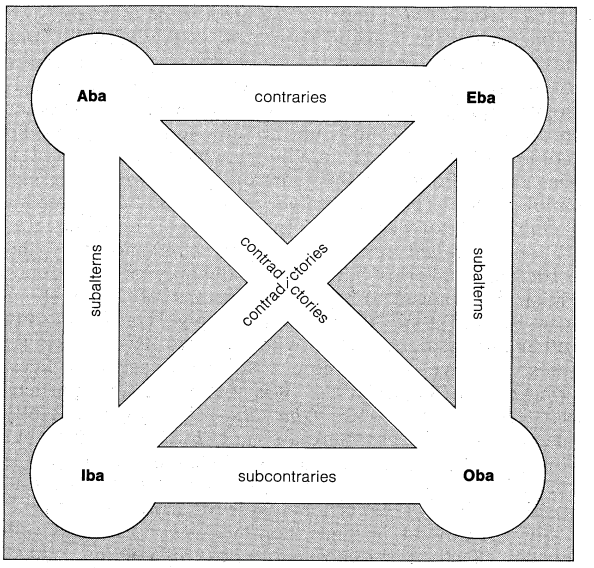


Figure 1: The square of opposition (see text).

If the extensions of general terms are symbolized by circles, it is possible to develop a series of so-called Euler diagrams (Figure 2)—named after the Swiss mathematician Leonhard Euler, who promulgated the method (though he had not invented it). The diagrams show how the five possible relations of terms may be visualized: α: exactly matching; β: *b* included in *a*; γ: *a* included in *b*; δ: partially overlapping; ε: entirely distinct. As long as negated term variables—*e.g.*, "not-man" or "nonsmoker"—are not considered, the five diagrams exhaust the possibilities. These figures will make clear the logical relations between pairs of propositions that are alike in quality but have the order of the terms *a* and *b* inverted: ("*Aba*," "*Aab*"), ("*Aba*," "*Iab*"), ("*Iba*," "*Aab*"), ("*Iba*," "*Iab*"), ("*Eba*," "*Eab*"), ("*Eba*," "*Oab*"), ("*Oba*," "*Eab*"), ("*Oba*," "*Oab*"). As the former logical relations were traditionally called those of opposition, the ensuing relations are those of conversion.

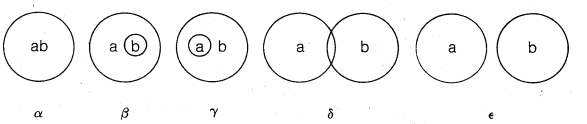


Figure 2: The Euler diagrams (see text).

Conversion

The identification of logically valid conversions is facilitated by the use of so-called truth tables (Table 1).

Table 1: Truth Table Correlating Propositional Forms with Euler Diagrams

Euler diagram	propositional form				converse propositional form			
	Aba	Eba	Iba	Oba	Aab	Eab	Iab	Oab
(α)	t	f	t	f	t	f	t	f
(β)	t	f	t	f	f	f	t	t
(γ)	f	f	t	t	t	f	t	f
(δ)	f	f	t	t	f	f	t	t
(ϵ)	f	t	f	t	f	t	f	t

A "t" in a given line and column signifies that the propositional form given at the top of the column will be true for the Euler diagram indicated at the left by Greek letters (referring to the different diagrams of term relations in Figure 2); an "f," on the other hand, signifies that the propositional form will be false. The columns for "*Eba*" and "*Eab*" are identical, as are those for "*Iba*" and "*Iab*." "*Eab*," therefore, may safely be inferred from "*Eba*," and

"*Eba*" from "*Eab*"; and "*Iab*" may also be inferred from "*Iba*," and "*Iba*," again, from "*Iab*." Traditionally, this logical fact is expressed by saying that *E*- and *I*-propositions are capable of simple conversion. Generally, a syllogistic proposition *Y* may be inferred from another proposition *X* built up from the same terms whenever the column of *X* has no "t" in a line in which the column of *Y* has an "f." Following this criterion it is shown that "*Aba*" implies "*Iab*" (but not vice versa) and that "*Eba*" implies "*Oab*" (but not vice versa). In this kind of inference (traditionally called *conversio per accidens*) not only the order of the terms but also the quantity of the proposition is changed. From the columns of Table 1, it can be seen that "*Oba*" cannot imply "*Oab*," nor any other propositional form with inverted terms, because these all have one or more "f's" in the last three lines, where "*Oba*" has only "t's." It is possible that, for particular terms *a* and *b*, both "*Oba*" and "*Oab*" may be jointly true (*e.g.*, if *a* = "German" and *b* = "blond"); but this joint truth depends on the accidental content of the particular terms, not on the logical structure of the propositions. Thus, "*Oba*" is not eligible for any conversion.

Equipollence, obversion, and contraposition

Since, in all of its lines, column "*Aba*" exhibits truth values exactly the opposite of those in column "*Oba*," another logical connection—traditionally called equipollence (near in sense to the modern term "equivalence")—can be identified, asserting that "*Aba*" is logically equipollent to "*Not-Oba*" (or "*NOba*") and vice versa. The same relation holds between "*Eba*" and "*Iba*" (or "*Iab*"). Using "≡" as a sign for equivalence, the logician can write "*Eba* ≡ *NIba*" and "*Eba* ≡ *NIab*." Furthermore, negated term variables can be introduced for terms like not-man or nonsmoker by employing a superposed bar, $\bar{a}$, $\bar{b}$, . . ., provided only that, just as terms defining the null (empty) class were earlier excluded, all terms defining the universal class shall be disqualified—in short, that, for every *a*, $\bar{a}$ is not empty. Thus every syllogistic proposition is equivalent to the proposition that results by changing its quality (affirmative to negative or vice versa) and at the same time negating the second (predicate) term: thus, "*Aba*" is equivalent to "*Eb*$\bar{a}$," and "*Iba*" to "*Ob*$\bar{a}$," a change traditionally called obversion. Another type of immediate inference, called contraposition, rests on the fact that some syllogistic propositions imply the proposition that results from the original proposition when both of its term variables are negated and their order is reversed. Only in the case of "*Aba*" and "*Oba*" is the resulting proposition logically equivalent to the original one. It is evident, for example, that "If every member is a smoker, then every nonsmoker is a nonmember"; and that "If some member is not a smoker, then some nonsmoker is not a nonmember"; *i.e.*, "Some nonsmoker is a member," or "There are members besides smokers."

After the foregoing definitions of equipollence and obversion, it is possible to define each of the operators *A*, *E*, *I*, and *O* by negation plus any one of the other three operators. "*Aba*" may be taken as basic, for example, and "*Eba*" defined as "*Ab*$\bar{a}$," "*Iba*" as "*NAb*$\bar{a}$," and "*Oba*" as "*NAba*." Corresponding definitions can be framed with other operators as basic. Instead of employing *A*, *E*, *I*, and *O*, for instance, modern logic uses for the same task only two operators, $(\forall x)$ and $(\exists x)$, the universal and existential quantifiers, which express only quantity ("all" and "some," respectively), leaving the quality, in the case of negative syllogistic propositions, to be expressed by obversion.

Summary of inference relations. Summing up (and using "⊃" for "implies" and "≡" for "is equivalent to"), the logical laws of Table 2 can be stated, each of which provides the basis for an immediate inference (from one syllogistic proposition to another).

Table 2: Summary of the Laws of Immediate Inference

law no.	law	law no.	law
From the square of opposition		**By conversion**	
2.1	Aba ⊃ NEba	2.11	Eba ≡ Eab
2.2	Eba ⊃ NAba	2.12	Iba ≡ Iab
2.3	NIba ⊃ Oba	2.13	Aba ⊃ Iab
2.4	NOba ⊃ Iba	2.14	Eba ⊃ Oab
2.5	Aba ⊃ Iba	**By obversion**	
2.6	Eba ⊃ Oba	2.15	Aba ≡ Eb$\bar{a}$
2.7	NIba ⊃ NAba	2.16	Iba ≡ Ob$\bar{a}$
2.8	NOba ⊃ NEba	**By contraposition**	
2.9	Aba ≡ NOba	2.17	Aba ≡ A$\bar{a}\bar{b}$
2.10	Eba ≡ NIba	2.18	Oba ≡ O$\bar{a}\bar{b}$

RULES AND LAWS OF THE SYLLOGISM

Terms, moods, and figures

Categorical syllogistic inferences. A categorical proposition contains just one subject term, one predicate term, and one syllogistic operator to join them. A categorical syllogism is an inference of one such proposition from two others. Each premise must have one term, called the middle term, in common with the other premise and one term in common with the conclusion. The term functioning as predicate of the conclusion is called the major term, and the one functioning as subject of the conclusion is called the minor term of the syllogism. The premise containing the major term, conventionally stated first, is called the major premise, and the other is called the minor premise. In the syllogistic example:

syllogism	proposition type
No man is perfect	*E*
Every logician is a man	*A*
Therefore, no logician is perfect	*E*

"perfect" is the major, "man" the middle, and "logician" the minor term.

The quantity and quality of the three propositions—expressed, as before, in terms of the proposition types *A*, *E*, *I*, and *O*—determine what are called the moods of the syllogism. The example just given thus belongs, as shown in the right-hand column, to the mood *EAE* (reading down). Since the three propositions may take four different forms each, there are $4^3 = 64$ different moods. Another way in which the forms of syllogisms differ is in their figure, which is the pattern in which its terms *S*, *M*, and *P* (subject, middle, and predicate) are arranged in the premises (see Table 3). This pattern is effectively determined by the function of the middle term, which may be subject in the major and predicate in the minor premise, as in the foregoing example (first figure); predicate in both premises (second figure); subject in both premises (third figure); and, finally, predicate in the major and subject in the minor premise (fourth figure).

Table 3: The Four Figures of the Syllogism

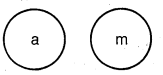
(1)	(2)	(3)	(4)
M-P	*P-M*	*M-P*	*P-M*
S-M	*S-M*	*M-S*	*M-S*
S-P	*S-P*	*S-P*	*S-P*

Analysis of the valid moods. Since there are 64 moods in four figures, there can be 256 forms of the categorical syllogism. It is one of the tasks of syllogistic to discover which of these forms are valid and to show that all others are invalid. Aristotle accepted 14 valid moods officially and five unoffically (those of the fourth figure). In the Middle Ages, these 19 moods were given Latin names, such as *Barbara*, *Camestres*, or *Datisi*, which indicated the moods through their vowels, names that were listed in mnemonic verses. Since five of these 19 syllogisms have universal conclusions, from which corresponding particular propositions can be inferred (by passing from "all" to "some" by laws 2.5 and 2.6), there will be five additional valid syllogisms—traditionally called *Barbari*, *Celaront* (first figure); *Cesaro*, *Camestrop* (second figure); and *Camenop* (fourth figure). Upon examination, it would seem that a syllogism in *Barbara* is self-evident; and Aristotle, in fact, accepted all first-figure syllogisms as self-evident and deduced all other syllogisms from them (see below *Reduction to the first figure*). But Euler diagrams (see Figure 2) can also be employed. Taking as an example *EIO*, in the second figure (*Festino*), written

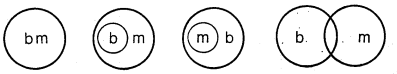
Identification of valid moods

Eam · *Ibm* ⊃ *Oba*,

(using · for "and"), the relation of *a* and *m* can be shown by the Euler diagram (ε):



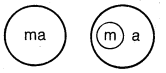
The second premise, "Some *b* is *m*," is satisfied by each of the diagrams (α) through (δ):



since, in each case, at least some of the "*b*'s" must be inside the extension of "*m*." Because, by the first premise, *a* is wholly excluded from *m*, however, at least some members of *b* will be outside of *a*; and this is just what "*Oba*," the conclusion, asserts. Similarly, it can be shown that *AEE*, in the third figure, is not valid; *i.e.*,

$$Ama \cdot Emb \supset Eba \text{ (invalid)}.$$

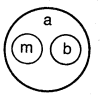
"*Ama*" is satisfied by the Euler diagrams (α) and (β),



and "*Emb*" by diagram (ε),



Clearly, the arrangement



is possible, in which case "*Aba*" would be true. Because *b* might equally well fall extensionally wholly outside of *a*, however, or *a* and *b* might merely overlap, all four propositional forms concerning *a* and *b* are compatible, and thus no conclusion follows.

Rules for testing validity

To shorten the task of testing the 256 possible syllogistic forms, three sets of general rules for the validity of categorical syllogisms have been formulated.

There are, first, the rules of distribution. A term is said to be distributed in a given proposition if that proposition implies all other propositions that differ from it only in having, in place of the original term, any other term whose extension is a part of that of the original term. Thus the term "student" is distributed in the propositions

Every student is a member;
No general is a student;
No student is a general;
Some boy is not a student;

inasmuch as each proposition remains true if one inserts, say, "law student" or "girl student" instead of "student." "Student" is not distributed, however, in the propositions

Every member is a student;
Some student is a boy;
Some boy is a student;
Some student is not a girl;

which may become false under a corresponding insertion for "student." In general, negative propositions distribute their predicate term, and universal propositions distribute their subject term. Therefore, in "*Eba*," both *b* and *a* are distributed, in "*Aba*" only *b*, and in "*Oba*" only *a*; "*Iba*," however, distributes neither of its terms.

There are two rules of distribution:

1. The middle term must be distributed at least once in the premises, and

2. No term may occur distributed in a valid conclusion unless distributed in the premises.

Thus "*Aam* · *Ibm*" has an undistributed middle term. In the syllogism

$$Ama \cdot Obm \supset Oba \text{ (invalid)},$$

however, the middle term is distributed; but *a*, distributed in the conclusion, is distributed in neither premise, a fault called illicit process.

The rules of quality are the following:

3. One premise, at least, must be affirmative, and

4. If one premise is negative, the conclusion must be negative; and a negative conclusion requires one negative premise.

Though rules 1–4 are sufficient to test any syllogism, three more rules facilitate the testing: one more rule of quality, viz.,

5. If both premises are affirmative, the conclusion must be affirmative,

and two rules of quantity:

6. At least one premise must be universal, and

7. If either premise is particular, the conclusion must be particular.

All syllogisms satisfying all of these rules are valid; and a syllogism violating at least one of them is invalid.

Analysis by law, case, and application

Underlying logical structures. Though inferences framed in the first figure have usually seemed particularly natural or evident and have thus been given prominence, the logical structures lying at the roots of all syllogistic figures are the same; the different figures only bring into focus their different aspects. The logical forms can be approached from various angles:

1. Given a general law that some class has (or lacks) a certain property, one may then establish that something has (or lacks) this property by showing that it is a member of that class (first figure);

2. Given the same general law, if one wishes to establish that something is not a member of that class, he may do so by showing that it lacks (or has) that property (second figure);

3. If it has been shown that something has (or lacks) a certain property and that it is a member of a certain class, one can then infer that no general law asserting that all members of this class lack (or have) that property can be valid (third figure).

If the different functions of the syllogistic propositions are described as "the statement of a law," "the establishment of a case," and "the application of the law to the case," the special nature of the figures may then be schematically summed up as in Table 4. In all three figures, the proposition of the application step is the only one capable of assuming all four propositional forms, *A*, *E*, *I*, and *O*.

Table 4: Functional Summary of the First Three Figures of the Syllogism

figure	law	case	application	case	law
first	affirmed	affirmed	affirmed		
second	affirmed		denied	denied	
third			denied	affirmed	denied

As for the fourth figure, although, like the first, it conforms to the law-case-application model, it does so with special emphasis not on the class to which the property is applied but on the property itself, of which the argument shows that it has (or lacks) application to something, viz., to all or some of the members of a certain class.

Reduction to the first figure. Aristotle, in *Analytica priora A* 5–7, reduced all other syllogisms (called by him "imperfect") to those of the first figure, the "perfect" (*i.e.*, self-evident) syllogism. He proved their validity, as would be said today, by deriving them as theorems from the axioms of his system—thus qualifying as the first scholar known to have developed an axiomatic system.

Direct reduction

Aristotle used two main methods of proof: direct and indirect reduction. The first has its logical basis in an *a fortiori* argument: if a syllogism *X* is valid, he argued, the logician may also infer the validity of every syllogism *Y* such that the premises of *Y* imply those of *X*, and the conclusion of *X* implies that of *Y*. The reduction, for example, of *Cesare* (second figure),

$$Eam \cdot Abm \supset Eba$$

to *Celarent* (first figure), $Ema \cdot Abm \supset Eba$, proceeds thus: "*Eam*" implies "*Ema*" (by 2.11); the second premises of *Cesare* and of *Celarent* are the same; therefore, the premises of *Cesare* imply those of *Celarent*. But *Celarent*'s conclusion, being identical, implies that of *Cesare*; therefore, *Celarent*'s validity implies that of *Cesare*. More complicated is the reduction of *Camestres* (second figure),

$$Aam \cdot Ebm \supset Eba$$

to *Celarent* (first figure). "$Aam \cdot Ebm$" implies (by 2.11) "$Aam \cdot Emb$." After changing the order of premises, one obtains "$Emb \cdot Aam$," from which, by *Celarent*, "*Eab*" can be inferred. But this implies (again by 2.11) the conclusion of the original *Camestres*. Thus, *Camestres* has been reduced to *Celarent*, the one having been logically derived from the other.

Indirect reduction

Baroco (second figure) and *Bocardo* (third figure) cannot be reduced by conversion because *AO* premises (or *OA* premises) imply only *IO* (or *OI*) premises by conversion, which do not even form a valid syllogism, still less a perfect one. Therefore, Aristotle fell back on another proof, indirect reduction (or *reductio ad impossibile*), based on another logical law: that if a syllogism is valid, the negation of its conclusion (posited for the sake of argument) together with either of its premises will imply, as a conclusion, the negation of the remaining premise; *i.e.*,

1. $(p \cdot q \supset r) \supset (p \cdot \bar{r} \supset \bar{q})$

and

2. $(p \cdot q \supset r) \supset (\bar{r} \cdot q \supset \bar{p})$.

To show that *Baroco* (second figure),

$$Aam \cdot Obm \supset Oba$$

is valid, its conclusion "*Oba*" is hypothetically negated (by 2.9) to "*Aba*." But "$Aam \cdot Aba$" implies by *Barbara* (first figure) "*Abm*," which is the contradictory of the original second premise. The hypothesis that *Baroco*'s premises may be true and its conclusion false for the same terms leads to a contradiction ("$Abm \cdot Obm$"); therefore, this hypothesis must be false; *i.e.*, *Baroco* is valid. The reduction of *Bocardo* (third figure) is strictly analogous.

Aristotle had already stated (*Analytica priora B* 11–13) that indirect reduction can be used not only for *Baroco* and *Bocardo* but for all moods of the second and third figures and that, even within the first figure, *Darii* and *Ferio* may be reduced via *Camestres* and *Festino* (second figure) to the first-figure mood *Celarent*, using one step of direct and one of indirect reduction in each case—which reduces the number of axioms, or forms commended only by their self-evidence, to just two: *Barbara* and *Celarent*. In fact, all of the 24 valid syllogisms may be arranged into groups of three, with each member derivable from the other two by indirect reduction.

Development from antilogisms

Inasmuch as the validity of a syllogism means that the truth of the premises is incompatible with the falsity of the conclusion, then the fact that, for example, $Ama \cdot Abm \supset Aba$ is a valid syllogism may be expressed in another way by saying that "*Ama*," "*Abm*," and "*Oba*" (the negation of the conclusion) cannot be jointly true, or, equivalently (using "$\sim$" to symbolize "not"), that $\sim(Ama \cdot Abm \cdot Oba)$. Thus, *AAO* is defined as a (basic) antilogism. Other antilogisms are *AAE*, *EAI*, *AIE*, and *EIA* and, for the fourth figure, *AAE* and *AEI*. Since in any conjunction the order of terms is logically indifferent, the number of antilogisms from which all 24 categorical syllogisms may be developed can be reduced to just three: *AAO*, *AAE*, and *AIE*.

Several other types of categorical syllogism have been described (see also RELATED ENTRIES under SYLLOGISTIC in the *Ready Reference and Index*); a sorites, for example, is a polysyllogism in the form of a chain (with intermediate conclusions suppressed), with the conclusion uniting the subject of the last link with the predicate of the first, thus: $Aba \cdot Acb \cdot Adc \ldots Eyx \supset Eya$. An enthymeme, as used in post-Aristotelian times, is a syllogism in which one of the premises (or even the conclusion) is suppressed, being thought too obvious to need expression. And, finally, in *Analytica priora B* 23, Aristotle introduced the inductive syllogism, which establishes a general rule by making an assuredly complete examination of all cases (one by one) and showing it to be true for every one.

Noncategorical syllogistic inferences. Besides categorical syllogisms, there are also noncategorical syllogisms of several types—hypothetical, disjunctive, and (in some treatments) copulative syllogisms. Since the fundamental difference between propositional logic, which treats unanalyzed entire propositions as its units, and predicate logic, which deals with terms in combination, was not appreciated by logicians before the rise of modern mathematical logic, there has been widespread confusion on this topic. The famous and influential *La Logique; ou l'art de penser* (1662; Eng. trans., *The Port-Royal Logic*, 1851), for example, failed to recognize the difference between implication between terms and implication between propositions and threw together arguments containing propositions such as "If the Chinese are Muslims, then they are infidels" and "If one of the elect perishes, then God is mistaken," the first being of the type expressed in modern logic by $\forall x[Fx \supset (Gx \supset Hx)]$, in which such expressions as Fx signify that "x has the property F," and the second being of the form "$p \supset q$" (in which p is called the antecedent and q the consequent). In 1887 the logician John Neville Keynes, long the registrar at the University of Cambridge, insisting on these important formal differences, suggested the name conditional for the first kind of proposition and hypothetical for the second. Noncategorical syllogistic has mainly and wisely been restricted to the important argument forms of propositional logic.

Hypothetical and disjunctive syllogisms

Hypothetical syllogisms are called pure if they consist of propositions all of which are of the form "$p \supset q$"; *e.g.*, (1) "If p then q, and if q then r; therefore if p then r." Mixed hypothetical syllogisms, however, which have one hypothetical and one categorical premise and a categorical conclusion, have two valid moods: (2) "If p then q, and p; therefore q" (called *modus ponens* from the Latin *ponere*, "to affirm"), and (3) "If p then q, but not-q; therefore not-p" (called *modus tollens*, from *tollere*, "to deny"). (For relevant fallacies, see LOGIC, APPLIED: *The critique of forms of reasoning: Formal fallacies.*)

Disjunctive syllogisms, which contain propositions built up with the "either . . . or" operator (symbolized $\vee$), have two important moods: (4) "Either p or q, but not-p; therefore q"—or: " . . . but not-q; therefore p" (called *modus tollendo ponens*); (5) "Either p or q, and p; therefore not-q"—or: ". . . and q; therefore not-p (called *modus ponendo tollens*). Mood (5) is valid, however, only if "either p or q" is taken in the sense of "at least one of p or q is true, but not both"; it is invalid, on the other hand, if "p or q" is taken in the nonexclusive sense, which does not rule out the joint truth of p and q.

A more complex form of noncategorical syllogism is the dilemma. It has two moods, one constructive, the other destructive: (6) "If q then p, and if r then p, and either q or r; therefore p"; (7) "If p then q, and if p then r, and either not-q or not-r; therefore, not-p." By an operation known as conditionalization, it is also possible—in what is essentially a reversal of the case in the dilemma—to draw hypothetical conclusions from categorical premises.

Complex and compound syllogisms

Syllogisms can also have as their terms not simple concepts like a, b, or c but terms built up from such simples by conjunction ("$a \cdot b$") or disjunction ("$a \vee b$"). Such terms, which are called complex, combine into complex propositions (example: "Every student or professor is eligible"). Compound propositions, on the other hand, are those constructed by combinations of other propositions by operators such as "If p then q," "Either p or q," or "Both p and q." Compound propositions may contain simple or complex propositions or both. "Either every student is efficient or popular, or all professors and schoolteachers are frustrated," for example, is a compound proposition formed from complex propositions.

The logical analysis of inferences from complex and compound propositions has been brought to a satisfactory level only in the 20th century by the recent development of the necessary formal methods and powerful calculi by modern mathematical logicians.

INTERPRETATION OF SYLLOGISTIC AS AN AXIOMATIC DEDUCTIVE SYSTEM

Aristotle's methods of reduction achieved something approaching an informal axiomatic system of syllogistic. At the turn of the 18th century, Gottfried Leibniz, a German Rationalist philosopher and mathematician, improved upon this method by showing how the laws of conversion and subalternation may be proved syllogistically using syllogisms of the second and third figures, and how these syllogisms, in turn, may be derived indirectly from the perfect moods. But neither of them stated rigorously the logical laws according to which these reductions and proofs were made. A formalized axiomatic system of syllogistic has, in fact, been made possible only by the methods of modern logic.

Auxiliary theory and base. For such an axiomatic representation of the syllogistic system, the following main elements are needed: As an auxiliary theory, the laws of the logic of propositions are employed (some of which were mentioned earlier). The logician may either accept as axioms (*i.e.*, without proof) the three to six axioms of the propositional calculus as variously presented by the German mathematician Gottlob Frege, by Bertrand Russell, or by Jan Łukasiewicz, three pioneers in fields of modern logic, and deduce from them all of the propositional laws that he may need; or, alternatively, he may specify certain laws (usually a dozen or more) after discovering which he will actually use.

The system must, moreover, have a base from which to derive its theorems, a list that should specify (1) primitive expressions (letters or other symbols); (2) rules of formation, according to which the logician may build up other expressions from the primitive ones; (3) axioms of the system; *i.e.*, propositions that are accepted without proof; (4) definitions, which enable the logician to substitute one form of expression for another (optional); and (5) rules of inference, which state all of the steps he may take in deriving further expressions from axioms or from other expressions already derived.

Elements of the base

For the axiomatization of syllogistic, the following list (from Łukasiewicz) contains a sufficient base:

A. Primitive expressions:
1. Three term variables, *a*, *b*, and *m*.
2. Primitive syllogistic operators *A* and *I*, each of which may be joined to any two of the term variables.

B. Rules of formation (for well-formed formulas):
1. An expression consisting of *A* or *I* followed by any two term-variables *a*, *b*, or *m* is a well-formed formula or a syllogistic expression, designated by alpha (α), beta (β), and so on.
2. If α is a syllogistic expression, $N\alpha$ (read "Not-α") is also a syllogistic expression.
3. If α and β are syllogistic expressions, $\alpha \cdot \beta$ and $\alpha \supset \beta$ are also syllogistic expressions.

C. Axioms of the system:
1. *Aaa*
2. *Iaa*
3. *Ama* · *Abm* ⊃ *Aba (Barbara)*
4. *Ama* · *Imb* ⊃ *Iba (Datisi)*

D. Definitions:
1. $Eba =_{df} NIba$
2. $Oba =_{df} NAba$

E. Rules of inference (for deriving expressions from others):
1. If α is an axiom or a derived expression (theorem) of the system, so is any expression that is the result of replacing one term-variable of α by another term-variable in all places of its occurrence in α (rule of substitution for term variables).
2. If α is a law of propositional logic, the result of replacing any propositional variable $p, q, r \ldots$ in α by a syllogistic expression may be asserted as a theorem provided that the same expression is inserted in all places of its occurrence (rule of substitution for propositional variables).
3. In any derived expression, the expressions that are defined by rules D1 and D2 may be substituted for any occurrence of the expressions that define them (rule of substitution for definitions). This rule does not require, however, that the replacement cover all occurrences.
4. If α and $\alpha \supset \beta$ are derived expressions, β is a theorem (rule of detachment). This rule is an application of *modus ponendo ponens* (see above *Noncategorical syllogistic inferences*).

The rules of formation generate all of the syllogistic expressions, without regard to their validity or invalidity. To select from these expressions those that are valid (theorems of the system), the foregoing axioms and rules of inference are sufficient. To show that all 24 valid syllogisms may be derived formally, it is sufficient to show how *Barbara*, *Celarent*, the rules of contradictory opposition, and the rules of conversion can be derived in the system, since this is the basis used by Aristotle.

Properties of syllogistic as a system. It is interesting to note what deductive power is possessed by each of the axioms in the system; *i.e.*, which theorems would be lost by dispensing with a given axiom and what other expression might serve as a substitute for it. Axiom C1 ("*Aaa*") is needed to derive the theorem "*Iba* ⊃ *Iab*" and the other rules of conversion. Since the derivations of exactly half of the 24 moods depend on conversion, only the rules of opposition and 12 of the syllogisms would remain as theorems. In the place of C1, one might add C5, "*Iba* ⊃ *Iab*," as an axiom to restore the full deductive power of the set of axioms. Without axiom C2 ("*Iaa*"), which guarantees that all of the terms are nonempty (*i.e.*, have application to something), the logician is left without the rules of subalternation and the other rules of opposition (except contradiction). Therefore, with axiom C2 would go all of the nine moods of the syllogism that depend on subalternation for their proof. Addition of the theorem C6 *Aba* ⊃ *Iba* instead of C2, however, would restore full deductive power to the system. Leaving out axiom C3 (*Barbara*) would cost only two syllogisms besides *Barbara*, viz., *Baroco* and *Bocardo*, each of which could thus serve as a substitute for C3. Since axiom C4 (*Datisi*) is needed in the proofs of C5 and C6, the result of leaving it out would be very damaging; it would equal the results of dropping C1 and C2 together, so that $9 + 12 = 21$ syllogisms would be blocked, leaving only *Barbara, Baroco*, and *Bocardo*. Any one of the syllogisms *Ferio, Cesare, Dimaris, Fresison*, or *Camenes*, however, could, taken as an axiom, repair the damage. If all four syllogistic operators *A*, *E*, *I*, and *O* were admitted instead of just two (*A* and *I*), the logician could also take *Barbara* and *Celarent*, as Aristotle did, as axioms together with C1 and C2. If, instead of C1 and C2, he would take C5 and C6 as axioms, the resulting set of axioms would be the base corresponding best to Aristotle's approach.

Deductive power of axioms

Since derivations in the system of axiomatic syllogistic depend entirely on rules for the manipulation of concatenations of signs, and not on any intuitive insights or knowledge of the meaning of these signs, it may be regarded as a formal deductive system in the modern technical sense. Thus, the formal properties of such a system can be investigated without giving the symbols any particular interpretation. One such formal property, internal consistency, is attributed to a system if it is impossible to derive a contradiction in it. Another formal question, the independence of its axioms, is at issue when the logician asks whether one of the axioms is redundant because it can be proved from the other axioms. Finally, the logician may ask whether the system is complete, in the sense that it is always possible, for any well-formed formula α, to devise a proof either that it is a theorem or that its negation is a theorem.

Consistency of a system

If a system is inconsistent, it is, for all purposes of deduction, without value, since any proposition whatsoever must then be derivable in it. That this is so is easily shown: A contradiction must be of the form "$p \cdot \bar{p}$." If "$p \cdot \bar{p}$" is a theorem, then according to the propositional law "$p \cdot \bar{p} \supset p$," one infers, by the rule of detachment, p. With p, it follows also that "$p \vee q$"—because "$p \supset (p \vee q)$" is again a law of propositional logic. If "$p \vee q$" is a theorem, however, it is possible to derive $\bar{p}$ again from "$p \cdot \bar{p}$" and to have, by an application of *modus tollendo ponens*, also q. This means that in an inconsistent system any proposition that one likes is derivable as a theorem, and the system is therefore deductively useless.

A system is sufficiently shown to be consistent if its axioms are consistent; for the rules of inference guarantee that no false proposition will be derivable from true ones. Therefore, in case of inconsistency anywhere, the

axioms themselves already must have been inconsistent. Hence, a proof of the consistency of a formal system must cite an interpretation for the axioms that makes them all true propositions. For the system of syllogistic, various such interpretations, mostly from arithmetic or propositional logic, have been given.

Independence and completeness

In a similar way—and in a manner resembling Aristotle's method of invalidating moods by counterexamples—it may be shown that each of the axioms is independent of the others (*i.e.*, is not derivable from them) by producing an interpretation that makes the axiom in question false and the other axioms true. It is then obvious that this axiom cannot logically follow from the others, since its falsity would be incompatible with their joint truth. Simple proofs of independence have been devised —with interpretations taken from the propositional calculus and arithmetic—for all of the axioms except *Barbara*, for which the proof is complicated. These proofs have shown that none of the axioms can be regarded as being the one and only principle of syllogistic.

The preceding system is strong enough to derive all valid theorems of syllogistic from the stated axioms. For completeness, however, it is further required that a formal proof be provided for the negation of any invalid syllogistic expression. But, since this cannot be done, the system is not complete. The system may be strengthened by adding two axioms, viz., the negations of "*AAI*" and "*EEI*" in the second figure, and certain rules of rejection, corresponding to the rules of substitution and the rule of detachment for inference. Though with these additions to the base, it is possible to reject all invalid syllogistic moods, it is still not possible to reject such invalid syllogistic expressions as those having more than three terms and some of those having two terms. By an ingenious method, devised by the Polish logician L. Słupecki, a stronger rule of rejection has been formulated that makes the system complete in the sense defined.

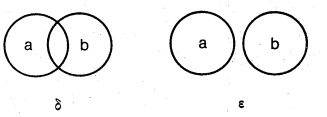
EXTENSIONS OF SYLLOGISTIC

Singular and negative terms. Expressions like "Socrates" or "the founder of psychoanalysis," called singular terms because they refer to but one unique particular, yield propositions that are formally different from general ones. All syllogistic laws, however, are formulated for general propositions only. Though Aristotle seems to have thought that singular propositions were a special kind of universal proposition, the rules of opposition fail for them and they cannot be converted or changed by contraposition, because proper names can only take the subject place in a proposition. "Socrates is mortal" can be replaced, of course, by "Everything identical with Socrates is mortal," which defines a class of only one member; but this solution does not really introduce singular terms into syllogistic but keeps them out while taking care of all statements that contain them. Moreover, if proper names can only replace subject terms, the only allowable syllogisms incorporating singular premises are those in which singular terms occur twice in the subject place. One may therefore substitute singular terms instead of b in the first and second figures and in place of m in the third figure. These requirements limit the syllogisms that can have singular terms to *Barbara, Celarent, Cesare, Camestres, Darapti,* and *Felapton.*

Confinement to subject position

Aristotelian syllogistic avoids negative terms like $\bar{a}$, $\bar{b}$, and $\bar{m}$. The modern way of expressing syllogistic propositions, however, is to treat *E*- and *O*-propositions as *A*- and *I*-propositions with a negated predicate term. Because the modern functional calculus allows the negation of subject terms as well, the foregoing syllogistic can only cover a fragment of the class of possible expressions of predicate logic. The question arises, therefore, whether the axiom system for syllogistic might be extended to cover propositions with negated term-variables—which would make the so-called exclusive propositions, such as "Only students are members," and exceptive propositions, such as "All but (except) heroes are frightened," also amenable to syllogistic treatment. The exclusive proposition can be translated into "Every nonstudent is a nonmember" and represented as "$A\bar{b}\bar{a}$," which is equivalent to "Aab" by contraposition. Moreover, the exceptive proposition, being translatable into the conjunctive proposition "Every nonhero is frightened and no hero is frightened," is of the form "$A\bar{b}a \cdot Ab\bar{a}$," which amounts to saying that a and b are disjunctive terms in the mutually exclusive sense. Because the terms may be combined in four ways, ba, $b\bar{a}$, $\bar{b}a$ and $\bar{b}\bar{a}$, each of which can take the four operators *A*, *E*, *I*, and *O*, there are 16 logically different forms of syllogistic propositions and 16 further converse forms; *i.e.*, 32 forms of propositions in this system. If one now disqualifies empty terms, he must also exclude universal terms, which apply to everything (like "self-identical" or "mentionable") since, if a is a universal term, $\bar{a}$ must be empty. A problem now arises in the interpretation of the Euler diagrams (δ) and (ε):

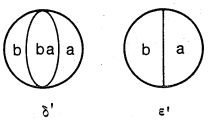
Exclusion of universal terms



δ ε

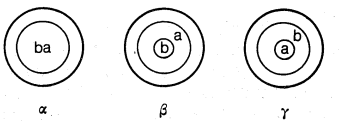
There can be objects that are $\bar{b}$ and objects that are $\bar{a}$; but it is possible that there are no objects that are both $\bar{b}$ and $\bar{a}$, because b and a might be complementary classes that between them exhaust the universe. There may then be cases of "$Ib\bar{a}$" and "$I\bar{b}a$," but not of "$I\bar{b}\bar{a}$," or there may be cases in which no $\bar{b}$ is an $\bar{a}$ (*i.e.*, "$E\bar{b}\bar{a}$," which is equivalent to "$A\bar{b}a$"; *i.e.*, to "Everything that is not b is a," there being nothing else in the universe but things that are a or b).

Two more diagrams must therefore be introduced, presenting the special subcases of (δ) and (ε):



δ' ε'

in which the circle represents the total universe of discourse. To bring the rest of the diagrams up to this level of abstraction, such a circle should then be added also to the diagrams (α)–(γ):



α β γ

By introducing negative terms, the rules for the validity of syllogisms are somewhat altered. Since "Aba" is equivalent to "$Eb\bar{a}$," the valid syllogism "$Em\bar{a} \cdot Eb\bar{m} \supset Eb\bar{a}$" arises—merely another way of expressing *Barbara.* Moreover, "$Em\bar{a} \cdot Eb\bar{m} \supset Aba$." Since these syllogistic forms seem to violate the rules of quality, it should be stipulated that those rules should apply only to a syllogistic without negative terms. An axiom system called $CS(n)$ has been devised, which differs from the system described above by admitting negative terms, by dropping *Barbara* from the axioms, by using as definitions "$Eba =_{df} Ab\bar{a}$" and "$Iba =_{df} NAb\bar{a}$," and by having a special rule of inference for double negation.

Empty terms and existential import. The fact that no empty terms may replace any of the term-variables is a deficiency of traditional syllogistic that calls for certain precautions. In the early 1970s, for example, no one knew whether gravitation waves existed; but, from "Every gravitation wave is a physical phenomenon," there is a danger that, without this ban on empty terms, the proposition "Some gravitation wave is a physical phenomenon" might be inferred. This statement, however, implies by conversion that "Some physical phenomenon is a gravitation wave," which in turn would most naturally be interpreted as asserting the existence of gravitation waves.

Modifications owing to emptiness

In order to admit terms regardless of their emptiness or nonemptiness, certain revisions of the earlier laws are required. It is important to specify, for example, which syllogistic expressions are to be false if containing empty terms. Three different specifications have been discussed by logicians: (1) that *A*-, *I*-, and *O*-propositions, but not

E-propositions, presuppose the existence of both terms for their truth; (2) that *A*- and *I*-propositions, but not *E*- and *O*-propositions, presuppose nonemptiness of their terms; (3) that *I*- and *O*-propositions presuppose nonemptiness, whereas *A*- and *E*-propositions do not.

In cases (1) and (2), all of the laws $CS(n)$, with the exception of inversion and contraposition ("$Aba \supset I\bar{b}\bar{a}$" and "$Aba \supset A\bar{a}\bar{b}$") would remain valid. Since these theorems are not used by $CS(n)$ for the derivation of syllogisms, however, all syllogistic moods remain. Specification (3) has much more to commend it. *A*- and *E*-propositions may be defined as negations of the corresponding *O*- and *I*-propositions. But "*Iba*" is most naturally interpreted as asserting the existence of objects that are both *b* and *a* and must therefore be false if *a* or *b* is empty. The same applies, correspondingly, to "*Oba*" in relation to *b* and $\bar{a}$. Since "*Aba*" and "*Eba*" are negations of "*Oba*" and "*Iba*," respectively, they must be true if "*Iba*" and "*Oba*" happen to be false through lack of reference of their terms. In this way, the laws of contradictory opposition remain valid. But subalternation is invalid: "*Aba*" may be true merely because "*Oba*" fails for lack of existential reference of *b*; but in this case "*Iba*" is also false. Contrary opposition and subcontrary opposition now vanish from the logical square: "*Aba*" and "*Eba*" may both be true and "*Oba*" and "*Iba*" may both be false in case there is a lack of reference for *b*. The most important law that is sacrificed, however, is "$Aba \supset Iba$," since the derivation of all moods having universal premises and particular conclusions depend on this law. The effect is the same as that of dropping the axiom "*Iaa*" from the system described earlier: *Barbari, Camestrop, Felapton, Celaront, Cesaro, Darapti, Bramantip, Fesapo,* and *Camenop* are then lost and only 15 moods remain. Since specification (3) for the existential import of *A*-, *E*-, *I*-, and *O*-propositions is equivalent to those adopted for general and particular propositions in the modern predicate calculus, the 15 remaining moods are the theorems of the syllogistic when interpreted in the predicate calculus. This is not to say, however, that Aristotle's syllogistic is incorrect, but only that he and his followers never formulated the existence postulates for their syllogistic terms. If these were made explicit and added to the premises of all syllogistic laws, they would all again become laws of predicate logic also.

Boolean algebra and Venn diagrams

When negative and empty terms are admitted, syllogistic becomes a special case of Boolean algebra, which can be defined as any system incorporating the concepts of a universal class, of a null class, and of the complement of a class, and those of class union and class intersection. The null class is denoted by 0, the universal class by 1; and if *a* is a class, $\bar{a}$ is its complement class. Emptiness is expressed by saying that the class is identical with the null class or that "$a = 0$"; universality, expressing the fact that the term *a* applies to everything, is represented by "$a = 1$." Clearly, if a term is empty, its complement class is identical with the universal class. The class that contains the common members of the terms *a* and *b* is denoted by $a \cap b$ and is called the intersection of *a* and *b*. The union of classes *a* and *b*, expressed by $a \cup b$, is the class of all of those things that belong either to *a* or to *b*. If two classes have the same objects as members—that is, if *a* and *b* are extensionally identical—then this fact is expressed by "$a = b$." Nonidentity of classes is expressed by "$a \neq b$."

In this language, the syllogistic propositions may now be expressed in the following way:

$$\begin{aligned} Aba &= {}_{df}\ b \cap \bar{a} = 0 \\ Eba &= {}_{df}\ b \cap a = 0 \\ Iba &= {}_{df}\ b \cap a \neq 0 \\ Oba &= {}_{df}\ b \cap \bar{a} \neq 0. \end{aligned}$$

Barbara, in this notation, is then:

$$(m \cap \bar{a} = 0 \cdot b \cap \bar{m} = 0) \supset b \cap \bar{a} = 0$$

and *Datisi* becomes:

$$(m \cap \bar{a} = 0 \cdot m \cap b \neq 0) \supset b \cap a \neq 0.$$

To test such Boolean syllogisms for validity, the Venn diagrams, so-called after the British logician John Venn, who devised them in 1880, provide an ingenious method. He represents the extension of the three terms by three interlocking circles as shown in Figure 3. A region that is said to be empty in a premise is shaded out; and a region said to be not empty is marked by a cross inside the

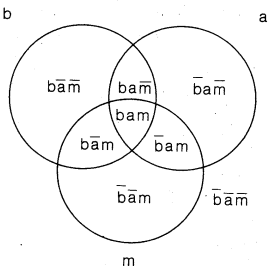


Figure 3: Basis of the Venn diagrams.

region. The Venn diagram for *Barbara* is shown in Figure 4. Its first premise states that the intersection of *m* and $\bar{a}$ is empty; therefore, one shades out all of *m* that is outside of *a*. The second premise states that the intersection of *b* and $\bar{m}$ is empty; therefore, one shades out all of *b* that is outside of *m*. The conclusion, finally, states that the intersection of *b* and $\bar{a}$ is empty. If, accordingly, the logician wants to shade out the region of *b* that is outside of *a*, he finds that this has been done already by the stipulations of the premises: the conclusion states nothing that has not been stated by the premises taken together; therefore, *Barbara* is valid.

Venn proofs of *Barbara* and *Datisi*

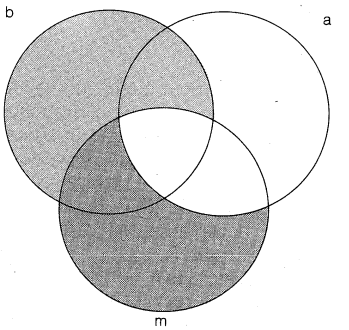


Figure 4: Venn diagram for *Barbara*.

The proof of *Datisi* proceeds as follows (see Figure 5). The first premise is illustrated by shading out the region of *m* that is outside of *a*. The second premise requires that a cross be placed in the region where *m* and *b* overlap. Since part of this region has been shaded out already, the cross can be put only into the part still left white; *i.e.*, the part where *m* and *a* and *b* overlap. The conclusion calls for a cross to be put into the overlapping region of *a* and *b*. The logician finds, however, that there is a cross already in this part; therefore *Datisi* is valid.

Modal and other special operators. Aristotle devoted the greater part of his *Analytica priora* to the analysis of modal syllogisms; *i.e.*, syllogisms complicated by the addition of such modal operators as "possible" and "necessary," as in the statements "It is possible that some of the soldiers who retreated are not cowards" and "Every cat is necessarily a mammal." Aristotle's treatment of modal syllogistic, however, was unsatisfactory: among other de-

Problem of a syllogistic of possibility and necessity

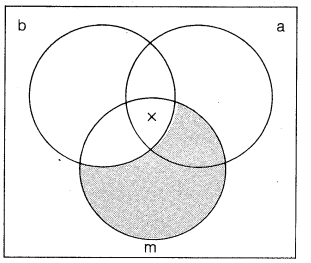


Figure 5: Venn diagram for *Datisi*.

fects, it was confused by his failure to decide whether a modal operator should determine the whole proposition (first example above) or only its predicate (second example). This problem reflects the dilemma that, if modal operators modify predicates only, then no special *modal* syllogistic is required, because the propositions have a normal structure (but with peculiar predicates); but if they modify whole propositions, then again no modal *syllogistic* is needed, because the special rules for the modal operators are independent of the intrinsic character of the statements. Thus modal syllogistic would best be developed in the propositional logic and not in the predicate logic of ordinary syllogistic, and it would then be merely a special case of general modal logic. Aristotle's treatment of it under syllogistic must then be attributed to his overrating of this new method rather than to the intrinsic aptness of the allocation.

Over the past decades new fields of logic have been drawn up under such names as the logic of assertion, the logic of epistemic propositions, deontic logic, and tense logic, to mention only a few. One may thus ask whether there is also a syllogistic that employs the special operators of these branches instead of the copula and the operators *A*, *E*, *I*, and *O*. In a role parallel to that of the modal functors, deontic operators might now be introduced such as "it is obligatory that," "it is forbidden that," or "it is allowed that"; the copulae might be varied according to tenses like past, present, and future; or epistemic functors like "____ is known to be a ____" might be employed. Here again, however, it is likely that in due time these new fields will be found subject to the same criticisms just raised against a modal syllogistic and that they might best be discussed in the field of propositional logic.

PRESENT STATUS OF TRADITIONAL LOGIC

Since the rise of modern mathematical logic alongside traditional logic, the relation between these two branches has sometimes been strained. Admirers of mathematical logic have asserted (1) that traditional logic is without useful application in the sciences, (2) that it covers only a very small fragment of logic, and (3) that it contains fundamental errors; *e.g.*, the acceptance of inferences from general to particular propositions.

The admirers of traditional logic have sometimes replied (4) that mathematical logic is merely a branch of mathematics, not of philosophy, and (5) that the results of mathematical logic cannot be compared in philosophical significance with the classical syllogistic theory.

Rejoinders to critical points

Point (4) is merely a matter of definition and can be neglected; point (5), however, seems to be confuted by the profound influence that modern logic has already exerted on the course of philosophy. Point (1) must be conceded: syllogistic theory was important for science at a time when science was endeavouring to bring particulars under general concepts and to establish a system of classification. In modern science, however, there is not much scope for such activities. Syllogistic arguments are used, however (in enthymeme form), when scientific results are applied, as in the application of law or in medical therapy. Point (2) is true: syllogistic is, indeed, a very small part of logic; but it is a very fundamental and important one. Almost all new ideas in the history of logic have been developed in connection with some special aspects of syllogistic theory. Point (3) is simply false: traditional logic is as correct as modern formal logic. But it is not formalized and, thus, not all of its presuppositions have been made explicit.

To sum up, syllogistic is a special part of logical theory. Its systematic importance has been overrated, however, by those who have tried to bring all of logic into the compass of Aristotelian logic. Because syllogistic is the first deductive system ever devised and was constructed by Aristotle with so high a level of rigour and logical purity, it has proved to be very fruitful as a starting point for generations of later logicians. Modern logic is not in opposition to traditional logic; it is a continuation of it—with better instruments and methods and in a much larger field.

BIBLIOGRAPHY. The standard edition of Aristotle's texts on syllogistic is: *Aristotle's Prior and Posterior Analytics*, with introduction and commentary by W.D. ROSS (1949). For the interpretation of Aristotle's syllogistic, see the chapter "La Syllogistique," in L. COUTURAT, *La Logique de Leibniz* (1901, reprinted 1961); J. LUKASIEWICZ, *Aristotle's Syllogistic from the Standpoint of Modern Formal Logic* (1951; 2nd ed., with a chapter on modal syllogistic, 1957); G. PATZIG, *Die Aristotelische Syllogistik* (1959; 3rd ed. 1969; Eng. trans., *Aristotle's Theory of the Syllogism*, 1968); and STORRS MCCALL, *Aristotle's Modal Syllogisms* (1963).

Traditional logic: The best textbook is J.N. KEYNES, *Studies and Exercises in Formal Logic* (1884; 4th ed., 1906). Remarkable for discussion of traditional theories in the conceptual framework of mathematical logic is A.N. PRIOR, *Formal Logic*, 2nd ed. (1962), with bibliography. The best history of logic is W. and M. KNEALE, *The Development of Logic* (1962), with bibliography.

Modern developments, axiomatization, and extensions of syllogistic: I.M. BOCHENSKI, "On the Categorical Syllogism," *Dominican Studies*, 1:35–57 (1948); a development of an axiomatic system for syllogistic with negative terms is given in I. THOMAS, "*CS*(*n*): An Extension of *CS*," *Dominican Studies*, 2:145–160 (1949); C.A. MEREDITH, "The Figures and Moods of the *n*-term Aristotelian Syllogism," *Dominican Studies*, 6:42–47 (1953); A. MENNE, *Logik und Existenz* (1954); P. LORENZEN, "Über Syllogismen als Relationenmultiplikationen," *Arch. math. Logik GrundlForsch.*, 3:112–116 (1957); OTTO BIRD, *Syllogistic and its Extensions* (1964), the best introduction to the subject, with bibliography.

(Gü.P.)

Sylvester II, Pope

Sylvester II, pope from 999 to 1003, was renowned, as Gerbert, for his scholarly achievements, his advances in educational methods, his shrewd political judgments, and the exceptional ability that lifted him from his peasant origin to two archbishoprics—Reims and Ravenna—and supreme headship of the church.

Early life and career

Gerbert was born of humble parentage about 945 near Aurillac in the ancient French province of Auvergne. He was trained at Saint-Gerald in grammar, arithmetic, and music under the tutelage of Raymond Lavaur, who later became abbot of Aurillac, where Gerbert probably became a monk.

Gerbert was taken to Spain in 967 by Count Borrell of Barcelona and remained there three years. At the monastery of Santa María de Ripoll, which was noted for its fine library, he studied the quadrivium (the higher division of the liberal arts, which includes music, arithmetic, geometry, and astronomy) under Bishop Atto of Vich.

In 970 Gerbert accompanied Borrell to Rome, where his mathematical knowledge delighted Pope John XIII, who presented him to the Holy Roman emperor Otto I. Gerbert was then entrusted with his first royal pupil, Otto II. Lively Socratic discussions, however, convinced Gerbert that he lacked training in logic, and, when a skilled logician, Archdeacon Gerann of Reims, visited Rome, Gerbert gained permission to study in Reims under Gerann's tutelage.

Logic and dialectic so captivated Gerbert that, after being invited by Adalbero (who later ordained him) to teach at the cathedral school, he became famous for reorganizing logical and dialectical studies. Gradually, he eliminated simpler logical works and added monographs on logic by Boethius, an early-6th-century Roman philosopher.

Gerbert's fame aroused the jealousy of Otric, master of the cathedral school at Magdeburg in Saxony (presently in the German Democratic Republic), who denounced Gerbert to Emperor Otto II. In December 980 Otto provoked a debate in Ravenna between Gerbert and Otric on the subject of classifying knowledge. The vehement argument that resulted was terminated only when the Emperor intervened.

Otto subsequently gave Gerbert the headship of the wealthy monastery of St. Columban of Bobbio (southwest of Piacenza, Italy), and Benedict VII consecrated him abbot of the monastery. Gerbert, however, was inexperienced in administration and unfamiliar with the land law in the region of Lombardy. As a result he encountered violent hostility from monks, clerics, and nobles,

Opposition to Gerbert's monastic rule

who possessed much of Bobbio's extensive property through long-term civil-law land leases and refused to make contractual payments or supply troops demanded by Otto. Some even stole the abbey's agricultural products. Within Otto's palace in Pavia (west of Piacenza), nobles plotted against Gerbert and even against Otto himself.

Rebellion followed Otto's death (December 983), forcing Gerbert to flee from Bobbio and hasten to Reims. He travelled and wrote urgent letters to arouse Germans, Belgians, and Lorrainers against Duke Henry of Bavaria, who was plotting to become king by seizing the three-year-old King Otto III. Gerbert resumed teaching at Reims but he really longed to be summoned to the imperial court.

After the deaths of the last Carolingian kings (Lothair in 986 and Louis V in 987), Adalbero and Gerbert influenced the nobles to elect Duke Hugh Capet king against the claims of the remaining Carolingian, Duke Charles of Lorraine. As Hugh's secretary and adviser, Gerbert became involved in Hugh's resistance to Charles's attempt to dethrone him. Before Adalbero died, in January 989, he indicated Gerbert as his successor, but Hugh unwisely chose Arnulf, an illegitimate son of King Lothair. In September Arnulf betrayed Reims to Charles, his uncle, who forced Gerbert to remain in the city. After eight months Gerbert managed to escape to Hugh's court, where he was falsely accused of having been the leader in "enormous crimes" of collaboration with Duke Charles.

In June 990 the bishops and Hugh appealed to Pope John XV to condemn and depose Arnulf and approve their consecration of a new archbishop. After Hugh captured Charles and Arnulf in March 991, a council at Saint-Basle de Verzy, near Reims, degraded Arnulf. Gerbert was elected archbishop and soon became embroiled in controversies with the papacy. He administered his ecclesiastical duties capably, but the clergy's confidence in the legality of his election decreased. When a newly appointed bishop refused to be consecrated by him and travelled to distant Rome for consecration by the Pope, Gerbert hastened to Rome to defend the legitimacy of his election. In the meantime, however, Pope John had died and Gerbert was heard by his successor, Gregory V, who rejected Gerbert's plea.

Returning north through Pavia, he met Otto III, then 16 and recently crowned emperor. Gerbert's witty intelligence charmed Otto, who immediately made use of his talents for writing letters. They parted reluctantly after August 5, 996. In January 997 a council declared that all bishops involved in Arnulf's deposition were removed from office. By April Gerbert, no longer tolerating the clergy's acceptance of these decrees, left Reims for Otto's court in Aachen (in modern West Germany), never to return to France.

He became a teacher again and a musician in Otto's chapel; he also travelled with Otto to Magdeburg, where he displayed his inventiveness by constructing a timepiece (*horologium*). In October he followed Otto to Italy, where Otto had him appointed (*c.* April 998) archbishop of Ravenna, an important city for Otto's control over Italy. Gerbert instigated a new agricultural policy in Italy by having Bobbio's frustrating long-term leases quashed and restricting future leases to the lifetime of the grantor abbot or bishop; this was, perhaps, the most important legislation enacted for Italy during the reign of Otto.

Elected pope

Within a year after Gerbert took over the see of Ravenna, Gregory V died, and Otto was able to choose Gerbert as Gregory's successor. He was consecrated—the first Frenchman to be elevated to the throne of Peter—on April 9, 999, and took the name Sylvester II, declaring his cooperation with Otto's ideal of a renewed Christian Roman Empire.

Their unanimity is shown by numerous documents issued by papal and imperial chanceries that use similar terms. Each appeared as intercessor for a petitioner in documents of the other and also participated in councils of the other. The two men disagreed on only one important political decision, the crowning of Duke Bolesław as king of Poland. Sylvester permitted establishment of Gniezno as an archbishopric independent of the German hierarchy but refused Otto III's desire to elevate Bolesław to kingship. Sylvester did agree, however, to create Stephen king of Hungary and to erect an independent archbishopric there. Otto's sudden death, on January 23, 1002, terminated their close cooperation.

Leonard von Matt—EB Inc.



Sylvester II, detail from the ivory vessel used at the consecration of Otto III, 996. In the Domschatzkammer des Aachener Domes, Aachen, West Germany.

To extend papal influence, Sylvester communicated with Vladimir I, the grand prince and first Christian ruler of Russia; demanded that King Olaf I Tryggvason of Norway, who made Christianity the official religion of the country, abandon the use of runic writing; sent ambassadors to Dalmatia (a coastal district of modern Yugoslavia); reprimanded the Doge of Venice and the Patriarch of Grado (on the northeastern coast of Italy) for the Venetian clergy's loose morals; held semi-annual general councils; and restored Arnulf to the Reims archbishopric, ending the divisive Reims controversy. He granted Poland its first archbishop (in 1000); and he denounced simony (the buying and selling of ecclesiastical offices) and nepotism and demanded celibacy. Bishops and nobles often exerted authority over wealthy abbeys; the abbeys thus petitioned Sylvester to free them from all exterior control except that of the papacy. Included in their request was the important right to elect their own abbots. Sylvester's acquiescence reduced the power of the bishops and increased that of the pope. To strengthen the papacy still further, Sylvester began its feudalization by enfeoffing Count Daiferio of Terracina in return for military service.

Sylvester died on May 12, 1003, and immediately legends about his great learning surfaced, revealing the impression he made on contemporaries. Some attributed his learning to magical arts learned in Spain, some to the devil's coaching, some to an artificial head that answered his questions.

Assessment

Gerbert was a great scholar, and he was also a good teacher and devised several practical means for learning: a chart for learning rhetoric; his writing about the abacus, which became the basic work on the subject and included

the use of Hindu-Arabic numerals, which he had learned in Spain; celestial globes; a hemisphere for learning the imaginary celestial circles; and auxiliary spheres, one for identifying constellations and another with planetary orbits. He acquired an astrolabe and wrote about its uses. He prepared a work on geometry that attempted to fill the gap caused by the existence of a fragmentary Euclid and that included writings by the Roman surveyors. He had an extraordinary knowledge of music and seems to have constructed several organs and also a monochord for studying music theory. His philosophical tract *De rationali et de ratione uti* ("Concerning the Rational and the Use of Reason") emphasized problems of definition and classification of knowledge.

Gerbert was an avid collector of manuscripts, which he kept locked in four or five chests; only three of these manuscripts have been identified. He also kept copies of his letters, which were written in good Latin; his original collection has disappeared, but two copies have survived. Only one contemporary biography exists: a work written by Richer of Saint-Rémy, who believed that Gerbert was sent to Reims "by the Divinity Itself." (This account was discovered only in modern times, in the cathedral library in Bamberg, Germany.) Thus for a long time even learned men relied on legends for knowledge of Gerbert.

Eventually, however, scholars discovered and published critical editions of his writings that dispelled many of the fables that had accumulated around him.

Gerbert remains the dominating figure of the late 10th century for modern scholars, who are concerned especially with more precise dating of his letters and a more detailed understanding of his political roles, his teaching (especially in logic, dialectic, mathematics, and astronomy), his transmission of Arabic learning, his relations with Otto III, his papacy, and his influence on later ages.

BIBLIOGRAPHY. RICHER, *Histoire de France (888–995)*, ed. and trans. into French by ROBERT LATOUCHE (1937), by Gerbert's pupil, is basic and accurate but with omissions; it ends in 998. Editions of Sylvester's works include ALEXANDRE OLLERIS, *Oeuvres de Gerbert* (1867), with an extensive biography; JULIEN HAVET, *Lettres de Gerbert* (1889), critical edition, superseding previous editions; and HARRIET LATTIN, *The Letters of Gerbert, with His Papal Privileges As Sylvester II* (Eng. trans. 1961), based on manuscripts and the Havet edition of letters—dates letters and privileges, lists Sylvester II's non-extant and spurious documents, and contains an extensive bibliography. LA SALLE DE ROCHEMAURE, *Gerbert Silvestre II* (1914), is the best biography to date (illustrated); see also OSCAR G. DARLINGTON, "Gerbert, the Teacher," *Am. Hist. Rev.*, 52:456–467 (1947), which emphasizes practical aspects but misunderstands Gerbert's astronomical instruments; HARRIET LATTIN, *The Peasant Boy Who Became Pope* (1951), thoroughly researched and illustrated but slightly fictionalized; and E.R. CHAMBERLIN, "Pope Silvester II, 999–1003," *History Today*, 19:115–121 (1969), based on old scholarship.

(H.P.La.)

Symbolism and Iconography, Religious

Symbolism, the basic and often complex artistic forms and gestures used as a kind of key to convey religious concepts, and iconography, the visual, auditory, and kinetic representations of religious ideas and events, have been utilized by all the religions of the world since time immemorial.

NATURE AND SIGNIFICANCE

Religion, according to the traditional interpretation of 18th-century Rationalism, was understood as a system of experiences, ideas, teachings, and practices the central focus of which is a belief in supernatural beings. Because of an emphasis on dogmatic or rationalistic interpretations during that period, the symbolic aspects were de-emphasized. During the age of Romanticism (roughly 1760–1870) the importance of the symbolical and pictorial for religion was rediscovered and revived, as noted in the works of the English poet and artist William Blake and three Germans: the theologian Friedrich Schleiermacher, the philologist and archaeologist George Friedrich Creuzer, and the Roman Catholic writer Joseph von Görres. In the 20th century the symbolical character of religion has often been stressed and upheld in contrast to all attempts to present it rationally. To some extent, the symbolic aspect of religion is even considered by some scholars—because of psychological and mythological studies—as the main characteristic of religious expression. Scholars of the study of comparative religions and of ethnology and psychology have gathered and interpreted an abundance of material on the symbolical aspects of religion, especially in relation to Eastern and primitive religions. In recent Christian theology and liturgical practices a revaluation of their symbolical elements has occurred.

The importance of symbolical expression and of the pictorial presentation of religious facts and ideas has been confirmed, widened, and deepened both by the study of primitive cultures and religions and by the comparative study of world religions. Systems of symbols and pictures that are constituted in a certain ordered and determined relationship to the form, content, and intention of presentation are believed to be among the most important means of knowing and expressing religious facts. Such systems also contribute to the maintenance and strengthening of the relationships between man and the realm of the sacred or holy (the transcendent, spiritual dimension). The symbol is, in effect, the mediator, presence, and real (or intelligible) representation of the holy in certain conventional and standardized forms.

RELIGIOUS SYMBOLS AND SYMBOLIZATION

The word symbol comes from the Greek *symbolon*, which means contract, token, insignia, and a means of identification. Parties to a contract, allies, guests, and their host could identify each other with the help of the parts of the *symbolon*. In its original meaning the symbol represented and communicated a coherent greater whole by means of a part. The part, as a sort of certificate, guaranteed the presence of the whole and, as a concise meaningful formula, indicated the larger context. The symbol is based, therefore, on the principle of complementation. The symbol object, the picture, the sign, the word, and the gesture require the association of certain conscious ideas in order to fully express what is meant by them. To this extent it has an esoteric and at the same time an exoteric, or a veiling and a revealing, function. The discovery of its meaning presupposes a certain amount of active cooperation. As a rule, it is based on the convention of a group that agrees upon its meaning. Symbols, however, may also be individually and subjectively constructed.

Concepts of symbolization. In the historical development and present use of the concepts of symbolization, a variety of categories and relationships must necessarily be differentiated. Religious symbols are used to convey concepts concerned with man's relationship to the sacred or holy (*e.g.*, the cross in Christianity) and also to his social and material world (*e.g.*, the *dharma-cakra*, or wheel of the law, of Buddhism). Other nonreligious types of symbols have achieved increasing significance in the 19th and 20th centuries, especially those dealing with man's relationship to and conceptualization of the material world. Rational, scientific–technical symbols have assumed an ever increasing importance in modern science and technology. They serve partly to codify and partly to indicate, abbreviate, and make intelligible the various mathematical (*e.g.*, $=$, equality; $\equiv$, identity; $\sim$, similarity; $\parallel$, parallel; or $<$, less than), physical (*e.g.*, $\sim$, alternating current), chemical (*e.g.*,  , benzene ring), biological (*e.g.*, ♂, male; ♀, female), and other scientific and technical relationships and functions. This type of "secularized" symbol is rooted, to a degree, in the realm of religious symbolism. It functions in a manner similar to that of the religious symbol by associating a particular meaning with a particular sign. The rationalization of symbols and symbolical complexes as well as the rationalization of myth have been in evidence at least since the Renaissance.

The concept of the religious symbol in its literal sense and in the symbolical expression of religious realities, however, also embraces an abundantly wide variety of

types and meanings. Allegory, personifications, figures, analogies, metaphors, parables, pictures (or, more exactly, pictorial representations of ideas), signs, emblems as individually conceived, artificial symbols with an added verbal meaning, and attributes as a mark used to distinguish certain persons all are formal, historical, literary, and artificial categories of the symbolical. If one looks for a definable common denominator for the various types of symbols, one could perhaps choose the term "meaning picture" or "meaning sign" to best describe the revealing and at the same time the concealing aspects of religious experience. The symbol (religious and other) is intended primarily for the circle of the initiated and involves the acknowledgment of the experience that it expresses. The symbol is not, however, kept hidden in meaning; to some extent, it even has a revelatory character (*i.e.*, it goes beyond the obvious meaning for those who contemplate its depths). It indicates the need for communication and yet conceals the details and innermost aspects of its contents.

Varieties and meanings associated with the term symbol. Different forms and levels of the experience of and relationship to reality (both sacred and profane) are linked together with the concepts of symbol, sign, and picture. The function of the symbol is to represent a reality or a truth and to reveal them either instantaneously or gradually. The relationship of the symbol to a reality (whatever it may be) is conceived of as somewhat direct and intimate and also as somewhat indirect and distant. The symbol is sometimes identified with the reality that it represents and sometimes regarded as a pure transparency of it. As a "sign" or "picture" the representation of the experience of and relationship to reality has either a denotative or a truly representative meaning. The doctrine of the eucharistic (sacramental) presence of Christ in the teachings of Eastern Orthodoxy, Roman Catholicism, and the Protestant Reformers (Luther, Zwingli, Calvin, and Kaspar Schwenckfeld) concretely demonstrate the various and extensive levels of symbolical understandings. These levels extend all the way from the concept of physical identity in the transubstantiation theory of Roman Catholicism (in which the substance of breadness and wineness is believed to be changed into the body and blood of Christ, though the properties of the elements remain the same) through Luther's Real Presence theory (in which Christ is viewed as present, though the question of how is not answered because the question of why he is present is considered more important), and Zwingli's sign (symbolic or memorial) theory, to the concept of mere allusion. The concept of the symbol, however, includes all these interpretations.

Denotative and cognitive aspects

Furthermore, a symbol in its intermediary function has aspects of epistemology (theory of knowing) and ontology (theory of being). As a means of knowledge, it operates in a characteristically dialectical process of veiling and revealing truths. It fulfills an interpretative function in the process of effectively apprehending and comprehending religious experience. In doing so, the word, or symbol—with its meaning, contextual use, relationship to other types of religious expression, and interpretative connection with the various forms of sign, picture, gesture, and sound—plays an important part in the process of symbolical perception and reflection. Although the symbol is an abbreviation, as a means of communication it brings about—through its connection with the object of religion and with the world of the transcendent—not only an interpretative knowledge of the world and a conferral or comparison of meaning to life but also a means of access to the sacred reality. It may possibly even lead to a fusion, or union of some sort, with the divine. To this extent, the sacrament of the Lord's Supper, the liturgical and ritualistic mystery in Christianity—with its many symbolical signs, pictorial representations, significative actions, interpretative words, and various levels of approach to the divine reality—is an example of a highly developed form of a complex symbolic action. Here, the concept of analogy is important; the symbol functions in these ways because it has an analogous cognitional as well as existential relationship to that which it signifies.

The symbolic process. To trace the origin, development, and differentiation of a symbol is a complicated process. Almost every symbol and picture in religion is at first either directly or indirectly connected with the sense impressions and objects of man's environment. Many are derived from the objects of nature, and others are artificially constructed in a process of intuitive perception, emotional experience, or rational reflection. In most cases, the constructions are again related to objects in the world of sense perception. A tendency toward simplification, abbreviation into signs, and abstraction from sense objects is quite evident, as well as a tendency to concentrate several processes into a single symbol. A good example of this last tendency may be seen in ancient Christian portrayals of the triumphant cross before a background of a star-filled heaven that appear in the apses of many basilican churches. In these representations the Crucifixion, Resurrection, Ascension, exaltation, and Transfiguration of Christ are joined to apocalyptic concepts (centring on sudden interventions by God into history) inherent in the doctrine of the Last Judgment. An excellent example of such an apse mosaic is to be found in the S. Apollinare in Classe, near Ravenna (in Italy). On the other hand, there is a tendency to accumulate, combine, multiply, and differentiate symbolical statements for the same thought or circumstance, as seen, for example, on the sarcophagi (stone coffins) of late Christian antiquity—especially in Ravenna. Here, the same idea is symbolically expressed in various manners; *e.g.*, by means of persons, objects, animals, and signs, all appearing side by side.

Development and differentiation of a symbol

The forms and figures of symbolical thought can change into exaggerations and rank growths, however, and lead to transformations and hybrids—figures with several heads, faces, or hands—as exemplified in the statues and pictorial representations of the deities of India (*e.g.*, the multi-armed goddess Kālī) and of Slavic tribes (*e.g.*, the four-headed Suantevitus). The meaning of individual symbols can change and even be perverted. The lamb that in ancient Christian art symbolizes Christ may also symbolize the Apostles or mankind in general. The dove may symbolize the Holy Spirit or the human soul. The wheel or circle can symbolize the universe, the sun, or even the underworld. The encyclopaedic Christian allegorism (symbolism) of the Middle Ages offers many interesting examples, as noted in the writings of Isidore of Seville, a 6th- to 7th-century Spanish theologian, and Rabanus Maurus, a 9th-century German abbot and encyclopaedist.

The foundations of the symbolization process lie in the areas of the conscious and the unconscious, of experience and thought, and of sense perception, intuition, and imagination. From these arise the structure of religious symbolism. Sensation and physiological and psychological processes participate in the formation of the symbol structure. Extraordinary religious experiences and conditions, visions, ecstasy, and religious delirium brought about by intoxication, hallucinogenics, or drugs that produce euphoria and changes in consciousness must also be taken into consideration. The symbol itself, however, is intended as an objective concentration of experiences of the transcendent world and not as a subjective construction of a personally creative process. In cultic and mystical visions and trances, the forms and processes of the external world and of the religious tradition are condensed and combined with mythical images and historical events and take on a life of their own. The process of rational conceptualization and structuralization, however, also plays a part in the origin and development of many symbols. There is a correlation between sense perception, imagination, and the work of the intellect.

The foundations of the symbolization process

Symbols in the religious consciousness. The formation of religious symbols that occur when unconscious ideas are aroused or when a process of consciousness occurs is principally a matter of religious experience. Such symbols usually become intellectual acquisitions, and, as religious concepts are further elaborated upon, the symbols may even finally become subjects of major theological questions. In Christian theology, for example, summaries of dogmatic statements of faith are called symbols (*e.g.*,

the Apostles', Nicene, and Athanasian creeds and the confessional books of Protestantism, such as the Augsburg Confession of Lutheranism). This particular use of the term symbol is exceptional, however.

In the development of the symbol, religious experience, understanding, and logic are all connected, but each places different accents on the individual categories and species of symbol. Occasionally, religion is regarded as the origin and the product of certain established (or fundamental) symbols. In such cases the outcome of the process of the structuralization of religious consciousness would then be the establishment of a symbol that is generally applicable to a particular historical species of religion. Conversely, one could ask whether the experience and establishment of an individual or collective symbol by a creative personality or a community is not itself the establishment of a religion. If so, the classical symbol that was developed at the time of the foundation of any one particular religion would then be constitutive for its origin and further development (*e.g.*, the T'ai Chi or the combination of Yin and Yang for the Chinese, the cross for the Christian religion). In any event, the symbol belongs to the essence of man's coming of religious consciousness and to the formation of history's institutional religions. It plays a fundamental and continual part in the further growing of such religions and in the mental horizons of their followers.

Religion as both origin and product of symbols

THE RELATION OF THE SYMBOL AND THE SACRED

Symbols as the incarnate presence of the sacred or holy. Whatever the experience of reality that lies behind the religious symbol may be, it is above all the experience of the sacred or holy, which belongs essentially to any concept of religion. The historical study of religions has shown that it is fundamentally the symbol that mediates and forms for man's religious consciousness the reality and the claim of the holy. Religion is a system of relationships, a system of reciprocal challenges and responses the principal correspondents of which are the sacred or holy and man. Though there are many forms of experience in which the sacred or holy is distinctly known and felt, the experience is often acquired in worship, in which this system of relationships is realized and continually renewed and in which the sacred or holy supposedly makes itself present. The details of worship serve to objectify and regulate in a perceptual and material manner the presupposed presence of the sacred or holy, of which the symbol and the picture are intended to be its materialization. In its material manifestation the sacred or holy is adapted to man's perceptual and conceptual faculties. Viewed from the aspect of its holiness, the symbol originates in a process of mediation and revelation, and every encounter with it is supposed to bring about a renewed actualization and a continual remembrance of this revelation.

The actualization of the presence of the holy by means of symbolic representation can, in extreme cases, lead to an identification of the physical manifestations with the spiritual power symbolized in them. The symbol, or at least an aspect of it, is then viewed as the incarnated presence of the holy. The sacred stone, animal, plant, and drum and the totem symbol or the picture of ancestors all represent the sacred or holy and guarantee its presence and efficacy. The origin of many such symbols clearly indicates the identity that was presumed to have existed between the symbol and the sacred or holy. The Greek god Dionysus as a bull, the Greek goddess Demeter as an ear of corn, the Roman god Jupiter as a stone, the Syrian god Tammuz-Adonis as a plant, and the Egyptian god Horus as a falcon all are viewed as manifestations of the deities that were originally identified with these respective objects of nature.

The symbol as the incarnated presence of the holy

Symbols as indicators of the sacred or holy. The symbol is understood to have a referential character. It refers to the reality of the sacred or holy that is somewhat and somehow present. When the symbol is an indicator of the sacred or holy, a certain distance exists between them, and there is no claim that the two are identical. Short of actual identification, various degrees of intensity exist between the symbol and the spiritual reality of the sacred or holy. The symbol is a transparency, a signal, and a sign leading to the sacred or holy. The objects, gestures, formulas, and words used in meditation—for example, the Buddhist *mudrās* (gestures), *pratimās* (images), *mantras* (magic formulas)—and in mysticism—for example, the crystal or the shoemaker's ball in the contemplative experience of Jacob Böhme, a 16th- to 17th-century German cobbler and mystic; the navel in Omphalicism (a method [called Hesychasm] of contemplating the navel in order to experience the divine light and glory in medieval Greek Christian mysticism of the monks of Mt. Athos); and the pictures of the deity in the language of Hindu, Islāmic, and medieval Christian mysticism—all of these are truly symbols, but nonetheless they have at most only an indirect mediating relationship to the divine, a purely noetic (intellectually abstract) significance with regard to the reality and presence of the sacred or holy.

Symbols of sacred time and space. The symbolical forms of representation of the sacred or holy are to be understood as references to or transparencies of the sacred or holy. The sacred manifests itself in time and space, so that time and space themselves become diaphanous indications of the holy. The holy place—a shrine, forest grove, temple, church, or other area of worship—is symbolically marked off as a sacred area. The signs, such as a stake, post, or pillar, that delimit the area themselves are endowed with sacred symbolic meanings, which often can be noted by their particular designs. The ground plan of the sacred building and its orientation, walls, roof, and arches are all utilized to symbolize the sacred or holy. Prehistoric places of worship—*e.g.*, Stonehenge (in England) and other megaliths of Europe and the shrines and holy places of ancient Egypt, Babylon, China, and Mexico—were invested with symbolical meanings.

Sacred places are often pictorial reflections of the universe and its design and partake of its holiness. The domes of Christian churches are symbols of heaven, the altar a symbol of Christ, the Holy of Holies of the Temple in Jerusalem a symbol of Yahweh, the Holy of Holies in Shintō shrines (*honden*) a symbol of the divinity, and the prayer niches in mosques a symbol of the presence of Allāh. In many instances shoes may not be worn on holy ground (*e.g.*, Shintō temples), and hands and feet are to be washed before entering into a holy place. The woodwork of demolished Shintō shrines, when taken to private homes, makes the sacred or holy present in the homes of pious Japanese families.

Time as a transparent symbol of the sacred may be represented by means of the cycle of the sacred year and its high points—*e.g.*, New Year's (as in ancient Near Eastern religions), the times of sowing and reaping, and the solstices and equinoxes. Or the lapse of time may be represented in signs and pictures. Cosmic, mythical, and liturgical time and destiny are portrayed, for example, in the Buddhist symbol of the wheel of life, *bhava-cakra*, with its causal chain of human deeds and succession of existences, entwined by the claws of a devouring monster; the figures of Aion (Time) in late Greco-Roman and Persian antiquity show a figure with a winged lion's head standing on a globe and encircled by a snake. Time itself, its course, division, and fixed points, is both an allusion and the bearer and mediator of the sacred or holy (see also FEAST AND FESTIVAL; CHURCH YEAR; JEWISH RELIGIOUS YEAR).

The cycle of the sacred year

Ceremonial and ritualistic objects as indicators or bearers of the sacred or holy. Liturgical and ceremonial objects can also indicate or lead to the sacred or holy. Not only holy pictures and symbols (*e.g.*, the cross in Christianity or the mirror in Japanese Shintō) but also lights, candles, lamps, vessels for holy materials, liturgical books, holy writings, vestments, and sacred ornaments are indicators of the sacred or holy. Liturgical vestments and masks are intended to transform the wearer, to remove him from the realm of the this-worldly, and to adapt him to the sphere of the sacred or holy; they help him to come into contact with the divine—for example, by obscuring his sexual characteristics. The vestments

may be covered with symbols, such as those worn by Arctic shamans (medicine men with psychic transformation abilities). They are signs of the function of the wearer and his relationships to the sacred or holy and to the profane world. Such vestments are frequently derived from those of rulers or from ceremonial court dress; *e.g.*, Japanese Shintō and Roman Catholic and Eastern Orthodox Christianity. They are supposed to create a fitting atmosphere of solemnity and dignity. In Western Christianity, the liturgical vestments have a very specific symbolism: the alb (a tunic) symbolizes purity of heart; the stole, the raiment of immortality; and the chasuble (an outer eucharistic, or holy communion, vestment), the yoke of Christ. The liturgical vestments of the Eastern Christian churches have a similar symbolism. The ritual headdress and the crown express the sacred dignity of the wearer. The vestments of the various religious orders (Oriental and Occidental) express the holiness of the members of the community, their nearness to the sacred or holy, and the significance of religious life for them. In the reception ritual of Jainism and Buddhism, the monastic vestments are put on as a sign of an entrance in a new state of life. This ritual in Jainism resembles that of a wedding ceremony. The taking over of the monastic garb is an essential part of becoming a *sādhu*. The monks of the Jainistic Śvetāmbara sect wear five objects (*e.g.*, shells) as symbols of the five monastic virtues. In early Christianity the white baptismal vestment was a symbol of rebirth, new life, and innocence (see also RELIGIOUS DRESS AND VESTMENTS; CEREMONIAL AND RITUALISTIC OBJECTS).

Other relations between the symbol and the sacred. The sacred or holy as represented or manifested in the symbol has, generally speaking, a sanctifying function (elevating one to a closer relationship to the sacred or holy) and an exorcising function (decreasing or eliminating those aspects that hinder one's relationship to the sacred or holy). Remembrance (*anamnēsis*) and imitation (*mimēsis*) are the analogous and associative means of representing the reality and indestructibility of the sacred or holy and its power, which defends, protects from injury, bans evil, and guarantees salvation. Symbolic signs and pictures (*e.g.*, masks; sex, animal, or plant symbols, such as the skulls or horns of animals) are placed on houses and sacred places to make present the saving and sanctifying power of the sacred or holy.

RELATION OF RELIGIOUS SYMBOLISM AND ICONOGRAPHY TO OTHER ASPECTS OF RELIGION AND CULTURE

Relation to myth and ritual. The symbol has a long-established relationship with myth (sacred stories that define the human condition and man's relation to the sacred or holy). Often containing a collection of symbolic forms, actions, expressions, and objects, myths describe gods, demons, men, animals, plants, and material objects that are themselves bearers of symbolical meanings and intentions. Thus, it is sometimes difficult to distinguish between a myth and a coherent complex of symbols brought together in story form. Examples are myths of cosmogony (origin of the world), theogony (origin of the gods), and anthropogony (origin of man). The details and contexts of religious teaching, dogma, and theology also produce or form symbolic values or refer to traditional symbolic representations. Symbol structures and pictorial representations are brought into connection with dogma and theological statements—*e.g.*, the Buddhistic *karma-saṃsāra* (law of cause and effect and reincarnation) theory and the *bodhisattva* (buddha-to-be) theory or the Christian teaching of the Last Judgment, punishment of sin, hell and purgatory, and eternal reward (Paradise). In worship, individual actions and objects used in the ritual are given a symbolic meaning that transcends their immediate practical purpose. Magic, in its ritual, also uses various formations of symbols, pictures, and symbolical actions that may be seen as parallels to the distinctively religious use of symbols (see also MYTH AND MYTHOLOGY; RITUAL).

Relationship of symbolism and doctrine

Relation to meditation and mysticism. The spiritualization of religious experience in forms of meditation and mysticism assimilate and rework the existing symbols and pictures of an older historical period of religion, giving to some symbols a higher value and placing others in the centre of focus. At the same time it develops new forms the appearances of which stem especially from the visionary experiences of the mystic and from his need for a suitable means of expression and from the objects of meditation training; *e.g.*, holy sounds and words (*om*), the lotus flower, the *vajra* (ritual object shaped like a thunderbolt), and the wheel in Buddhist meditations or the ladder, the heart, and the letters IHS (the first three letters of the Greek word for Jesus) in Christian mysticism. In contemplation, colours, forms, sounds, signs, and pictures become ways and means of penetrating to the centre of the mystical union. Jacob Böhme's work is characteristic of the development of an especially rich mystical language of symbols. Mysticism supplies conventional and customary religiosity with new pictures and symbols (see also MYSTICISM).

Relation to the social realm. The field of symbolism and iconography shows a strong interdependence that existed between religion and other areas of culture that were later to become autonomous and profane (or secularistic). The social domain under the influence of religion develops its own symbolism for expressing its values and objectives. Conversely, religion often draws its symbols and pictorial forms from the social, political, and economic domains. Persons (*e.g.*, king, father, mother, child, slave, brother) and conditions and structures in society and the state (*e.g.*, government, a people, family, marriage, occupation) all receive meaning as symbolical and pictorial motifs in myth and cult. Examples of such motifs are throne, crown, sceptre, standard, arms, instruments, the figures of the father, mother, and child, and symbols of familial relationships. The morals, law, administration of justice, and the customs and habits of a society contain religious symbols and symbolical actions, as in the anointing of a king and in the administering of the oath or ordeal or in the observance of traditions and customs associated with birth, marriage, and death.

Reciprocal influences of culture and religion

Relation to the literary and visual arts. Religious symbols and pictures may be identical with, related to, or similar to those of language (metaphors) and to pictorial expressions in prose and poetry. They are related in allegory, parable, fairy tales, fables, and legends in which they can appear in a form that is closely related to that of religious symbolism. Religious symbols are used in the plastic arts, in architecture, and in music. Symbols also have been developed in those arts and then introduced into religion. A few examples of such symbols are house, room, door, column, sound, harmony, and melody (as when Christ was viewed as the "new melody" in the words of Clement of Alexandria, a 2nd-century philosophical theologian). Here, also, the interdependence and the continual reciprocal influence of religion and culture may be observed.

Relation to other areas of culture. The formation of religious symbols and pictures has been stimulated by numerous other areas of human culture—such as the philosophy of nature, the natural sciences (especially botany and zoology), alchemy, and medicine (including anatomy, physiology, pathology, and psychiatry). In the works of Jacob Böhme, alchemy (*e.g.*, the elements, fire, salt, sulfur, mercury, tincture, gold, essence, the philosopher's stone, and the transmutation) found an all-inclusive symbolical use; and in the works of Robert Fludd, an English physician and mystical philosopher of the 16th and 17th centuries, medical, cosmological, alchemical, and theosophical (esoteric religious) symbols were fused together (*e.g.*, the contrast of light and darkness and the idea of man as a microcosm). Symbols, also religious and mythological (such as signs of astral gods for the planets in astronomy), have achieved new importance in the conceptual presentations of distinctively scientific systems; *e.g.*, in physics, cosmology, psychiatry, and psychology. Even spaceships bear symbolic or mythical names. Psychoanalysis and depth psychology have re-evaluated the role of the religious symbols and have used them in interpreting psychological processes, such as in the works of the Swiss psychiatrist Carl Jung. Jung

Relationship of religious symbolism to scientific symbols

interprets religious processes as symbolic ones and emphasizes the growing of individual and social symbols in the unconscious. According to his interpretations, many of the symbols, transforming the archaic libido into other functions, come out of dream experiences in a kind of intuition or revelation. Important symbols are duality (male–female, animus–anima), trinity, and quaternity.

Changes in symbolical relations and meanings. Symbols emerge and disappear and change in their value and function. Although symbols have a tendency to be normative, stable, and to have a fixed meaning value, the demise of old symbols and the genesis of new ones or changes in the meaning of existing symbols nevertheless occur. Many ancient Christian symbols (*e.g.*, the fish) had long lost their recognition value or had been pushed into the background. With the renewal of ancient Christian symbolism in modern times, they have had a re-evaluation. The triangle and the eye as recently used in Christianity are relatively new symbols for God. The old and formerly very meaningful religious symbolism of the axe and hammer has almost disappeared. The symbolism of kingship and sovereign authority has, on the other hand, been maintained in religious language and in the religious conceptual framework, although the political structures from which it originated have disappeared or lost their relevance. The disintegration of individual symbols and the change in the emphasis on the role of symbolism in general are partly consequences of cultural, intellectual, social, and economic transformations.

MODES OF SYMBOLIC EXPRESSION

In the long history of the forms of symbolical expression a narrower (exclusive) and broader (inclusive) idea of what a symbol is has gradually evolved. This evolution is reflected in the various manners of symbolical expression that may influence and combine with one another. Many scholars question whether a picture or a verbal expression, for example, strictly corresponds to the idea of a symbol. Just as the ideology and terminology of the ancient Greek mystery (salvatory) religions distinguished between that which is shown and seen (*deiknymenon*), that which is done (*drōmenon*), and that which is said (*legomenon*), so also can one make distinctions among three types of symbols: the visual symbol, the symbolical action that is dramatically enacted in worship, and the linguistic symbol, which includes music and other sounds. Viewed in these various aspects, the complex character of the symbol becomes apparent.

Diagrammatic and emblematic. Symbolic representations are usually depicted in diagrammatic or ideographic modes as signs, abbreviations, images, and objects of all kinds that indicate a larger context. In this category belong the simplified or abstract forms of objects of nature or other objects and geometrical forms, as well as colours, letters, and numbers. The circle, the disk, the rosette, or the swastika, for example, may symbolize the sun, universe, or a star. The square and the cross may symbolize the Earth or the four cardinal points; the wreath, the labyrinth, the spiral, the plait, and the knot may indicate eternity, the flow of time, or a magical spell.

Geometric, numerical, chromatic, and alphabetical symbolism

Ornamental designs in primitive artwork, those of the American Indians, for example, frequently have a symbolical meaning and embody fundamental figures, such as the straight line, circle, rectangle, rhombus, or ellipse. The cross in its varying forms: the Latin (†), Tau (T), ankh (☥), Saint Andrew's (X), and the forked (Y) may symbolize man and his extremities. Among various peoples and in different religions a number of basic colours have at times different and sometimes even opposite meanings. White, for example, may signify joy and festivity or death and sadness. Red has the most pronounced symbolical value: it refers to the liturgical, priestly sphere and also to life and death. In Christianity, colour symbolism is associated with the sacred year; in Buddhism with the picture of the universe, the regions of which are classified according to particular colours; and in the religion of the Maya of Mexico and Central America with the four world directions—east (red), north (white), west (black), and south (yellow). The symbolism of metals and precious stones also is related to their colours (*e.g.*, emerald with green).

The symbolism of the letters of the alphabet varies according to the alphabet (*e.g.*, A and Ω in Greek or *A* to *Z* in English) and is often connected with magic and prophecy—which is also true of the symbolism of numbers. In the picture writing of hieroglyphic systems and in the ideographic (idea-sign) writing of earlier times there is a direct relation between the word-sign and the object to which it refers. In alphabetic writing, numbers and letters are interchangeable if the letter has a number value, as, for example, in Greek and Hebrew alphabets. In some religions, the world of the gods is arranged according to a number system; *e.g.*, into enneads (nines), triads (threes), or dyads (pairs). The idea of oneness was then extended to a numerically arranged pantheon. Gnosticism (a Hellenistic esoteric dualistic system) and Kabbala (a Jewish esoteric mystical system) developed number symbolisms. The letters received a symbolic character in two ways: first, as components of a word for which they stood—*e.g.*, the Hebrew tetragram (four-letter) YHWH (Yahweh) for God, the Latin IOM for Jupiter Optimus Maximus (Jupiter the Best and Greatest), and the Greek IHS for Jesus—or, second through their numerical value, as in A and Ω, the beginning and end of the Greek alphabet, signifying Christ. They then became means of abbreviation—signs possessing a specific content and meaning.

Pictorial. Pictorial symbolism in its many forms is a further development of nonrepresentational, ideographic symbolism and also, to some extent, its origin. In depicting the world of nature, pictorial symbolism captures and mediates the religious experience of reality. The picture shows plainly and clearly the rich and intricate connections of its symbolic content. It may present a part for a whole (a head, a hand, a foot, or an eye for a complete figure) or the whole itself. Symbolic expression of religious experience by means of painting has had a long history.

Sculptural representations of the sacred or holy have their origin in cult. They range from Stone Age idols to the sacred sculpture of early Mesopotamian and Egyptian cultures, from the statues and reliefs of Greco-Roman gods, divinized heroes, and their deeds to the symbolic sculpture of India with its Hindu gods and demons, from the sacred sculpture of China and Japan with their respective pantheons to that of Mahāyāna (Greater Vehicle) Buddhism with its *bodhisattvas* (buddhas-to-be), saints, and spirits. These sacred figures, which may appear in statue or relief form, are sculptured out of various materials. The reliefs on the interiors and exteriors of temples have a decorative function similar to that of wall painting. They narrate a myth or tell a sacred history. Particular parts of the body and symbolical objects may also be sculpturally represented. They may be the male and female sexual organs (*e.g.*, the *liṅga-yoni* in Hinduism); the hand of Sabazius, a Greek god sometimes identified with Dionysus (the god of wine), whose hand is portrayed as raised in blessing and encircled by a number of rather bizarre appendages; and human limbs used as votive offerings for the cure of the part of the body represented. Representative symbolic sculpture tends either to simplify the figure in an abstract, geometrical, or expressionistic style or to imitate nature realistically.

Sacred figures in sculptured or relief forms

Gestural and physical movements. Gestures and bodily movements play an important part in religious ritual and in the religious conduct of man. Such behaviour derives its meaning from its relationship to the holy.

In proceeding to and from a holy place, a worshipper generally proceeds according to certain symbolic patterns: rectilineal, circular, and vertical. Rectilineal movement to and from a holy place is intended to gradually prepare the worshipper for the spatial encounter with the holy and after the encounter to remove him from the sacred sphere. Special streets for processions often are marked off or built to a temple or holy place, such as in ancient Egypt, Mesopotamia, and China. The great procession from Athens to Eleusis for participants in the mysteries possessed a symbolic meaning. Worshippers not only enter a holy place but may also walk around it.



Triumphant cross, symbolizing Christ, before a star-filled heaven, the eternal, surrounded by the symbolic words or letters ΙΧΘΥΣ (ICHTYS; fish), SALUS MUNDI (salvation of the world), and Α and ω (Alpha and Omega; beginning and end); one set of symbols bringing together several Christian ideas of the relationship between man and God, the temporal and eternal, earth and heaven, the transfiguration and ascension of Christ, the Last Judgment. Apse mosaic in the church of S. Apollinare in Classe, near Ravenna.



Avalokitesvara, the compassionate *bodhisattva*, shown as a composite figure with eleven heads and eight arms, symbolic of his ability to sense man's needs everywhere in the universe. In the Rijksmuseum voor Volkenkunde, Leiden, The Netherlands. Height 35 cm.

The symbol: unity and multiplicity



Two-headed Janus, who sees forward and backward, personification of the month of January; romanesque high-relief sculpture of stone In the Museo del Duomo, Ferrara, Italy.



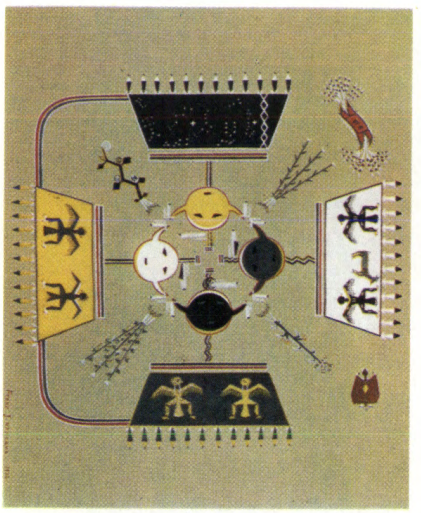
Buddha s footprint, the "sign of a grand man," containing the symbols of the swastika, the wheel of the law, the *tri-ratna* (the "three jewels" of the Buddhist creed), and the lotus. Stone relief from Gandhara, 2nd century AD. In the Prince of Wales Museum of Western India, Bombay.



The Trinity represented by Christ as man, the Holy Spirit as a dove, and God as a hand; Armenian miniature of the baptism of Jesus, 1273. In the Topkapı Museum, Istanbul. 16 × 11 cm.

Plate 1: By courtesy of (top right) the Rijksmuseum voor Volkenkunde, Leiden, The Netherlands, (centre) the Prince of Wales Museum of Western India, Bombay; photographs, (top left and bottom left) SCALA, N.Y., (bottom right) Ara Guler

Symbols of space and time



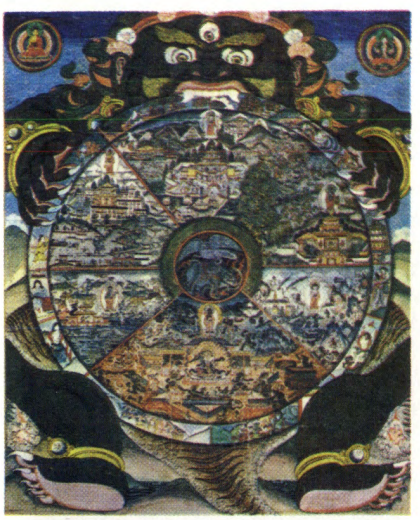
"The Skies," watercolour copy of a Navajo sand painting. Four holy plants radiate from the central four storm figures. The dawn, midday, evening, and night skies are shown in the east, south, west, and north. Paired guardians in the east are a medicine pouch and a bat. In the Museum of Navaho Ceremonial Art, Santa Fe, New Mexico.



Cathedral of Notre Dame, Chartrès, France, completed mid-13th century. The nave and apse form a cross; the steeple and bell tower may be interpreted symbolically as the "fingers of God."



Borobudur *stupa*, Buddhist monument in central Java, built in the form of a *mandala*, late 8th century. The surrounding galleries are decorated with reliefs representing scenes of the life and religious development of the Buddha.



Bhava-cakra ("wheel of life," circle of existences), Tibetan *tanka*, probably 19th century. In the Newark Museum, New Jersey. 109.2 × 86.3 cm.

Stonehenge, circular earthwork and stone religious site, Wiltshire, England; late Neolithic to Early Bronze Age (1800–1400 BC).

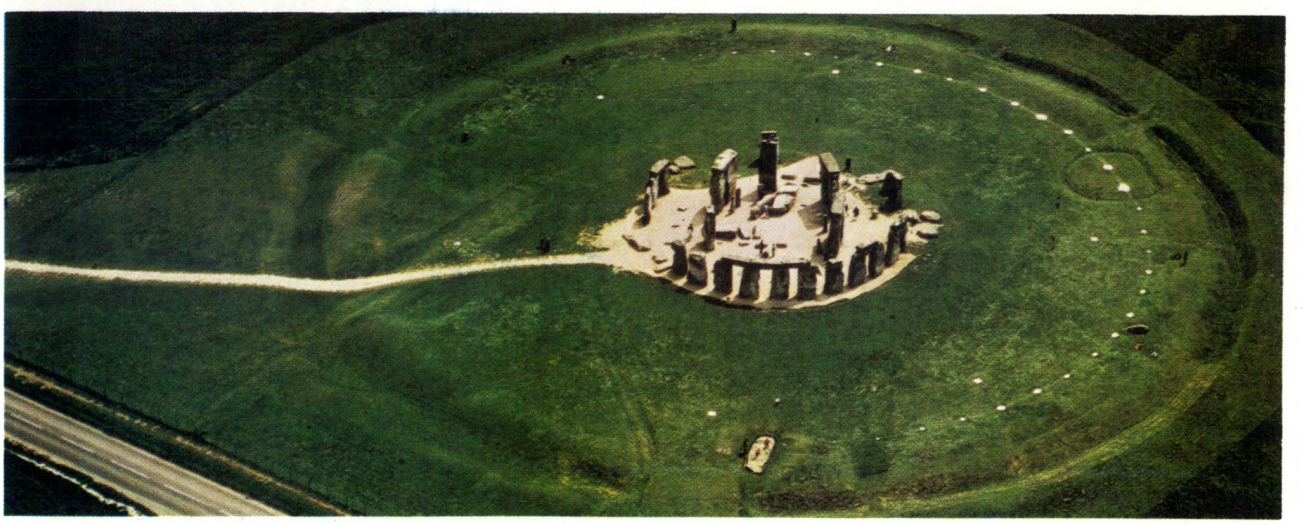


Plate 2: By courtesy of (top left) the Museum of Navaho Ceremonial Art, Inc., Santa Fe, New Mexico, Newcomb Collection, (centre left) the Newark Museum, New Jersey; photographs, (top right) Alain Perceval, Paris, (centre right) Brian Brake—Rapho Guillumette, (bottom) Aerofilms Ltd.



Poseidon, Apollo, and Artemis, detail from the marble east frieze of the Parthenon, Athens, *c.* 440 BC. In the Acropolis Museum, Athens. Height 109.2 cm.

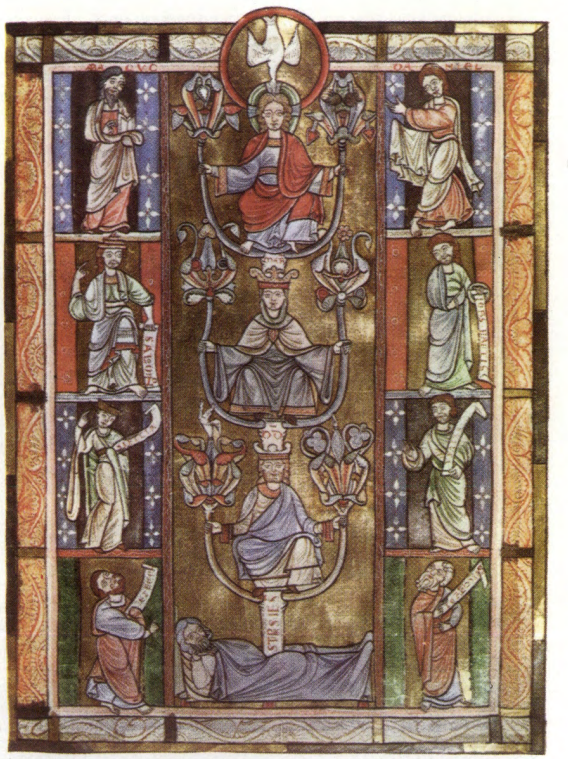
The sacred or holy in the image of man, animals, and plants



Horus as a falcon, Egyptian bronze, 26th dynasty to Ptolemaic dynasty (7th–3rd centuries BC). In the Brooklyn Museum, New York. Height 28.8 cm.



Buddha Amitabha and two Bodhisattvas on lotus blossoms, the "Amida-Trias." Japanese bronze known as the Tachibana Shrine, 7th–8th century. In the Horyu-ji, Nara Prefecture, Japan.



"Tree of Jesse," symbolizing Jesus' descent from the house of David, illuminated page from Rabanus Maurus's *De laudibus sanctae crucis,* from Anchin, mid-12th century. In the Bibliothèque Municipale de Douai, France. 29 × 20 cm.



The four Evangelists represented by their symbols—Matthew by a man, Mark by a lion, Luke by an ox, and John by an eagle—illuminated page (f.27v) from the Book of Kells, Hiberno-Irish, 8th century. In Trinity College Library, Dublin.

Plate 3: By courtesy of (top left) the Acropolis Museum, Athens, (centre) the Brooklyn Museum, New York, Charles Edwin Wilbour Fund, (top right) the Bibliotheque Municipale de Douai, France, (bottom right) the Board of Trinity College, Dublin; photographs, (top left) G. Nineatallali, (top right) Giraudon, (bottom left) Benri-do, Kyoto

Symbols from the realm of family and society



Krsna (right) with his beloved, Rādhā, painting in the Kangra style from India, 18th century. In the Prince of Wales Museum of Western India, Bombay.



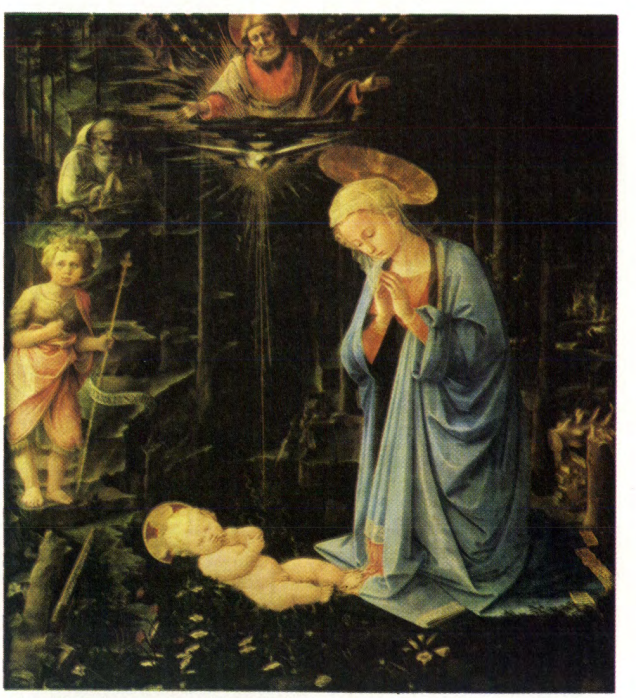
The child Krsna, bronze from India, 16th–17th century. In the Detroit Institute of Arts. Height 11.4 cm.



The warring archangel, shown in "St. Michael Triumphing Over the Devil," bronze by Sir Jacob Epstein, 1959. Coventry Cathedral, England. Height 7.6 m.



Christ enthroned as Lord of All (Pantocrator), with the explaining letters IC XC, symbolic abbreviation of "Iesus Christus." Mosaic in the cathedral of Monreale, Palermo.



"The Virgin Adoring the Child," painting by Fra Filippo Lippi, *c.* 1460. In the Staatliche Museen Preussischer Kulturbesitz, West Berlin. 127 × 116.8 cm.

Plate 4: By courtesy of (centre) the Detroit Institute of Arts, (bottom right) the Staatliche Museen Preussischer Kulturbesitz, Berlin; photographs, (top left) Pramod Chandra, (top right) A.C. Cooper Ltd., (bottom left) Farrell Grehan—Photo Researchers

Rectilineal and circular movement thus complement each other. Movement to and from a holy place may also be vertical as well as horizontal, as to and from a holy place on top of a mountain or pyramid. All these various types of movement give expression to the symbolism of the holy way or path.

The sacred dance combines rectilineal and circular movements and may also include hopping, jumping, and hand movements. Hand and finger movements in temple dances in Indian and other Asian cultures are strictly regulated and have a precise symbolic meaning. The liturgical dance in a rudimentary form was maintained for a long time in Christianity, as has been the procession. Dancing has not only a significative but also a magical function. It seeks to enchant the holy power.

Symbolism of hand movements

Hand movements are widely used in ritual and liturgical actions; the touching of holy objects, materials, or men is performed according to a canon (rule) that precisely regulates these gestures and their accompanying prayers and blessings. The gesture of blessing may imitate a symbolic form, such as that of the cross in Christianity. Here the position of the fingers is regulated and has a special meaning, as is also true in the Hindu and Buddhist practice of meditation (*mudrās*). Stroking, thrusting, striking, pushing, waving, and hand clapping also can be symbolical gestures. By raising his hands in prayer, the worshipper approaches the realm of the heavenly gods; by kneeling, the realm of the underworld. This apparently was the original meaning of kneeling before it became an expression of humility. The bow as an intimated genuflection generally indicates respect. The kiss and the embrace—and sometimes also the actions of breathing or spitting upon someone or anointing a person with spittle—were originally magical manipulations; in later usage, they indicated union with or a strengthening of the community or the transferral or communication of power. The holy kiss, whether practiced or only verbally depicted, plays an important part in many religions. Standing is a posture of respect; sitting expresses the reception and acceptance of the sacred word or teaching. It is also the position for meditation as it is practiced in Buddhist monasteries. Symbolic gestures may be either individually or collectively performed.

Verbal symbolism. Gestures are usually accompanied by words. The spoken and written word in religion generally is not thought of primarily as symbolic but rather as a form of rational communication, of communication of thought. Despite its predominantly rational character in modern times, however, language does develop expressions that extend into the area of the symbolical. In its origin, language most likely was richly symbolical. Linguistic symbolism, however, has always had a certain tendency toward rational transparency and logical coherence, and thus words, objects, and pictures—in their origin as symbols—are very closely related. The visual value of the object and picture is later translated into language and enhanced by it.

Linguistic symbolism generally is metaphorical; the allegory, a particular development of the metaphor, symbolically represents an idea by means of a coherent complex of metaphors. Specific genres of narration and literature, such as myth, belong in this category. In a figurative, interpretative, and cryptic sense, names and metaphors denote the person or thing in question. God sometimes is metaphorically called a "spring" or a "rock"; Christ, "the Beloved"; Mary (the mother of Jesus), "the Rose"; and Vardhamāna Malāvīra (the founder of Jainism) and the Buddha (Siddhārtha Gautama, the founder of Buddhism) are called "the Conqueror."

Symbolic syllables and names

Individual syllables or sounds may also have a symbolic quality. The *om*, which is used to introduce the holy texts of Hinduism and is a meditation syllable used in Buddhism, provides one example. Understood magically as an emanation of the divine, the word or a name or a part of a word can become an independent hypostatized (substantial) object, a representation, or even the incarnation of the divine, such as the Logos (Word) in the Gospel According to John or *hū* ("he") and *al-ḥaqq* ("the truth") in Islāmic mysticism or the name Metatron in Kabbala. A holy writing or book in its entirety may represent the divine in the same way, as the Bible in Christianity, the Qur'ān in Islām, and the *Ādi Granth* in Sikhism.

Musical symbolism. Music, like the word, also may have symbolic meaning. The basic elements out of which musical symbolism is built are sounds, tones, melodies, harmonies, and the various musical instruments, among which is the human voice. Sound effects can have a numinous (spiritual) character and may be used to bring about contact with the realm of the holy. A specific tone may call one to an awareness of the holy, make the holy present, and produce an experience of the holy. This may be done by means of drums, gongs, bells, or other instruments. The ritual instruments can, through their shape or the materials from which they are made, have symbolic meaning. The Uitoto in Colombia, for example, believe that all the souls of their ancestors are contained in the ritual drums.

The relationship between religious ideas and music is of special importance when the sacred word is set to music or when the music supports or interprets the sacred word by orchestral accompaniment. Medieval and modern Christianity in the West has made important contributions in this area. The symbolic word may be enriched, intensified, and increased in meaning when it is given a musical form. In the medieval technique of motet composition, different but parallel texts from the Bible or the liturgy would be simultaneously sung in various voices to appropriate but different melodies. This is an example of the structuralization of symbols into a coherent whole, a process that may sometimes also be encountered in the visual arts.

Conjunction and combination of various modes. In ritual, liturgy, liturgical and devotional art, and in religious literature and experience, many different types of symbolical expression are frequently combined. Pictorial art may be symbolically interpreted or its present meaning may be reinforced by the addition of a verbal explanation or possibly even by music. In ritual, symbolical words, tones, noises, gestures, signs, odours (*e.g.*, the odour of the sacrifice or the fragrance of incense as an expression of prayer and offering), colours, and pictures are combined. Pictorial art often depicts religious texts and ideas; in so doing it not only uses the human form but also objects of nature, scenery, sacred architecture, and particular symbols. A picture or sign on an emblem often receives its interpretation by the inscription of a verbal explanation. Conversely, in an illustration of religious texts, the picture or sign interprets the text. Over against verbal and musical symbols stands the sacred value of silence. It may indicate devotion, contemplation, or the presence of God.

ICONS AND SYSTEMS OF ICONOGRAPHY

Throughout the history of their development, religious iconography and symbolism have been closely interrelated. Many religious symbols can be understood as conceptual abbreviations, simplifications, abstractions, and stylizations of pictures or of pictorial impressions of the world of sense objects that are manifested in iconographic representations. In conceiving, describing, and communicating the experience of reality, the realistic picture and the nonrepresentational sign both have as their primary function the expression of this experience in religious terms. In religious pictures that are of a compound or complex nature, particular symbols occasionally reappear. These pictures may also include other types of symbolic representation, such as words, tones, gestures, rituals, and architecture.

Relationship of symbolism to experience

Iconographic forms. *Temples and other sacred places.* The architectural iconography of sacred buildings and places of worship is a field of its own. The place of worship, insofar as it is understood as the image of the universe and its centre, must be architecturally patterned according to a specific design of the universe. The place of worship may be considered to be the navel of the world; *e.g.*, the *omphalos*, a round stone in the temple at Delphi (in Greece), the holy stone in the Church of the

Holy Sepulchre, in Jerusalem, or the rock in the temple area of the Dome of the Rock (Mosque of Omar), in Jerusalem. A holy place usually is built around these holy points.

The cross-shaped ground plan of the Christian transept church is sometimes interpreted as an architectural portrayal of the crucified Christ, the apse with its altar representing Christ's head. The holy place as a structural creation together with its natural setting may create an idyllic or overwhelming effect, evoking in the beholder an experience of religious awe or devotion. The Shintō and Buddhist temples of Japan and the beauty of the landscape in which they are set, the mountain temples of ancient Greece, and Christian churches and chapels built in such dramatic settings as Mont-Saint-Michel in France all inspire a sense of wonderment. The Buddhist temple in all the splendour and richness of its form, trappings, and surroundings or the *stūpa* (a building containing relics of the Buddha) represents the presence of the Buddha.

Great importance, therefore, is often attached to the exterior form of the holy place, and its construction is governed by a canon of symbolical and iconological principles. The individual parts of the building—the walls, columns, ceilings, vaults, and towers—usually have pictorial and symbolic functions. Generally, the ceiling or vault presents a picture of heaven. Special accent is placed on the portals and the paths leading to them, on the position of the tables of offering, altars, sacred pictures, and relics. The bell tower, or campanile, is characteristic of Christian churches and is popularly interpreted as the finger of God. Ancient Christian basilicas (large, roofed buildings, generally with aisles) were viewed as images of the heavenly Jerusalem. The pictorial aspect of the place of worship extends not only to the building in the entirety of its architectural form but also to the painted, sculptured, and mosaic artwork that decorates it. The exteriors of Hindu and Buddhist holy places, such as the famous terrace temple of Borobuḍur on Java, and the pediments and friezes of Greek temples utilize an abundance of figures and reliefs representing scenes from myth and sacred history. The facades of Egyptian temples are covered with tableaus of the gods and depictions of ritual ceremonies. The facades and portal walls and sometimes the outside walls of Christian churches portray the main figures and events in the history of salvation, legends of the saints, and the Last Judgment. Inside the holy place, this pictorial and interpretative function is continued in the figures and scenes on its walls, capitals, and vaults. The adytum (sanctuary), the apses, and the altar may be decorated with symbols or pictures of the divinity or of other gods and saints.

Icons and images. Pictures are the main subject matter of iconography, which also includes free-standing sculptured forms and reliefs. Free-standing figures or statues are important in ritual as well as in partly serving magical purposes, which cannot always be separated from religious ritual. Such figures, which later became objects of personal devotion and meditation, include representations of the gods and demons in various prehistoric religions and of Buddha, Christ, and the various Buddhist and Christian saints. Generally, Judaism, Islām, and ancient Shintō have rejected any representation of the divine.

Pictorial representations of myths and historical events

Painted or sculptured tableaus of historical or mythical events originally belonged in a ritual setting. The function of a wall painting, wall or floor mosaic, or relief was or is to establish the ritual actions as authentic re-enactments of their mythical or historical prototype and to make these mythical or historical events continually present. These tableaus also may be found on the interiors and sometimes the exteriors of houses and on cemetery monuments. They are made to serve private devotion and a personal confession of faith. In the form of a framed picture, Oriental roll picture, print, or book illustration, such an iconographic tableau contains religious information, mediates, and stimulates contemplation and devotion.

Iconographic themes. In the religions of highly developed cultures and in the universal religions, complicated systems of iconography have been developed. In the course of time, however, these systems have been subject to change. Icons (images) may depict the divine in its oneness and in the plurality of its differentiations, emanations, and incarnations, as well as man in his various relationships to the sphere of the holy. They may also depict the world as the stage of divine action, as the realm of the diabolical, or as the battleground of these two warring forces. They may portray evil, the diabolical, and the Satanic (the negatively sacred); or, more positively, they may depict the offer of salvation, redemption, and damnation. Furthermore, icons may portray the ritual means of attaining salvation or moral relationships and duties. Icons may borrow from myths and other religious narrative material to depict the historical past and the present, as well as the future and the afterlife. Icons, finally, may represent religious doctrine and the theological treatment of dogmatic themes, as well as other religious beliefs, religious experiences, and conceptions of a more individualistic nature.

Iconographic types. There are many fundamentally different points of departure in the ways of conceiving the contents of religious pictures and of forming them. These differences, which go back to very early times, continue to exist side by side throughout the history of religions, some dominating at one time while others recede in importance.

Anthropomorphic motifs. The object that generally is depicted in religious pictures or sculpture is an anthropomorphic (human-form) representation. Man is shown as the image and likeness of the holy and as engaging in typically religious behaviour; conversely, the divine appears with anthropomorphic characteristics. This tendency is found quite early in the history of religions. Examples include the religious pictures used in ancestor worship; the spirit and soul idols of various primitive cultures in animism; the fetish, or charm, figures of West African fetishism; and the magical objects of hunter and agrarian cultures. This type of anthropomorphism reaches its high point in the ritual and mythical pictures of the great polytheistic religions and is especially characteristic of ancient Greek religion and also of Jainism in its pictures of the Tīrthaṅkaras (saviours).

Importance of human forms in representations of the holy

In universal religions, such as Buddhism and Christianity, anthropomorphic pictures of the divine were maintained despite criticism. They were not intended to be interpreted realistically but rather as symbolically representing the divine. Buddhism adapted the gods and anthropomorphic myths of the then popular Asian religions and developed the figure of the *bodhisattva* (buddha-to-be) to represent the attainment of Nirvāṇa (the state of extinction or bliss). In Christianity, the picture of Christ usually serves as a representation of the divine. God the Father also is anthropomorphically depicted, usually as an old man wearing papal or imperial insignia. Individual parts of the body may be depicted and serve as symbols of the divine: *e.g.*, the hand of God may stand for Christ, the creative power of God, God's covenant with man, or for God's fidelity and truth; the phallus or foot may symbolize Śiva (a Hindu god). Man may be portrayed as a miniature copy of the universe or as the recipient of salvation and also the bearer of the divine, as in the Christian iconography of Mary and the saints.

Theriomorphic, or zoomorphic, motifs. Besides animal demons in primitive religions and totemism (a belief system and social system based on animal symbolism), animal images frequently occur in other more sophisticated religions. The animal form as a representation of the divine (theriomorphism, or zoomorphism) is characteristic of polytheism. It has been maintained in Hinduism, to some extent in Buddhism, and occasionally in Christianity. Besides the theriomorphic (animal-form) representations of the holy (*e.g.*, the ancient Egyptian gods and animals that are symbols of the divine or the lamb symbolizing Christ in Christianity), there are also theriomorphic (animal-form) pictures of the universe and its powers and of the world of the demons. In many religions the animal kingdom is depicted as a part of creation, as in the portrayals of creation in ancient Greek myths and in

the Bible. Animals also play important roles in allegories. Various forms of the shepherd–flock motif have been developed to describe God's relationship to man.

Besides being represented in human form, the Christian Evangelists Mark, Luke, and John are symbolically depicted in animal form (lion, ox, and eagle, respectively). Byzantine iconography sometimes depicts St. Christopher (patron of travellers) with a dog's head. Parts of animals (skulls, horns, wings, and feet) also serve as symbols of the power of the divine or diabolical.

Phytomorphic motifs. Phytomorphic, or plant-form, representations of the divine also are rich in diverse examples and often enigmatic. Holy plants and plants considered to be divine are represented in connection with gods in human form. The god sometimes is the plant itself, as the Egyptian god Nefertum is the lotus, or begets the plant, as the Egyptian Osiris or the Greek Demeter as deities of corn, or the deity comes forth from the plant, as the Egyptian goddess Hathor from the sycamore or the *bodhisattva* from the lotus, or the god unites with or is transformed into the plant, as the Greek heroine Daphne changed into the laurel tree, which thus became sacred to Apollo. The genealogy of Christ from "the root of Jesse," the father of the Israelite king David, is represented as a tree the last blossom of which is Christ. The biblical story of creation describes the vegetative surroundings of man and his dependence on plants (*e.g.*, the tree of knowledge). The tree of life, the world tree, and the primeval cosmic plant all have characteristics related to the nature and origin of the cosmos.

Plants as representations of the holy

The grapevine is a prominent ritual motif. It is found, for example, in representations of Dionysus and Christ. Painted and sculptured leaf, flower, and plant motifs decorate Christian churches and many religious and funeral monuments. Plants bound into a wreath symbolically promise victory over death and the joys of heaven. In such instances, the simple forms of nature may sometimes be depicted in a nonrepresentational and ultimately abstract and stylized manner.

Hybrid motifs. In religious iconography, anthropomorphic, theriomorphic, and phytomorphic motifs may be combined. The result of this fusion of forms may be seen in the numerous hybrid figures of primitive culture (*e.g.*, totem poles, Uli figures of New Ireland, and ancestral tablets). Such combined motifs occur also in ancient Near Eastern figures of winged demons with human heads and animal bodies or in winged beings with animal heads and human bodies and in the winged Greek goddesses, as well as in the winged protectresses of the dead in ancient Egypt and the angels and demons in Christian art. In Christianity, the snake in the Garden of Eden is sometimes portrayed with a human head (the face of Satan). In the Middle Ages, representations of the living cross with its arms depicted as hands appear. The cross also has been combined with various other anthropomorphic and phytomorphic elements.

A composite picture of plants, animals, and men together with other natural objects and architectural structures often becomes a sacred scenic background against which the mythical and ritual action takes place. Such scenic depictions were developed in Hellenism and adopted by early Christianity. Paradise scenes including plants, animals, men, Christ, and the saints are later enriched by symbolic and diagrammatic elements. Renaissance painting and East Asian Buddhist and Taoist art also use such combinations when depicting sacred, mythological, and allegorical scenes.

Chrematomorphic motifs. Objects that are used, or chrematomorphic objects, provide another form of pictorial representation. Holy objects, especially those used in worship, fall in this category. The holy book, the cross, the throne and other insignia of power and majesty, lights, lamps, and canopies become representatives of the holy. Garments also may have a symbolic meaning of their own apart from their wearer, as, for example, the veil or the blue mantle of Mary as symbols for the tent of heaven.

Absence of representational forms. The absence of an expected object, person, plant, or animal in a picture or the absence of all pictorial representation may also represent the holy or divine. In the Holy of Holies of the Jewish Temple in Jerusalem there was no picture of Yahweh in or on the ark of the Covenant, although it was supposed to be a sort of portable throne for God. Ancient Christian art often depicted an empty throne on which perhaps lay a folded purple robe or a book (*hetoimasia*) as a symbol of the invisible presence of God. In mosques the empty prayer niche (*miḥrāb*), which is oriented toward Mecca, represents the presence of Allāh. Buddha apparently was not iconically represented in early Buddhist art in accordance with the theory of "emptiness" (*śūnya*) and the radically negative transcendence of the aim of salvation, Nirvāṇa. The rejection of a picture as a means of representing the holy also is a symbolical way of positively asserting the presence of God.

Depiction of the presence of the holy by nonrepresentation

Hostility toward and prohibition of pictures are found in ancient Shintō, Judaism, Islām, the various radical movements (*i.e.*, the iconoclasts, or image destroyers) of 8th-century Christianity that were influenced by Islām, and, centuries later, in some elements of Reformed Protestantism.

INFLUENCE OF MAN'S ENVIRONMENT ON RELIGIOUS SYMBOLISM AND ICONOGRAPHY

Influences from nature. The main streams of the influence from nature are derived from man's experience of nature itself, his position in the universe, and his attempt to master his world in religious terms. Man's sense of the holy influences the way he perceives and understands nature. The space that surrounds man provides him with the dimensional coordinates of his religious experience. Height, depth, breadth, direction, proximity, and distance are the spatial forms in which the holy manifests itself. The holy may reside on a mountaintop, in heaven, in a chasm, in the underworld, in watery depths, or in a desert. The holy way or path provides man with his direction to the divine and a means of approaching it. The spatial position of the holy and the direction to it may also be abstractly expressed—*e.g.*, by means of symbolical numbers or coordinates. The infinity of space may be represented by geometrical and linear figures.

Emptiness or fullness may characterize the utilization of spaces and surfaces that are usually intended for the reception of symbols and signs. Works of art may be totally absent in certain architectural structures; or all available space may be filled with a dense profusion of all kinds of figures and objects, all of which may sometimes be encircled by an ornamental network or web of branches, vines, leaves, and blossoms; an example of such embellishment is Islāmic art. The ebb and flow of time and things, the flow of water, and the cyclic recurrence of time are pictorially expressed—in symbols such as the wheel, spiral, wave, and circle. Time appears as the god of destiny—*kāla* in ancient Indian and Zurvān in ancient Persian religions. In late antiquity time takes the form of a demon entwined with snakes (Aion). The snake biting its own tail, the ring, and the spiral are frequently recurring symbols of fate and eternity; in Christianity, eternity is represented by the A and Ω, and the wreath.

Other physical, chemical, and physiological facts of nature also serve as sources of symbolic and iconographic concepts. Examples include the experiences of seeing, hearing, tasting, smelling, and touching; the myriad forms of plant and animal life; heaven and its astral and meteorological phenomena, which may be represented realistically or abstractly through symbols or personifications; and the colours and various colourful natural occurrences such as the rainbow (often symbolizing God or Christ) or the sunrise and sunset or minerals and precious stones. In the depiction of natural happenings, plant, animal, and human forms may blend into one another, as in the symbolic circle of the Mesopotamian god Tammuz in which the tree of life is combined with figures of fertile wild and domestic animals and the figure of a shepherd. All these symbols represent the preservation and regeneration of life. They also represent nature as a holy power.

Another area of nature symbolism is that of the micro-

cosm and macrocosm of heaven and Earth. Heaven and Earth are depicted as a dually or polarly related pair, which generally are theistically personified as a man and a woman. These roles may sometimes be reversed, as in ancient Egypt: the heavenly divinity a goddess, Nut, and the Earth divinity a god, Geb. In Greek myths on the origin of the gods (theogony), the world of the gods and man results from such pairing. Mother Earth is a central figure of many myths: she is the mistress of fertility and death.

The importance of social relationships in concepts of the holy

Influence of human relationships. Another group of pictures and symbols that are especially significant in depicting the relationship of God and man are those drawn from the area of family and social relationships, especially the roles of the father and mother. These relationships to some extent are determined by the structure of the society and its economy. The mother image is closely bound up with Earth symbolism, vegetation, agriculture, fertility, the reappearance of life, and the lunar cycle. The father image usually is associated with the sphere of heaven, authority, dominion, age, wisdom, and struggle. Love, betrothal, marriage, sexual union, family, and friendship also are significant in symbolization. The relationship between brothers and sisters is of importance, especially in the structure of religious communities, as in the totemistic organization of primitive societies and in the various fraternal groups and secret organizations of modern societies.

The image of the child, the subject, or slave again indicate man's relationship to God; the image of the ruler, king, or master expresses the power and authority of the deity. Even the structure of the world of the gods is explained in terms of family (*e.g.*, father, mother, and child as a trinity or divine twins).

Symbolism of sex and the life cycle. The symbols of sexuality and the life cycle perform a function similar to those of time and eternity in the higher religions. They indicate the permanence of the cycle of sexual functions and the return and renewal of individual and collective physical life that those functions bring about. The endless renewal of life is variously represented. It may be as realistic depictions or diagrammatic and stylized abbreviations of man and woman, god and goddess, masculine and feminine animals in the act of love and sexual union, as in reliefs on Hindu temples, or as depictions of sex characteristics (*e.g.*, in Indian *liṅga-yoni* symbolism). The theme of renewal also may be depicted in representations of woman with emphasis on her function as mother, as in the nursing-mother figures of ancient Greece. The life cycle also is represented by figures portraying the different ages of man or by depictions of pain and suffering, as in pictures of Buddha's death, which also indicate his breaking out of the endless chain of existence.

Cultural influences. Other cultural, political, social, and economic institutions and conventions also influence religious symbolism and iconography. Work and leisure, war and peace, and the myriad things associated with them—occupations, positions in society, classes and their functions, the tools of domestic and professional life, technical equipment, forms of international relations and strife—all play an important part in man's interpretation and understanding of religious reality and hence in his symbolization of this experience. Hunters, farmers, shepherds, artisans, and merchants and their activities are represented in religious pictures and appear in the verbal symbolism of religion. Tools and weapons become expressions of religious activity. In the universal and missionary religions, such as Buddhism, Christianity, and Islām, the believer is summoned to take up spiritual arms and fight for salvation. In Judaism, Christianity, and the religion of ancient Rome, the relationship between God and man is regulated according to the model of a peace treaty. In ancient German and Indian religions the military virtues of loyalty, duty, and comradeship are stressed. Man's religious activities may also be expressed in terms of play and sport, training, competition, and victory.

Conceptual influences. Ideas, theories, and structured systems of thought also are incorporated into religious symbolism. Abstract ideas—such as wholeness, unity, and the absolute—and the power of the spirit are concretely expressed in religious terms. The idea of unity plays an important part in expressing the oneness of the divinity. Mathematical principles expressed in number symbolisms are used to organize the world of the gods, spirits, and demons, to describe the inner structure of man, and to systematize mythology and theology. The concepts of duality or polarity find expression as the body and soul of man: the divine pair; the syzygy (paired emanations) in Gnosticism; the dualism of God and the devil, of the good and evil principles, of the ethically good and evil; and, finally, as the two natures of Christ. The number three, or triplicity, is represented in divine triads, the trinity, and the body–soul–spirit structure of man; as is number four, or quaternity, in the four cardinal points, the picture of the cosmic whole, the divine quaternity. Other higher number groups play similar roles. Time and eternity may be expressed in abstract symbolical terms as well as concretely in picture form.

Symbolism of numerical concepts

INFLUENCE OF RELIGION ON SYMBOLISM AND ICONOGRAPHY

Symbolism of religious figures and spiritual authorities. Religious figures and spiritual authorities themselves form a vast complex of symbols: gods, saviours, redeemers, heroes, the avatars (incarnations) and the Īśvaras (manifestations) of Hinduism, the heroes and gods of epics, the great founders, lawgivers, saints, and reformers of the great religions. The Christian saints are characterized by a very complicated system of symbols. Theologians, mystics, and contemplatives may also be symbolically and pictorially represented; the doctors (teachers) of Roman Catholicism and Eastern Orthodoxy and fathers of the early church have standard iconic forms, attributes, and symbols (*e.g.*, St. Augustine is represented by the heart; St. Jerome by the lion). Persons connected with ritual and representatives of the religious institution (*e.g.*, hierarchs, priests, assistants in the liturgy, male and female dancers, and musicians) may also be symbolically and iconographically depicted. Such functionaries—together with the major figures of a religion, such as a prophet or founder—represent the idea of mediation and thus have an important iconographic function in the artistic decoration of a place of worship.

The offering. The offering, the place of offering, the altar and its trappings, the instruments that prepare and destroy the offering, the fire that consumes it, the liquids and drinks used in the rite, the sacred meal, and the rites of communion all are objects of iconography and symbolism. The offering symbolizes the idea of submission to the ideals of a religion, the giving up of valuables and possessions for religious purposes and for the service of human brotherhood, and the giving up of one's life for religion.

Organized religions. A religious community recognizes itself and its ideas by symbols. Examples are the *yin-yang* (union of opposites) symbol bound by the circle of stability (*t'ai-chi*) in Chinese universalism; the swastika in Hinduism and Jainism; the wheel of the law in Buddhism; the *khaṇḍā* (two swords, dagger, and disk) in Sikhism; the star of David or *menora* (candelabrum) in Judaism; and the cross in its various forms in Christianity.

CONCLUSION

The further development of symbolism and iconography in the higher religions of the modern world that maintain complicated theoretical and doctrinal systems is an open question. In the years between 1920 and 1960 in the Christian communities, revivals of the liturgical traditions and of ritual symbolism were in progress, though criticized vigorously by many theologians. The value of symbols and pictures, mainly in worship, however, found wide recognition. Liturgical symbolism became valued anew and stabilized. Theological systems, like that developed by Paul Tillich, have been based on the concept of the symbol. On the other hand, during the 1960s some indifference toward symbols and pictures developed be-

cause of an emphasis on the moral and social tasks of religion. Symbols, myths, pictures, and anthropomorphic ideas of God in the terms of space and time were rejected by many theologians and philosophical structures of thinking (*e.g.*, the theory of the demythologization of the Bible; or, the "God-is-dead" theology) became substitutes for them. In the great non-Christian religions, this process seems to be less acute. In the view of a more rationalistic and technological thinking within the horizons of a secularized, sceptic, and agnostic society, however, religious symbols seem to be dispensable. In spite of these developments and attitudes, a new and increasing interest in symbols is appearing, especially among the younger generations of the latter part of the 20th century who have come into contact with both Eastern and Western religious and cultural traditions with their rich sources of symbolic images and modes of thinking. Even in the field of politics and social organization, new symbols have been arising. Thus, a resurgence of an understanding for the specific values of symbolism and iconography has been recognized in the latter part of the 20th century, in spite of all apparently opposite trends.

BIBLIOGRAPHY

General works: Bibliographie zur Symbolik, Ikonographie und Mythologie, ed. by MANFRED LURKER (1968–), contains fundamental and comprehensive bibliographies; *Symbolik der Religionen*, ed. by FERDINAND HERRMANN (1958–), a series of compendious and important monographs; J.E. CIRLOT, *Diccionario de Símbolos Tradicionales* (1958; Eng. trans., *A Dictionary of Symbols*, 1962), contains an excellent bibliography.

The idea and nature of symbols, symbolization, and culture: HAROLD BAYLEY, *The Lost Language of Symbolism: An Inquiry into the Origin of Certain Letters, Words, Names, Fairy-Tales, Folklore, and Mythologies*, 2 vol. (1951); ERNST CASSIRER, *Philosophie der symbolischen Formen*, 3 vol. (1923–29; Eng. trans., *The Philosophy of Symbolic Forms*, 3 vol., 1953–57), a pioneer work; HUGH DALZIEL DUNCAN, *Symbols in Society* (1968), a symbolic analysis of society and its expressions in reciprocal relations; MIRCEA ELIADE, *Images et symboles* (1949; Eng. trans., *Images and Symbols*, 1961); ERICH FROMM, *The Forgotten Language* (1957), contains psychoanalytical aspects; DESIREE HIRST, *Hidden Riches: Traditional Symbolism from the Renaissance to Blake* (1964); HANS JENSEN, *Sign, Symbol and Script*, 3rd rev. ed. (1969); FREDERICK ERNEST JOHNSON (ed.), *Religious Symbolism* (1955); CARL GUSTAV JUNG, *Psyche and Symbol*, ed. by VIOLET S. DE LASZLO (1958), and *Man and His Symbols* (1964); GYORGY KEPES (ed.), *Sign, Image and Symbol* (1966); SUSANNE K. LANGER, *Philosophy in a New Key: A Study in the Symbolism of Reason, Rite, and Art*, 3rd ed. (1957); LUCIEN LEVY-BRUHL, *L'experience mystique et les symboles chez les primitives* (1938); HANS LOOFF, *Der Symbolbegriff in der neueren Religionsphilosophie und Theologie* (1955), survey and critique of the recent discussion of symbolism from Goethe to Tillich, Guardini, and Jaspers; ROLLO MAY (ed.), *Symbolism in Religion and Literature* (1960); RUDOLF OTTO, *Das Heilege*, 9th ed. (1922; Eng. trans., *The Idea of the Holy*, 2nd ed., 1950); JOSEPH R. ROYCE *et al.*, *Psychology and the Symbol* (1967); THEODORE THASS-THEINEMANN, *Symbolic Behavior* (1968), a psychological understanding of symbolic language; PAUL TILLICH, "Existential Analyses and Religious Symbols," in HAROLD A. BASILIUS (ed.), *Contemporary Problems in Religion*, pp. 35–55 (1956); JOACHIM WACH, *The Comparative Study of Religions* (1958); HEINZ WERNER and BERNARD KAPLAN, *Symbol Formation; An Organismic-Developmental Approach to Language and the Expression of Thought* (1963); ALFRED NORTH WHITEHEAD, *Symbolism; Its Meaning and Effect* (1955); EDWARD C. WHITMONT, *The Symbolic Quest: Basic Concepts of Analytical Psychology* (1969).

Modes and derivation of symbolic and iconic expression and principles—recurrent themes in history of symbolism and iconography: LEROY H. APPLETON and STEPHEN BRIDGES, *Symbolism in Liturgical Art* (1959); JITENDRA NATH BANERJEA, *The Development of Hindu Iconography*, 2nd ed. rev. (1956); B. BHATTACHARYA, *Indian Buddhist Iconography*, 2nd ed. rev. and enl. (1958); E. DOUGLAS VAN BUREN, *Symbols of the Gods in Mesopotamian Art* (1945); MAURICE FARBRIDGE, *Studies in Biblical and Semitic Symbolism*, rev. ed. (1969); GEORGE FERGUSON, *Signs and Symbols in Christian Art* (1961), handy survey with many illustrations from paintings of the Renaissance; ANTOINETTE K. GORDON, *The Iconography of Tibetan Lamaism*, 2nd ed. (1959, reprinted 1967); DONALD E. MACKENZIE, *The Migration of Symbols and Their Relations to Beliefs and Customs* (1926, reprinted 1968); CHARLES PERCY MOUNTFORD, *Art, Myth and Symbolism* (1956); JOSEPH M. KITAGAWA and CHARLES LONG (eds.), *Myths and Symbols* (1969), collection of various general and special contributions; DORA and ERWIN PANOFSKY, *Pandora's Box: The Changing Aspects of a Mythical Symbol*, 2nd rev. ed. (1962); H. DANIEL SMITH, K.K.A. VENKATACHARI, and V. GANAPATHI, *A Source Book of Vaisnava Iconography According to Pancaratragama Texts* (1969); CHARLES ALFRED SPEED WILLIAMS, *Encyclopedia of Chinese Symbolism and Art Motives* (1960); HEINRICH ZIMMER, *Myths and Symbols in Indian Art and Civilization* (1946).

(K.M.A.G.)

Symphony

The term symphony (literally, "sounding together"), though it has many meanings in musical literature, is commonly understood to refer to a lengthy composition for orchestra, often employing a sonata form (see below) in one or more large sections called "movements." Symphonies of this sort began to be composed during the so-called Classical era in European music history, *c.* 1740–1820 (the early part of this period and the decade immediately preceding it are sometimes called pre-Classical, as are the symphonies written before about 1750). During the 19th century, which included the Romantic era, symphonies grew longer, and composers concerned themselves with ways of unifying the movements; extramusical programs and new approaches toward tonality (the major–minor system of chord progressions) were among the solutions to the problems of large-scale symphonic form. Late in the century, symphonies—and orchestras—had grown to such an extent that reaction set in, culminating in the Neoclassic movement of the early 20th century, in which composers turned again toward principles of balance and formal discipline, using new techniques to achieve dynamic coherence. Economic considerations forced a reduction in the size of orchestras and amount of rehearsal time available to mid-20th-century composers, further justifying a return to less extravagant symphonic thinking. But throughout the 19th century a number of outstanding symphonists were able to reconcile the demands of fashion with strict musical logic; these composers represent the mainstream of symphonic activity, and their works remain models for much 20th-century activity in the genre. Throughout the following article two concerns predominate: a survey of the chief symphonic works and composers and consideration of the evolution of symphonic thought.

THE CONCEPT OF SYMPHONY BEFORE C. 1750

Early works entitled symphony. The word *symphōnia* was used by the Greeks in reference to notes sounding together in harmony and by extension meant an "ensemble" or "band" rather than a musical form. The word implies a pleasant concord of different notes and has been used in fields other than music to denote a pleasing combination of various elements. In the New Testament Gospel According to Luke (King James Version), *symphōnia* is translated as "musick," as distinct from *choroi*, "dancing." In the Middle Ages the name was given to several musical instruments, among them a double-headed drum, bowed stringed instruments, a large hurdy-gurdy, and bagpipes. Mention is made in 1582 of *eine Symphonei*, evidently a stringed keyboard instrument.

Vocal and instrumental works

From the mid-16th century, symphonia (and related spellings) is a term often found in titles in which it simply indicated ensemble music, whether for instruments with voices or either alone. A collection of madrigals published in Antwerp in 1585 is entitled *Symphonia angelica . . . raccolta per Huberto Waelrant*. Later notable examples are the *Sacrae symphoniae* of the Venetian composer Giovanni Gabrieli (Book I, 1597; Book II, 1615), collections of elaborate vocal and instrumental music, often for multiple choirs; and the *Symphoniae sacrae* of his celebrated German pupil, Heinrich Schütz (1629, 1647, 1650). Schütz's collection reveals his debt to the colourful and brilliantly orchestrated Italian style in works ranging from several voices to large polychoral compositions with solo parts and instruments. His countryman Samuel Scheidt's 70 *Symphonien auff Concerten-Manir*

(1644) likewise combine instrumental and vocal ensembles to enrich the texture and heighten the drama of his music.

Symphonies for instruments alone during the early Baroque era (*c.* 1600–30) occur as independent pieces and as introductions or interludes in theatrical productions. The Italian Biagio Marini's Sinfonia *"La Orlandia"* (1617) is a duet for violin or cornetto (a wind instrument with finger holes and cup-shaped mouthpiece) and continuo in five brief contiguous sections, distinguished by contrasting metres and new melodic material in each section. (The continuo is a harmonic accompaniment improvised over the written bass line, usually played on a keyboard instrument and a bass viol or other bass melody instrument.) Early operas often include instrumental symphonies. Jacopo Peri's *Euridice* (first performed 1600) includes a sinfonia for three flutes; Claudio Monteverdi's lavish musical drama *Orfeo* (1607) is punctuated with five richly scored sinfonias, while a *sinfonia da guerra* ("sinfonia of war") accompanies a staged battle in his *Il ritorno d'Ulisse in patria* (*The Return of Ulysses to his Country;* 1641). Each act of Stefano Landi's opera *Il Sant' Alessio* (1632) opens with a sectional sinfonia. Many other opera and oratorio composers used short descriptive or introductory sinfonias, often of sectional form with contrasting metres and tempos.

Formal precursors of the Classical symphony. *The Italian overture.* It remained for a Neapolitan, Alessandro Scarlatti (1660–1725), to formalize the overture to his operas as a fast–slow–fast *sinfonia avanti l'opera*, as in his opera *Dal male il bene* (*Good from Evil;* 1681). The so-called Italian overture of this and later works, scored for strings and continuo, has been widely considered to contain the germ of the later three-movement symphony. In contrast with the more contrapuntal French overture, which begins with a pompous slow movement and continues in a fugal section (involving imitation of a melody among several voices), the Italian style is immediately tuneful and predominantly homophonic (chordal) in texture. The first fast movement may be trivial; its symmetrical phrasing is unexpressive. The contrasting second movement may be more lyrical, perhaps anticipating tunes heard later in the opera. The last movement, sometimes a minuet, is an exuberant curtain raiser. This format spread quickly outside Italy, even to France. Jean-Philippe Rameau's *Zoroastre* (1749), for example, includes such a fast–slow–fast overture. Rameau, indeed, was considered an exponent of the Italian style, particularly in his lucid harmonic treatment. This late-Baroque concern with tonal clarity prefigured the attitudes of early Classical symphonists. Among the devices used to assure clarity are melodies constructed of arpeggiated ("harplike," or broken) chords and passages in unison or in parallel thirds or sixths (sequences of harmonies formed by thirds, such as C–E or D–F, or sixths, such as C–A or D–B). These features are not common in Baroque music that is strictly contrapuntal (based on interwoven melodic lines) in texture.

Characteristics of Italian and French overtures

Evolving approaches to other forms. While the opera overture settled into a form that eventually inspired early symphonists, the term sinfonia, or symphony, as yet had no formal definition. As late as 1771 *Encyclopædia Britannica*, reflecting ancient Greek usage, defined symphony merely as " . . . a consonance or concert of several sounds agreeable to the ear, whether vocal or instrumental, called also harmony." Sinfonia was used interchangeably with concerto, consort, overture, suite, and so on. Commonly a brief instrumental interlude, as in a song, was called a symphony, even into the 19th century. In the late Baroque era (*c.* 1700–50) the term was applied to such dissimilar pieces as J.S. Bach's didactic *Three-Part Inventions* for keyboard, called *Sinfonien* in the 1723 copy, and the orchestral "Pastoral Symphony," a quasi-descriptive interlude in Handel's *Messiah* (composed 1741), said to have been based on an Italian shepherd bagpipe tune and very much in the tradition of earlier descriptive symphonies in opera.

Bach's *Sinfonia VII* in E minor and *Sinfonia XI* in G minor are interesting in that in each piece the opening material recurs at the end. In *VII* this repetition is merely suggested, but in *XI* the last eight measures of the piece virtually duplicate the first eight. The whole intermediate body of these pieces develops the motivic material presented at the beginning, and the initial material is transformed contrapuntally and harmonically. In the closing bars the tension thus aroused resolves and the rhythmic drive reins in. This suggestion of an expository unit moving from the home key to a different key, followed by an extended development that explores still more remote keys and the motivic and contrapuntal implications of the beginning, concluding with a recapitulation in which the energy of the development is somewhat dissipated by a return to the opening material, prefigures the sonata form of the Classical symphonists (see below). Bach uses this technique in some of his instrumental concerto movements; the concertos have other elements in common with early symphonies, especially in the mood of their lyric slow movements and fast duple-metre finales.

Exploitation of key relationships

The word sinfonia was applied to a trio sonata for flute, oboe, and continuo in Johann Joseph Fux's *Concentus Musico-instrumentalis* (1701), a collection of suites each comprising a number (as many as 15) of bipartite (two-section) dances and descriptive pieces. An intellectual and influential Viennese court composer, Fux departed in this sinfonia from the typical 17th-century suite, which is merely a collection of contrasting dances in the same key. The work falls into two major divisions, both comprising three short movements; the key scheme is F major, D minor, F major—F major, D minor, F major, and the last three movements have programmatic titles. Here is not merely a collection of various dances but a conscious attempt to relate movements tonally and thereby create larger hierarchic units. F major and D minor are closely related keys, and it would not be possible to omit a single movement without destroying the symmetry of the whole (not that either group of three, or even each dance, does not sound good by itself). By means of this simple, balanced harmonic structuring Fux advanced beyond the looser architecture of the typical suite; and by framing a minor-key movement between two movements in the same related major key he anticipated the overall form of many early symphonies.

Importance of tonal harmony

Both Fux and Bach were products of the evolution of tonal harmony, a system of key relations which brought with it the possibility of basing large-scale forms not only on melodic variation or counterpoint, as earlier, but on harmonic tension and modulation. (Modulation, unlike simple change of key, implies the establishment of a new tonic, or tonal centre, by means of progression through a number of related keys.) The wide-ranging modulations and affective harmonic progressions of German Baroque composers depended on well-tempered tuning, a system that permits exploration of keys distant from the tonic without the necessity of retuning to accommodate the remote harmonies. Bach exploited this system to the utmost, as did many of his North German contemporaries; but their rich harmonic palette was foreign to the south, where many important symphonists arose. Concerned less with powerful emotions (*Affekten*) and more with clarity, the southerners avoided intricate counterpoint and convoluted harmonic progressions, preferring a restricted chord vocabulary and clear-cut symmetrical phrasing dominated by tuneful melody.

The Neapolitan intermezzo. Besides the suite and opera overture, the short humorous intermezzo, which originated in Naples and flourished *c.* 1685–1750, strongly influenced pre-Classical symphonists. Neapolitan composers, headed by Alessandro Scarlatti, concerned themselves in the intermezzo with dramatic, comic interplay between two singers in two or three short acts made up of arias, recitatives, and duets. Because the texts demanded clear articulation and careful declamation, they influenced the melodic phrase structure, giving rise to repeated-note figures and brief rhythmic or melodic motives. These phrases normally fall into two-measure units. Counterpoint was abandoned, for it tended to obscure the text; and harmonies became simple and slow-moving. Intermezzo melodies abound in ornaments, sudden ac-

cents, syncopation (displaced accents), and playful leaps reflecting the text declamation and lack the broad, spun-out arch and driving rhythm of typical Baroque melodies. Rather, they are made up of short motives joined one to another and give rise to frequently articulated phrase groups. This word-derived idiom furnished the melodic impulse of the early symphonies.

THE SYMPHONY PROPER

Early Classical period. *Italy.* Chord-generated melodies (those arising from arpeggiated triads) abound in 18th-century symphonies, among which a number of stereotyped "theme families" can be distinguished. These furnished raw material for further development. In fact, a composer's originality found expression not so much in his original theme as in his realization of the implications of the theme later in the composition. Certain tunes are by nature not highly implicative; they are perfect closed units that cannot be easily developed and so are superficially inappropriate for symphonic use. Such, for example, are many folk tunes; and this explains why great symphonists rarely use folk tunes without at least distorting them so as to open them to development. On the other hand, motivic melodies, such as those in the intermezzi, coupled with slow harmonic motion, lend themselves well to fragmentation, recombination, extension, elision, reharmonization, and other developmental techniques. By the 1740s, Italian symphonists had learned to sustain interest by these means and to obtain contrast by dramatically apposing tunes of different character in different keys (usually the tonic, or home key, and the dominant, located five tones above the tonic, or related major or minor keys).

Early uses of sonata form

Among Italians influenced by these factors was Tomaso Albinoni (significantly, a composer of 48 operas). The third movement of his *Sinfonia in D Major*, fifth of the *Sei sinfonie a quattro* (1735), displays a simple sonata form (also known as "first movement form," though not so limited). Sonata form, crucial in the symphony's evolution, is based on the dramatic apposition and eventual reconciliation of contrasting keys. In essence it consists of an exposition in which one or more themes are presented, the first (often forceful in character) in the tonic key and the second (often lyric) in the dominant. Sometimes a third, closing theme follows in the dominant. The two key areas contrast not only harmonically and melodically but often in instrumentation, loudness, and texture. So-called monothematic sonata movements lack a contrasting second melody; indeed, it is not so much the character of the tunes but the dynamism of the opposed key areas that is essential to the dramatic structure. The transition between tonic and dominant areas was to become a focus of interest to later composers, but in early symphonies the transitions were brief and simple.

The exposition, often marked to be repeated, comes to a close on a key other than the tonic (usually the dominant) and is followed by a development section, beginning on the dominant, in which themes previously heard are reharmonized, fragmented, or otherwise reshaped. Again it is not melodies so much as harmonies that arouse tension in the development. The composer confronts the problem of returning to the tonic via more distant chords, and this is sometimes accomplished by modulations that bear no thematic relation to the exposition. In early symphonies this process is only tentatively exploited, and developments are brief, sometimes involving merely transposition of the original first theme to a new key. Later the development assumed the character of the meat in the sandwich, as it were.

Following the development comes a recapitulation of the exposition, this time all in the tonic key (before *c.* 1770 the recapitulation sometimes retained the key scheme of the exposition, except for the closing bars), resolving the harmonic tension of the development. The recapitulation may be simply a virtual repetition of the exposition, with appropriate key changes, or may be truncated, expanded, or otherwise varied so as to continue developmental processes. Many early symphonies take advantage of the implications of a varied recapitulation, literal repetition being abhorrent to imaginative composers. It will be observed that when both themes appear recapitulated in the tonic, the function of the transition between themes differs from that in the exposition, in which it leads from one key to another. After the 1750s, however, the first theme was often omitted in the recapitulation. Obviously, the recapitulation's tonal scheme allows extended treatment of the tonic, but sometimes a coda (tail) is added after the recapitulation to consolidate further the focal nature of the tonic.

It is important to understand that sonata form was subject to great variation and deviation from "textbook" norms throughout the symphony's history. Ideas of a "typical" sonata form did not evolve until *c.* 1830. Nevertheless, the vocabulary of exposition, development, recapitulation, transition, and so on is capable of wide application and will simplify the remainder of the present article.

Albinoni's *Sinfonia in D Major* as a whole is forward-looking. The orchestra lacks a continuo, harmonic orientation replaces counterpoint, and the movements are of larger scale than typical Baroque dance movements. The central minuet-and-trio movement, surprisingly modern-sounding, contrasts lightheartedly with the sonata form. Minuets and song-form movements relieve the dramatic tension of sonata movements; and from this time on minuets appeared frequently in symphonies. Gradually losing dance character, they moved from the last or middle position to penultimate place when, later, a fast fourth movement appeared as finale.

The symphonies of Sammartini

A leading early symphonist, the Italian Giovanni Battista Sammartini, is known from some 77 extant symphonies, some of them available in modern editions. A prolific composer of instrumental chamber music, his use of incipient sonata form, restricted harmonic vocabulary, and motivic, not highly ornamented melodies pointed to the future. Though dependent on Baroque models for inspiration during his first period of activity (to *c.* 1740), he had abandoned the continuo by about 1760, filling the harmony instead with horns or trumpets—which, however, were restricted in their ability to play in certain keys and therefore operated to constrict the available harmonic vocabulary. More than many contemporaries, Sammartini infused inner parts with contrapuntal life, especially in earlier works, although he rarely used extended imitation. His rhythmic energy was appreciated by opera-writing contemporaries, from whom (especially the Italians Niccolò Jommelli and Rinaldo di Capua) he perhaps drew inspiration for the long, expressive cantabile (singing) phrases that, along with a firm grasp of modulation, characterize his later works. His slow movements are especially rich. Another late characteristic is the extended role of wind instruments, which are given independent, idiomatic parts. Sammartini wrote for oboes, flutes, and bassoons among the woodwinds (bassoons doubling the bass line); clarinets were also in use at this time, usually interchangeably with oboes. Significantly, Mozart heard Sammartini's music during a visit to Milan.

A third Italian, Padre Giovanni Battista Martini, renowned as a teacher and music historian, left 24 symphonies dating roughly between 1736 and 1777. There is remarkable consistency among this corpus. All but one are in the then-normal three movements (Sammartini wrote a number in four), and all have outer movements in a major key, whereas the middle slow movements are nearly always minor. Music of the Classical period greatly favours major keys; according to one investigator all but about 2 percent of 18th-century symphonies are major (excepting their expressive slow movements). Martini, although trained in counterpoint, avoided it in these works, favouring instead a treble-dominated texture over a simple, slow-moving bass. His chord vocabulary was restricted, his orchestration elementary. His melodies reflect a number of banal formulas in current use, but his manipulation and development of these formulas was skillful. Mozart visited Martini several times in Bologna in 1770.

In the hands of these and other Italian composers, symphonic style evolved considerably by 1740, and there was

continued fruitful experimentation through the 1750s. That Italians influenced contemporary and later German and, especially, Austrian symphonists cannot be doubted; but the extent of indebtedness and mutual influence cannot now be fully determined because of the scarcity of available scores. Relatively few of the surviving early symphonies have appeared in reliable modern editions, and those known to survive in manuscript represent but a fraction of the thousands composed. The early history of the symphony remains, therefore, a matter of speculation and debate, despite the enormous progress in research and publication in the mid-20th century.

German and Austrian composers. Germany and Austria were important centres of symphonic composition after about 1740. In Mannheim, Germany, the Bohemian Johann Wenzel Stamitz developed a remarkably well-trained orchestra that by 1756 comprised (in addition to 30 strings) four horns, pairs of flutes, oboes, clarinets, bassoons, trumpets, and timpani. With this ensemble, independent wind writing and creative orchestration flourished. Stamitz—himself a violinist and composer of more than 70 symphonies, chiefly mature works—and his contemporary Ignaz Holzbauer evolved a bold style born of the confluence of Italian melody and German seriousness. Counterpoint is abandoned; expression arises from orchestral crescendos and diminuendos (anticipated by Jommelli), characteristic melodic effects such as sighing appoggiaturas (falling figures) and rocketing arpeggiated (broken) chords, and strong dynamic, thematic, and textural contrasts. This style degenerated into mannerism with Stamitz's son Karl (composer of about 80 symphonies) and Johann Christian Cannabich, who enlarged the Mannheim orchestra.

The Mannheim composers

Mannheimers paid special attention to their development sections, which have been described as a "mosaic of fragments." They tended toward incomplete recapitulations, wrote tender slow movements, extended the number of movements to four: allegro (lively), andante (slow), minuet, and presto (rapid). Insisting on disciplined and precise performance, they wrote all parts out in full instead of leaving some to be realized by the players. With the rise of independent wind parts after *c.* 1770, the continuo became redundant and was abandoned; in outdoor music and in huge orchestras it had become inaudible anyway. Musical textures, which in early works were equally appropriate for string quartet or orchestra, became truly symphonic. Holzbauer occasionally features bassoon solos (as in the *Symphony in E Flat Major,* Opus 4, No. 3), liberating that instrument from simply doubling the bass. The elder Stamitz and Holzbauer were particularly highly regarded in Paris and strongly influenced symphonic activity there.

In Berlin and northern Germany another school arose, dominated by Carl Philipp Emanuel Bach, Johann Gottlieb Graun, and other musicians reared in a tradition of rigorous counterpoint and formal conservatism. Retaining three-movement format and avoiding strongly contrasting themes, they maintained contrapuntal interplay in the prevailingly homophonic texture. Even more than the Mannheimers, they concerned themselves with melodic development and ornamentation and with emotional expressiveness—an aesthetic approach they termed *Empfindsamkeit* (sensitivity). C.P.E. Bach's set of six symphonies commissioned by Baron Gottfried Bernhard van Swieten (completed 1773) aroused enthusiasm for their humour, technical challenge, and novelty of harmonic invention. Even more intensely passionate are the late *Orchestral Symphonies for Twelve Obbligato Parts* (1780), with their clever instrumentation and affective harmonies punctuated by unison passages. Reflecting the literary *Sturm und Drang* (storm and stress) movement, the Berliners powerfully influenced Haydn and Mozart and even Beethoven.

Another son of J.S. Bach, Johann Christian, the so-called London Bach, was perhaps yet more important for the future, influencing Mozart, who met him in London (1764) and again in Paris (1778). His 50 or so symphonies lack the passionate gestures of the Berliners' but are finely wrought and sophisticated in melody and orchestration. The *Sinfonias for Double Orchestra,* Opus 18, reflect in their *style galant* tunes the influence of his study with Martini in Italy, as well as French and Mannheim characteristics (*galant* was the 18th-century term for modern, light, elegant style). Numbers two and three of the set served as opera overtures. Simple in form, complex in texture, J.C. Bach's works are now considered to epitomize elegant, fashionable chamber music.

The style most influential on Haydn, Mozart, and Beethoven was that of the Viennese school, headed by Georg Matthias Monn and Fux's pupil Georg Christoph Wagenseil. The Viennese experimented further with orchestration and tone colour, emphasized the violin's melodic role, and displayed popular influence in their playful minuets. Monn, though perfectly at ease with counterpoint, ignored it in his symphonies. His four-movement *Symphony in E Flat Major* for pairs of oboes and horns, strings, and continuo is noteworthy for its wide dynamic range and gracious melody so characteristic of Viennese music. The first movement's first theme recurs in the recapitulation transformed into a virtuoso horn duet, reflecting the skill of the players at his disposal and the liberation of horns from a mere harmony-filling role. Elsewhere in the movement are found the slow, repeated-note bass and inconsequential viola part common in early symphonies. As yet the development is quite short and thematically limited, and the movement's scale is small—only a little more than three minutes. But as is often the case, the first movement is the most extended of the four, first movements having gained importance beyond the first movements of opera overtures. In later works Monn dropped the continuo and extended the role of the winds. Together with Johann Stamitz he has been considered highly influential in establishing modern symphonic style, although perhaps Wagenseil, an experienced keyboard player and opera composer, was a more successful composer, especially notable for skillful thematic construction and manipulation, working ornamentation into structure. Also active in Vienna was Haydn's friend Karl Ditters von Dittersdorf, who contributed no fewer than 115 symphonies.

The Viennese school

The Viennese style became widely disseminated in Europe. As many as 25 Viennese composers migrated to Paris; others moved to Dresden, Mannheim, and Munich and even as far as Dublin, St. Petersburg, and North America. Their style thus became truly international, and the innovations of these minor masters paved the way for the fully developed symphonic style of Haydn and Mozart. The music of the Viennese composers, mainly intended for elite entertainment, reflects the virtuosic attainments of many private orchestras and the cultured taste of aristocrats who commissioned a constant stream of symphonies and lighter chamber music.

The mature Classical period. *Haydn.* Joseph Haydn, despite his isolation from urban musical centres for much of his life, was revered throughout Europe, beloved by Mozart and Beethoven, and widely published and copied —so much so that the authenticity of many works attributed to him remains in question. One hundred and eight symphonies are thought to have been written by him; one of these is lost. Few composers show such remarkable growth as Haydn; from his insignificant youthful pieces, entirely dominated by the style of his pre-Classical elders, to the towering achievement of his last works, his symphonies display an evolution in form and content that had tremendous effect on his followers.

Viennese in style, some of his early symphonies display originality in the use of nonstandard phrase lengths and in their monothematic tendencies. *Symphonies Nos. 1* and *2* are in three movements, lacking a minuet. These works require a continuo (the slow movement in *No. 2* consists only of a bass and treble part), and horns and oboes are as yet not independent. *Symphony No. 3* and others incorporate contrapuntal movements. The sonata recapitulations are subtly altered; but unlike Stamitz', they are generally complete. Melodically, Haydn drew on folk music for inspiration, especially in minuets but also on *galant* and operatic styles. His work reveals a gradual growth in appreciation of the idiomatic qualities of wind

instruments, especially in trios of minuets (*e.g., Symphony No. 22, Symphony No. 40*); in *Symphony No. 5* he included winds in the slow movement, unusual at this time, and in *Nos. 6–8* he wrote independent wind solos, recalling the instrumental dialogue found in the Baroque concerto grosso. *Symphonies Nos. 6* and *11* begin with slow introductions, a characteristic which became common in Haydn's symphonies after *No. 84*.

With his appointment to the service of Prince Esterházy in 1761, Haydn's individuality began to emerge, partly because of his opportunity to experiment with the Esterházy orchestra. The bulk of his symphonic production dates from these years before 1771. Although humour and good nature pervade these works, stronger emotions and tension also begin to appear, as in the minor-key *Symphonies Nos. 26, 39*, and *49*. The *Farewell Symphony* (*No. 45*), with its adagio (slow) coda, displays Haydn's wit and is one of the best of his symphonies from the decade before 1780. This transitional period shows him striking out into more remote keys, introducing new themes in development sections, and growing more confident in formal craftsmanship and orchestration. Powerful and concentrated, the symphonies of the so-called *Sturm und Drang* period recall the *Empfindsamkeit* of C.P.E. Bach. By turns rigorously contrapuntal and lucidly witty, the vitality evident in the forms reflects Haydn's overflowing adventurousness. Contredanse (country dance) melodies may have inspired some of his themes, for example in the finale of *Symphony No. 88*.

Haydn's later works

The late *London* and *Paris Symphonies* reflect the influence of Mozart and show Haydn at the height of his power. No two movements are alike; the "mosaic" of theme elements pervades even transition sections and codas; each instrument shares in the melodic development; minuets grow in fire or dignity while finales exploit varieties of rondo form (see below). His slow movements, often straightforward sets of variations, engage in artful modulations prefiguring the romantic aspect of Beethoven. *Symphony No. 103* is especially thematically economical, and its movements are related by thematic resemblances, foreshadowing the cyclic nature of many 19th-century symphonies.

Haydn, though by no means the "father of the symphony," contributed enormously to a definition of the harmonic basis of Classical form, the dramatic role of key relationships, and the expressive capabilities of the winds. Major–minor contrasts, wide-ranging modulations, and reconciliation between counterpoint and homophony underlie his unambiguous moods, so different from those of Mozart. An eclectic architect, he amalgamated all the styles of his time in uniquely free and expressive shapes.

Use of rondo-form movements

Rondo form (in which a recurrent theme alternates with other material, as A B A C A) had been found especially in Italian opera and French instrumental music before *c.* 1770; in the 1770s and '80s it became second only to sonata form in symphonic importance. Exploited already by C.P.E. Bach, Stamitz, and others, rondos became a favourite last-movement form with Haydn and Mozart after about 1773. Haydn wrote a number of slow movements as rondos (*Symphonies Nos. 73, 74, 76*) and employed rondos 12 times in his last 17 symphonies. Mozart avoided the rondo in his last symphonies, perhaps because of the light nature of the form. Its vogue seems to have been brought about by the public's demand for structural simplicity and repetitive tunefulness. In the hands of Haydn and Mozart, however, the rondo increased in complexity, demonstrating in the so-called sonato-rondo the characteristics of sonata form, such as developments of earlier-stated material by means of fragmentation and modulation.

The symphonies of Mozart. Wolfgang Amadeus Mozart raised the symphony to heights never surpassed. Of his 50-odd symphonies, produced between 1764 and 1788, the earliest ones are conventional but precocious, reflecting influences of J.C. Bach, Sammartini, and Haydn. An invigorating first movement predominates, followed by a light cantabile movement and a fast finale or minuet (minuets in his symphonies date mostly after 1767). The *Symphony in B Flat Major*, K. 22 (1765; Köchel numbers are the standard way of identifying Mozart's works) contains a lovely chromatic (containing notes foreign to the key) slow movement in the key of G minor.

Mozart's synthesis of Italian and German structures

Mozart's exposure to Europe's main musical currents led him to synthesize the playful Italian homophonic and operatic style with serious German polyphony. This is evident in the agitated *Symphony in G Minor*, K. 183 (1773)—a *Sturm und Drang* work and his first minor-key symphony—and in the cheerful *Symphony in A Major*, K. 201 (1774). In these works the balance of interest shifts to the last movement. The addition of codas ("tails," extended closing sections reaffirming the tonic), the increased length and scope of slow movements and minuets, and a growing orchestral sensitivity all point toward maturity. In contrast with those of Haydn, Mozart's slow movements lean toward the sonata form with its inherent drama.

Mozart, unlike Haydn, was not a formal experimenter; he re-used successful structural formulations in later works. It was his treatment of melody that set him apart. He preferred to ignore monothematic structure; and his first and second themes, neither folklike nor mosaic-like, contrast strongly. His harmonic range is narrow compared with Haydn's, but within his range he constantly transformed thematic material. Development sections expand with the introduction of new thematic material and modulations over a wider tonal field. His recapitulations tend to be straightforward. In this mature period, Mozart's symphonies became unified thematically and expressively, using fuller imitation, more singing figuration, freer instrumentation (the *Paris Symphony*, K. 297, in D major, introduces clarinets). Mozart rejected Mannheim gesture in favour of better integrated dynamics.

Mozart's last 10 years saw him further exposed to Haydn's influence and very aware of J.S. Bach's music. The monumental last six symphonies reflect his experience as an opera and chamber music composer. The *Symphony in C Major*, K. 425, has a rare, slow chromatic introduction, while K. 504 in D major (*Prague Symphony*) dispenses with the minuet, has all three movements in sonata form, and uses canonic development (development by means of exact imitation). The last three symphonies (K. 543, in E flat major; K. 550, in G minor; K. 551, in C major, or the *Jupiter Symphony*), summits of the Classical genre, are bold in their harmonies and counterpoint; the serious minuet of K. 550 foreshadows the scherzo of Beethoven's *Fifth Symphony*. (The scherzo is a rapid, rhythmic, minuet-derived form.)

Mozart was no revolutionary. Receptive to the influence of others, he rejected more than he assimilated, transforming all into a uniquely personal idiom. Several of his symphonies were used as opera overtures, but the best ones are so complete in themselves as to make their use as incidental music unthinkable to modern taste. Mozart's and Haydn's mature symphonies are comprehensive in mood and design. The various movements balance one another so well that those who are accustomed to hearing them would find it difficult to accept the substitution of other movements. This tendency toward intimate relation among the standard four movements reflects the urge of these composers to seek unity on the highest hierarchic level—a trend foreign to most of their lesser contemporaries but a basic factor in the symphony's evolution throughout the next two centuries.

Beethoven. With Ludwig van Beethoven the symphony became no longer entertainment music but an expression of monumental intellect and innermost feeling, as in Haydn's and Mozart's late works. The *Symphony No. 1 in C Major* (completed 1800) is Haydnesque, particularly in the opening theme of the finale (comparable to the finale of Haydn's *Symphony No. 88*), but full of originality. Its four classically structured movements reflect Beethoven's concern with expressive woodwind writing and dynamics. The third movement (Menuetto) is prophetic of Beethoven's later whirling scherzos. The slow introduction to the first movement is remarkable for its avoidance of the tonic, a technique used often in later works to arouse tension.

The *Symphony No. 2 in D Major* (1802) is transitional and, like the *Symphony No. 1*, somewhat diffuse. A long introduction announces a work of grand dimensions. The lyric slow movement is rich in themes that are organically unified. A dynamic scherzo, only slightly dancelike, and an expanded sonata finale (with an enormous coda introducing a new theme) point toward the revolutionary length and structure of the *Symphony No. 3 in E Flat Major* (*Eroica;* completed 1804), a work that many consider to herald the dawn of musical Romanticism. The *Eroica* (Beethoven's title) no longer aims at an elite audience. Its first movement employs a multitude of themes, again drawn together into a cohesive organism and developed in a context of great harmonic tension. The tonic, E♭, is avoided near the beginning for 14 measures. A pathetic funeral march, replacing the ordinary slow movement, is followed by a vigorous scherzo; this leads to a variation finale, based on a theme from his *Creatures of Prometheus* ballet and full of contrapuntal development. The symphony marked a new fusion of old formal structures with Beethoven's dynamic outlook.

The *Eroica* symphony

The cheerful *Symphony No. 4 in B Flat Major* (1806) and fateful *Symphony No. 5 in C Minor* (1808), so unalike in character, were composed side by side. The *Fifth*, like the *Third*, a visionary work, is unified by the famous four-note motif that permeates all four movements in one form or another. The scherzo and finale are joined, and an explosion of C major in the last movement is celebrated with three trombones (possibly their first use in a symphony), piccolo, and contrabassoon. This grandiose edifice is constructed with relentless logic and rhythmic drive, hallmarks of Beethoven's mature style.

The *Symphony No. 6 in F Major* (1808), called the *Pastoral*, is in five movements, the first two and last in sonata form, each, according to Beethoven, expressing an aspect of rustic life. The whole has a unity of character that reflects a deeper rhythmic unity. A descriptive "Storm" movement links the scherzo ("Merrymaking of the Peasants") with a calm "Thanksgiving after the Storm" finale, which, incidentally, incorporates a Swiss yodel tune. The relaxed human and poetic qualities of the *Sixth* set it apart from the *Fifth* and from the demoniac *Symphony No. 7 in A Major* (1812), with its expanded scherzo and trio, blazing finale, and spirited first movement preceded by a long modulatory introduction.

The small scale of the first three movements of the *Symphony No. 8 in F Major* (1812) leaves one unprepared for its breathtaking finale. Its minuet is a subtle parody of the classical minuet of Mozart and Haydn.

The *Symphony No. 9 in D Minor* (*Choral*) found Beethoven deaf at its first performance in 1824. It marked a turning point in music history, not only for its novel inclusion of chorus and vocal soloists in the last movement and the extraordinarily variegated sonata form of that movement—incorporating a Turkish march, double exposition, double fugues, strophic (stanzaic) variations—but for the scope of the whole, a summary of Beethoven's ethical and symphonic achievements.

In his development of motifs and variation of entire themes Beethoven went unchallenged. He expanded the limits of Classical form, particularly in his finales, and increased the length of the symphonic process to more than four times the 15 or so minutes required for a pre-Classical symphony. Further, his orchestral sensitivity allowed all instruments a structural role while simultaneously making new demands on player and listener alike. Besides widening the scope of the orchestra with extra winds and percussion, he made it more than ever a cohesive single instrument, bequeathing to the 19th century a standard against which composers measured the effectiveness of their own orchestrations. Finally, through the immense concentration of his symphonies he made it impossible for his followers to equal the sheer quantity of production of the Classical composers; far too much effort went into creating a symphony to allow pouring them out by dozens.

Other composers. So overwhelming was the impact of Beethoven's symphonies, along with that of Mozart's and Haydn's mature ones, on later generations that they utterly obscure the productions of many other worthy symphonists. François Joseph Gossec, an early French symphonist, and Pierre van Maldere came to grips successfully with the dominating German–Italian idiom and were important among the Parisians influenced by Stamitz and his school. Van Maldere was eulogized for his imaginative thematic structures as well as for the unusually serious nature of his compositions, which strongly contrasted with the more lighthearted style characteristic of the Mannheimers.

An English composer, William Boyce, eclipsed by the London Bach, wrote eight sinfonias that betray in design the strong influence of theatre music. Basically merely overtures in French or Italian styles, they show none of the modern characteristics being formulated at the time in Germany; England, in general, was not quick to adopt the new symphonic style.

Eastern Europe produced revolutionary composers of whom until recently little has been known. Stamitz, Bohemian by birth, overshadowed such competent composers as Jiří Benda. Benda's symphonies, dating mostly between 1750 and 1765, are generally brief, in three movements, and close to the Italian overture in form and feeling. The sonata form is not exploited, although characteristics such as contrasting themes and contrast within a single theme (a technique used also by Mozart) suggest a Mannheim influence or at least a revolt against Baroque conventions.

Revolutionary composers of Eastern Europe

Luigi Boccherini, Giovanni Giuseppi Cambini, Michael Haydn (Joseph's brother), Leopold Mozart (father of Wolfgang), and many other important chamber music composers contributed numerous symphonies well worth performance. Later composers included the conservative Swede Franz Berwald and a brilliant but short-lived Spaniard, Juan Crisóstomo Arriaga, influential mostly in their own countries; and Muzio Clementi, Luigi Cherubini, Louis Spohr, and Carl Maria von Weber, who, although better known for work in other genres, were nevertheless popular symphonists. Spohr wrote a number of highly pictorial programmatic symphonies, going well beyond Beethoven's *Sixth*.

The Romantic era. *The period between Schubert and Liszt.* Among 19th-century symphonists several trends can be distinguished. Concerned to some extent with self-conscious emotional expression, they often tended to use looser forms and slower paces than the Classical composers. Sometimes this led to lax discipline but not in the case of the finest composers, among them Schumann, Brahms, and Dvořák, who were all very conscious of their debt to Beethoven, Mozart, and Haydn. With later composers, such as Anton Bruckner, Peter Ilich Tchaikovsky, and Gustav Mahler, the normal balance of form was sometimes upset in favour of romantic license, but they too derived their basic goals from the Classical composers, with a more or less heavy admixture of the influence of Richard Wagner.

Franz Schubert is known primarily as a songwriter. His nine symphonies stand in the shadow of Beethoven's but are revolutionary and Romantic in a way utterly different from Beethoven's. Whereas Beethoven wrestled with melodic problems, Schubert was a born melodist and consequently concerned himself more with the harmonic basis of form. He was likewise the more sensitive orchestrator, and in the last three symphonies he greatly expanded the role of the brasses.

His Symphony No. 1 in D Major (1813) and *Symphony No. 2 in B Flat Major* (1815) illustrate Schubert's departure from Classical models. Although the first movements are in sonata form, their pace is slower than the ordinary Classical allegro and is supported by long nonthematic passages that expand the harmonic arch. In the youthful sonata-form movements the second theme group is often set in an unexpected key before the music turns to the dominant at the end of the exposition. In recapitulations, too, Schubert shies away from harmonic simplicity and Classical expectation; his phrasing also is often irregular. Schubert's slow movements, scherzos, and minuets are not as strikingly original. Clear references to move-

ments and themes of Beethoven occur in these early works, and in key scheme and major–minor contrast Schubert often betrayed his indebtedness to Beethoven. He was unembarrassed to borrow melodic material, which he transformed in an utterly personal way. This is particularly the case in the *Symphony No. 4 in C Minor* (the *Tragic;* 1816). The *Symphony No. 5 in B Flat Major* (1816), scored for a smaller orchestra, more strongly recalls Mozart and Haydn. The highly emotional *No. 6 in C Major* (1818) is of larger scale, based as usual more on rhythmic and harmonic impetus than melodic development. The incomplete draft of the *Symphony in E Minor-Major* (1821) has inspired attempts at completion. But it is the last two (the *Symphony in B Minor* or *Unfinished,* 1822, and *Symphony in C Major,* or *Great,* 1828) that raise Schubert to high rank among symphonists. Composed for large orchestras, they nevertheless reflect Schubert's experience in writing for voice and piano.

Schubert's *Unfinished* and *Great* symphonies

The *Unfinished* consists of two complete movements in $\frac{3}{4}$ and $\frac{3}{8}$ time and a sketch for a scherzo. The complete movements form a convincing unity; masterful in harmonic organization and orchestration, they are expressive without being diffuse, a criticism often levelled against passages in Schubert's earlier works. The *Great* is of Beethovenian scale, partly because of extensive repetition. The scherzo and related slow movement, no longer simply rustic pieces, are both sonatas. Irregular phrases, modulatory schemes, and rhythmic force give evidence of Schubert's concern with form based on slowed-down and far-reaching harmonic motion. His rhythmic manipulation was un-Classical, his themes personal and of more than Classical significance.

Berlioz and Liszt

With the first group of symphonists born in the 19th century the Romantic style was fully fledged. The French composer Hector Berlioz and the Hungarian Franz Liszt contributed large symphonic works that to some extent departed in form from the Classical sonata-centred model. The literary program to Berlioz' *Symphonie fantastique: Épisode de la vie d'un artiste* (1830) was not written until the music was well along toward completion. The symphony was thoroughly planned out thematically and formally and stands as a musical unity without regard to the program, which Berlioz himself eventually withdrew. A very personal expression nevertheless, the *Fantastique* introduces a structural *idée fixe,* a theme (representing his mistress?) recurring throughout the five movements in various rhythmic forms, serving to unite the "scenes" musically as well as dramatically.

Harold en Italie (1834; after Byron's poem), like the *Fantastique,* makes use of pre-existent material and is unified not only by a program but by a recurrent theme, a viola solo representing Harold. This theme is not subject to the kind of variation given the *idée fixe* in the *Fantastique;* yet from it springs much of the melodic inspiration of the whole work. Berlioz' third symphonic work, *Roméo et Juliette* (1839), rarely heard in its entirety, incorporates chorus and vocal soloists into its five large sections, which are programmatically derived from episodes of Shakespeare's drama. Not coincidentally, Berlioz was a great admirer of Beethoven. Beethoven's unity of moods, thematic development, and dramatic orchestration were models for Berlioz to extend, although he did so outside the formal confines of the sonata and with even more explicit passion.

Liszt owed much to Berlioz, both in his handling of enlarged orchestral forces and in thematic transformation (as opposed to development). The three movements of his *Faust Symphony* (1854) bear the names of Goethe's characters: Faust, Gretchen, and Mephistopheles; and the final movement parodies themes of the first two in a satisfyingly diabolical manner. Characters aside, the music is highly effective and balanced; Liszt revised the score over several decades. The score is dedicated to Berlioz.

Liszt's other symphonic work, the *Symphony to Dante's Divina Commedia* (1856), depicts the Inferno, Purgatory, and Paradise. Liszt, at times a devout Catholic, portrayed Dante's scenes with great imagination and passion, cleverly suiting his melody—sometimes simple and tranquil, sometimes chromatic and writhing—and harmony to the special characters of the three levels. The symphony is dedicated to Wagner, who suggested the third-movement setting of the Magnificat for female chorus and orchestra. As do many operas of Wagner, Liszt's work uses the leitmotiv, an extension of Berlioz' *idée fixe.*

If Berlioz and Liszt represented a trend toward freedom and extramusical content in symphonic writing, Schumann and Mendelssohn were more conservative though not strictly comparable. All four were deeply concerned with formal discipline, but Schumann and Mendelssohn departed less widely from Classical norms and made less point of extramusical associations.

Mendelssohn's symphonic works

Felix Mendelssohn wrote 16 symphonies and a symphony-cantata. Twelve of the symphonies are immature works; but the remainder fairly exemplify his style: facile, full of light melody and brilliant orchestration, occasionally oversentimental, according to some critics. He is best known for his *Symphonies No. 3* (*Scottish*) and *No. 4* (*Italian*), both in A major–minor. The *Scottish* (also called *Scotch*), completed in 1842, although not programmatic, is expressive of Mendelssohn's poetic nature. Its beginning was sketched during a visit to Scotland in 1829. In structure the work consists of four movements played without pause, with a slow introduction. Its fairy-like scherzo, which incorporates part of a Scottish folk song, exemplifies the delicate moods that Mendelssohn excelled in creating. The other movements are well developed, the many contrasting themes integrated contrapuntally and extended with interesting modulations. But although it is full of good music, the symphony is less powerful than its companion, the *Italian* (finished 1833). This happy work, inspired by visits to Rome and Naples, is particularly colourfully orchestrated. It ends with a dance movement incorporating three themes; the minor tonality does not detract from its vivacity. The first movement, too, has three themes, the third introduced in the development section. The second movement, recalling a religious procession, and the third, a quasi-minuet and trio, are picturesque without being descriptive and represent Mendelssohn at his finest—uncomplicated, lush, and vigorous.

Schumann's miniaturist style

Robert Schumann, like Mendelssohn and Mozart, wrote his symphonies at an age when most longer-lived composers are just beginning to mature and wrote only a few truly great ones. Like many first-generation Romantic composers Schumann was essentially a miniaturist, most at home in songs and short piano works. His orchestral style reflects these qualities; rhythmically restless, often repetitive, not sensitively scored, they have been praised more for their harmonic subtleties and wonderful lyric melodies than for development of these ideas.

The *Symphony No. 1 in B Flat Major* (*Spring;* 1841), based on a poem by Adolph Böttger, originally had titles given each movement; these were soon rejected by Schumann and indeed are irrelevant to the music. The first movement, opening with a slow introduction (a tradition since the days of Haydn), incorporates three contrasting themes (the third introduced toward the conclusion of the movement) as well as the opening dramatic figure of the introduction. The slow movement and unusual scherzo (it has two different trios, rather than one) are linked thematically and played without a pause between them. The impulsive progress of the finale is interrupted before the recapitulation by slower passages for flute and hunting horns, perhaps intended by Schumann to be descriptive.

The *Symphony No. 2 in C Major* (1846) is tightly organized and owes something in design to Beethoven. It has been overshadowed by more frequent performances of the last two symphonies, *No. 3 in E Flat Major* (*Rhenish;* 1850) and *No. 4 in D Minor* (1841, rev. 1851). The five-movement *Rhenish* is less "classical" than the *Symphony No. 2.* Inspired by a ceremony at Cologne Cathedral as well as by the appearance of the cathedral itself, the polyphonic grandeur and harmonic richness, especially of the fourth movement, are tempered by the relaxed pace and rustic character of the scherzo and

following short, quiet slow movement. The outer movements are related both thematically and in mood, and the last two movements also share material, forming a large cohesive structure.

Even more cohesive is the plan of the *Fourth Symphony* in which all four movements are played, as in Mendelssohn's *Scottish Symphony*, without pause. A single theme recurs in various guises in all four movements; this thematic transformation is a hallmark of Schumann's style, as of Berlioz' and Liszt's. The last movement introduces new material but without destroying the cyclic nature of the whole work. Cyclic structure, which relates separate movements by means of re-use of thematic material, is a feature of much symphonic writing after Beethoven. As composers gradually departed from repetitive forms, cyclic construction became a chief mode of achieving unity over a large time span and greatly enlarged harmonic vocabulary. An advanced form of cyclic construction may be seen in the Belgian composer César Franck's influential single *Symphony in D Minor* (1888).

Brahms and his contemporaries. Although Johannes Brahms's four symphonies are popularly considered to be no less important than the greatest earlier symphonies, the contribution of his contemporary Anton Bruckner is controversial. Bruckner, a devout Catholic whose church music is among the finest of his generation, is noteworthy not only for the excessive length and heavy orchestration of many late movements but for his Wagnerian harmonies, large-scale repetitions, and (at its best) monumental conceptions of form. Bruckner gathered much from studying late Beethoven and Schubert. Yet his style evolved little in the course of nine symphonies (he was over 40 when he wrote his first symphonies; two other unnumbered early ones are never heard, and his last was incomplete at his death). Entirely personal in expression, his symphonies underwent frequent revision. They are in four movements and are basically unprogrammatic, even conservative. Despite his devotion to Wagner and Beethoven, Bruckner remained provincial; his technique was grounded on traditional studies, and his movement types seem to follow a set of typical formulas that derived from Classical patterns, especially the sonata. Within these formal types he develops themes powerful in their simplicity and monolithic in harmonic expanse. Chords reminiscent of German chorale (hymn melody harmonizations) and tremolo or pizzicato (plucked) accompaniments occur along with organ-like counterpoint and pedal tones (sustained notes against changing harmonies). Hardly imbued with youthful vigour, some scherzos still have roots in the fertile Austrian popular music that nourished Haydn and Beethoven. These movements, however, are not sufficient to lighten the overall impression of density. Unlike those of Brahms, Bruckner's symphonies are not immediately rewarding; yet connoisseurs, including Mahler, respect Bruckner's heroic finales, in which themes from earlier movements are sometimes combined. *Symphony No. 4 in E Flat Major* (1874, first performance of revision, 1881), with a Beethovenian andante and scherzo recalling the hunt, is noteworthy for the use of four themes in the first movement. *No. 7 in E Major* (1881–83, rev. 1885), well received at first hearing, is Wagnerian in orchestration (Wagner tubas play in the adagio) and makes use of contrapuntal techniques developed in the Renaissance. In the last two symphonies the adagio movements occur after the scherzos rather than before.

Traits of Bruckner's symphonic style

Individuality of Brahms's symphonies

The symphonies of Brahms, each highly individual, appealing on first hearing, and rewarding to rehear, could hardly be more different from Bruckner's. Yet Brahms, his technique grounded on thorough study of Classical and Baroque works, was no less essentially conservative. He retained the four-movement format and familiar methods of thematic development and rigorous contrapuntal craftsmanship, avoided programmatic content, and was always concerned with aural effect. These concerns are reflected in his orchestration, which is never merely flashy, never astonishing. His entire remarkable skill and attention to detail served the lyrical, spontaneous flow of his melody. His tunes, Romantic as Schubert's, are developed with consistency, refinement, and a harmonic interest foreign to many lesser contemporaries (especially those seduced by Berlioz and Liszt but without understanding their talent). Brahms is a master of understatement. Finished in 1876, 20 years after it was begun, his *Symphony No. 1 in C Minor* was the fruit of a mature man's experience. It carried on where Beethoven left off, drawing inspiration especially from Beethoven's *Fifth*. The triumph of major over minor, epitomized in the finale, underlies the whole. The key scheme of the movements is cyclic, based on a succession of rising major thirds: C minor, E major, A♭ (= G♯) major, C minor–major. The third movement, formally a scherzo with trio, slackens tension as though to prepare for the marvellously developed finale, which is preceded by a broad introduction (as is the first movement) that sets harmonic and thematic goals for the remainder.

Clear goal orientation also characterizes the *Symphony No. 2 in D Major* (1877). Unlike the *First* in mood, it is pervaded in all movements by optimistic calm. First and last share a three-note motif that is joined by other serene themes in the first movement, which includes a long horn solo near the end. The second movement begins in B major, a third lower than the first, and the fast third movement begins a lower third still, in G major; the finale returns to D. This third-based tonal scheme, like that of the *First*, marks Brahms as a true Romantic, as do the tempo changes within movements, the sensuous modulations that circumscribe harmonic goals, and the intense major–minor conflicts.

Modal tension—major versus minor—characterizes the cyclic *Symphony No. 3 in F Major* (1883), the movements of which are related by material derived from Brahms's "motto" motif, FAF, *frei aber froh* ("free but joyful"). Winds are prominently featured, particularly clarinet and horn, as elsewhere in these symphonies. Brahms retained Classical outlines as usual; but, as in the *First*, the scherzo with trio serves to throw the vigorous, stormy finale into relief. The first movement plunges to the heart of things from the opening chords, harmonizing the motto; there is no introduction, no coyness in exposing the main themes. Chromatic harmony and contrapuntal development are fully exploited.

Brahms's architectural skill is nowhere more in evidence than in the finale of the *Symphony No. 4 in E Minor* (1884–85), an extended chaconne, or set of variations over an (eight-bar) repeated bass melody. This movement is almost Baroque; and elsewhere in the work Brahms employs Baroque contrapuntal techniques, chromatic labyrinths, and modal melody that hovers between major and minor but is neither. In this work particularly, but throughout the symphonies, Brahms epitomized the tendency of the later Romantics to seek a balance between the expressive forms of the early 19th century and older traditional technique, to apply to the wealth of available harmonic and orchestral colour constructive methods consciously founded upon study of Beethoven (particularly *Symphonies No. 4*, *No. 7*, and *No. 8*), Handel, and other models. These disciplined composers (including Schumann and Mendelssohn) reacted partly against what they felt to be the extramusical emphasis and compositional excesses of Berlioz, Liszt, and lesser figures, who took as their point of departure the less conventional Beethoven of *Symphonies No. 5*, *No. 6*, and *No. 9*.

Dvořák and Tchaikovsky

Both trends found reflection in the symphonies of Dvořák and Tchaikovsky, composers who were products of growing nationalistic tendencies in music. Antonin Dvořák continued a distinguished line of Bohemian symphonists stretching back to Stamitz. Conscious of his musical heritage, Dvořák infused his music with folk-derived elements, particularly dances; his last symphony, *No. 9 in E Minor* (*From the New World*, 1893), even incorporates American tunes; but these are almost incidental to the strong Slavonic character of the work. An early devotee of Wagnerian sonorities, Dvořák in his later symphonies returned to the more conservative models and orchestrations of Beethoven and Brahms. It is these later works, through which Dvořák is known today, that have led detractors to call him a "second-rate Brahms." In fact, Dvořák's melodic invention, often based on irregular

folklike scale forms, and his captivating irregularity of phrase length, surprising variety of orchestration, and impetuous rhythms are entirely personal.

Peter Ilich Tchaikovsky, on the other hand, was not comfortable working with pre-established formal models but was at his best in ballets and tone poems in which his somewhat extravagant nature found fuller scope for expression. Of his eight symphonies, only *No. 4 in F Minor* (1877), *No. 5 in E Minor* (1888), and *No. 6 in B minor* (*Pathétique*, 1893), actually fourth, sixth, and eighth in order of composition, are well-known. These are controversial works, partly because their novel structures are not easily analyzed (or heard) in standard formal ways. Some feel that Tchaikovsky's freedom and tendency to musical autobiography were inimical to purely abstract musical expression and that understanding of his music depends on knowledge of his state of mind at various times or upon some extramusical imagery or program. This attitude conflicts with an essential determinant of symphonic idiom, which is that the establishment and working-out of tensions in the piece are primarily occasioned by purely musical, formal means; and that extramusical data, interesting though they may be, are not directly relevant to apprehension and appreciation of the symphonic process. If Tchaikovsky's symphonies are to be considered successful as symphonies, they must make purely musical sense—and the three mentioned fulfill this condition.

Tchaikovsky's kind of musical logic, however, is quite different from that exemplified by the main-line German symphonists. Isolated in his formative years from the influence of Brahms and Wagner, he learned instead by hearing Mozart and Italian opera, characteristics of which he fused with elements of non-European melody, harmony, rhythm, and colour; in this he followed Aleksandr Borodin and other Russians. He strongly favoured the minor mode, no doubt partly because of its inherent instability. This unique confluence of stylistic sources produced a new model for later symphonists, particularly in regard to orchestration and a re-evaluation of sonata form based on a fresh conception of tonal harmony.

Mahler as the culmination of the Viennese tradition

Mahler. With Gustav Mahler, the central path, if not the culmination, of Viennese symphony was regained. In importance Mahler's nine completed symphonies (a 10th was left unfinished at his death) stand equal to any corpus since Beethoven's. That this has been so recently recognized may have something to do with historical perspective but is more probably a matter of relevance of his music to the mood of the mid-20th century. Nowhere else in music is found such explicit cynicism, such deliberate distortion of the familiar; by the same token no symphonist exceeds him in desire for reconciliation. Mahler's melodies and harmonies are strongly goal oriented; that his goals are so frequently frustrated or beset with obstacles reflects a truly contemporary outlook.

Mahler's symphonies, like those of Tchaikovsky and Berlioz, suffer from being too often considered solely in the light of extramusical associations. Mahler himself suppressed the programs of his early symphonies and subjected their music to frequent revision. Structurally, the symphonies are entirely logical, even simple, beneath the multitude of themes and wealth of colouristic detail. Mahler was a fastidious and brilliant orchestrator. The seeming superfluity of orchestral resource he called for, especially in the later symphonies, is handled with restraint and sensitivity—Mahler was a conductor as well as a composer and knew well the capabilities of the instruments.

Enormous in time scale, Mahler's symphonies contain sufficient variety and contrast to maintain interest. Underlying this stylistic multiplicity—including parodies of folk song, waltzes, fanfares, marches, text painting (four symphonies include voices), chorales, borrowings from other composers as well as his own songs—was a leaning toward cyclic structure, with themes or motives shared among movements, as in his song cycles and those of Schubert and Schumann. Mahler also experimented with tonal structure to the extent of combining movements in unrelated keys, so that the ear never tires of a single tonal area. *Symphonies No. 5, No. 6,* and *No. 7,* themselves a huge cycle, show him coming to grips with Classical sonata form, greatly expanded though it is. Unity on a lower level is achieved through extensive counterpoint. These later works show an economy of structure that is foreign to Bruckner, from whom Mahler nevertheless learned much.

Nielsen and Sibelius

The 20th century. Important symphonists of the early 20th century include many non-Germans. Carl Nielsen and Jean Sibelius, the former Danish, the latter Finnish, both owe much to the Viennese symphonists but acquired individual styles that resulted in new conceptions of symphonic form. Nielsen's six symphonies display a kind of unity based on "progressive" or "emergent" harmony, one key moving on to the next in such fashion that the gamut of harmonies in a single symphony (or movement) does not totally relate to a single tonic. His harmonies sometimes fluctuate between two or more goals and incorporate chromatic and modal features. This untraditional harmonic tension is an aspect of the breakdown of normal Classical ways of establishing tonality, for which the 19th century (and Wagner in particular) is largely responsible. (Nielsen, however, was an anti-Wagnerian.) This "destructive" impulse, elements of which may already be heard in Haydn, led to constant re-evaluation of the basis of sonata form and hence of the symphony as a whole. As the traditional harmonic foundation weakened—partly because the enlarged time scale makes long-term harmonic relationships hard to hear—symphonists sought new unifying techniques, among them cyclic forms, extramusical plots, progressive harmony, etc. None of these is incompatible with dramatic sonata form, taken in its broadest sense. Nielsen retained and emphasized key conflict as a dynamic force and experimented with counterpoint and conflicting rhythms (*e.g., Symphony No. 4, The Inextinguishable,* 1916; and *Symphony No. 5,* 1922), joined movements, and the rest.

Sibelius wrote his seven symphonies between 1899 and 1925. Like Nielsen's, they departed from Classical models (notably Beethoven's *Fifth*) and reflected especially the advances of Brahms and Bruckner. Sibelius was a restrained orchestrator, a non-Wagnerian. Although nationalistic, he used no folk songs and incorporated no programs. He had little immediate influence, isolated as he was after World War I, but is highly regarded presently in his native country, England, and the United States. Sibelius, like Nielsen, is not tied to Classical tonic–dominant opposition. He sometimes introduces modal scales and polar harmonic goals between which the music oscillates. A chief unifying device is a repeated bass line or sustained pedal tone. His themes are often groups of melodic fragments, meaningless when out of context, that are capable of being combined in various ways. The gradual integration of these motifs is an important means of development. In this Sibelius resembles Borodin; the effect is quite unlike that of Mahler's long melodies. Structurally Sibelius is terse and simple. He eliminated transitions and introductions, avoided simple recapitulation, and kept harmonies static or ambiguous over long stretches. Rarely was he merely playful. He was fond of low, dark sounds. Sibelius is hardly unemotional but certainly is not effusively romantic. His symphonies represent a great contrast with Mahler's, and he strikes a new path away from the sonata.

Elsewhere in the early 20th century important symphonic contributions were being made by the Frenchman Albert Roussel, whose four symphonies are elegant and classical in form; the American Charles Ives, who quotes well-known tunes in his highly dissonant four symphonies; and the Englishmen Sir Edward Elgar and Ralph Vaughan Williams. Their works are not often heard outside their native countries, and the extent of their influence on the growth of symphonic thinking remains to be determined. A more important innovator, the Viennese Arnold Schoenberg, considered his *Chamber Symphony,* Opus 9 (1906), "the climax of my first period." Basically still sonata-oriented and tonal, this work departed from the gargantuan orchestrations of Mahler and Richard Strauss and is scored for 15 soloists, 10 of whom play

winds. The idiom is highly dissonant and motivic, straining at the limits of traditional tonality and pointing to Schoenberg's formulation of dodecaphony, or 12-tone serialism (deriving melody and harmony from the composer's ordering of the 12 tones of the chromatic scale).

Stravinsky's imaginative force

Igor Stravinsky likewise inherited a living tradition and transformed it with great imaginative force. Like Schoenberg, he was not primarily a symphonist; his early *Symphony in E Flat Major* (1905–07) is no more original than anything of other Russians such as Borodin or Aleksandr Glazunov. But the *Symphony in C* (1940) and *Symphony in Three Movements* (1942–45) are unique. The former, a Neoclassical work, reinterprets in Stravinsky's language the thematic construction and sonata form of the Classical era. The result, far from a simple parody of Classical style (such as in Sergey Prokofiev's *Symphony No. 1 in D Major* or *Classical Symphony* of 1917), was an altogether fresh and revealing insight into the implications of Haydn's work. The *Symphony in Three Movements*, inspired by wartime impressions, is independent of models, yet in outward form the movements appear traditional. Two other of Stravinsky's works deserve mention here: the *Symphonies of Wind Instruments* (1920, rev. 1947) and the *Symphony of Psalms* (1930), which includes chorus. These works are symphonies only in the 17th-century sense of ensemble pieces; in both form and structure they are dissimilar from the other symphonies.

Chief among other Russians are, besides Prokofiev (seven symphonies), Nicolay Myaskovsky (also seven) and Dmitry Shostakovich (14 by 1972). Shostakovich's large corpus marks him as one of the most important and prolific living symphonists. His works are uneven in quality. A number are programmatic, dealing with Russian political and social upheavals. Shostakovich at times concerns himself with incorporating popular and folk tunes, giving expression to national feelings, and, more basically, with problems of achieving cyclic unity by such means as leitmotivs and thematic recurrences and joining movements without pause.

The mainstream of American symphonic production resides with Roy Harris, Walter Piston, William Schuman, and Roger Sessions. Important as teachers, these men have influenced younger symphonists such as Easley Blackwood and helped create a distinctive style of composition that resulted in the first important symphonic output in the Western Hemisphere. Although many younger composers on both sides of the Atlantic are turning to avant-garde techniques, including experiments in formlessness, and to small instrumental and vocal ensembles as well as to electronic sources, the future of the symphony seems secure. Extremely large orchestrations are not widely favoured, partly because of the difficulty and expense of mounting a performance; yet Blackwood has written one symphony for an orchestra larger than Mahler's as well as one for chamber orchestra. Audiences tend to favour works in familiar forms. Prizes and commissions assure a steady output of symphonies and many have been written for college and university orchestras. Flourishing amateur orchestras also furnish outlets for the young composer. As regards the masters of the 18th and 19th centuries, there seems no possibility that their symphonies will become museum pieces in the foreseeable future. On the contrary, the recording and broadcasting industry has brought about renewed interest in even the lesser known symphonists. Scholars and public alike have now almost unlimited access to the treasures of three centuries of symphonic production.

BIBLIOGRAPHY. FRIEDRICH BLUME *et al.*, "Symphonie," in *Musik in Geschichte und Gegenwart*, vol. 12, col. 1803–99 (1965), important historical and regional surveys of symphonic production, with extensive bibliography; NATHAN BRODER, "The Wind-Instruments in Mozart's Symphonies," *Musical Quarterly (MQ)*, 19:238–259 (1933), a study of the changing role of winds in 18th-century orchestration; HOWARD BROFSKY, "The Symphonies of Padre Martini," *MQ*, 51:649–673 (1965), a discussion of Martini's pre-Classical symphonic style; BARRY S. BROOK, *La Symphonie française dans la seconde moitié du XVIIIe siècle*, 3 vol. (1962), an important survey of over 1,200 works by 150 composers; ADAM CARSE, *Eighteenth-Century Symphonies: A Short History . . .* (1951), with emphasis on pre-Classical and early Classical forms and the overture; MALCOLM S. COLE, "The Vogue of the Instrumental Rondo in the Late 18th Century," *Journal of the American Musicological Society (JAMS)*, 22:425–455 (1969), evidence for the rise and passing of a formal fashion; CHARLES L. CUDWORTH, "The English Symphonists of the Eighteenth Century," *Proceedings of the Royal Musical Association*, 78:31–51 (1951–52), a survey of a neglected national school; PHILIP G. DOWNS, "Beethoven's 'New Way' and the *Eroica*," *MQ*, 56:585–604 (1970), an examination of the symphony in the light of a crisis in Beethoven's life; FRANK E. KIRBY, "Beethoven's Pastoral Symphony as a *Sinfonia caracteristica*," *MQ*, 56:605–623 (1970), traditional pastoral elements related to Beethoven's symphonic form and content; H.C.R. LANDON, *The Symphonies of Joseph Haydn* (1955, suppl. 1961), a thorough analysis of Haydn's evolution toward greatness; JAN LARUE, "Major and Minor Mysteries of Identification in the 18th-Century Symphony," *JAMS*, 13:181–196 (1960), on problems of authenticity and attribution, mostly among minor masters; and "Significant and Coincidental Resemblance Between Classical Themes," *JAMS*, 14:222–234 (1961), discusses and illustrates "theme families" and elements of melodic formation; GORDANA LAZAREVICH, "The Neapolitan Intermezzo and Its Influence on the Symphonic Idiom," *MQ*, 57:294–313 (1971), an examination of a strong determinant of early symphonic style; ERNEST SANDERS, "Form and Content in the Finale of Beethoven's Ninth Symphony," *MQ*, 50:59–76 (1964), a study of the structure of this famous movement; ROBERT SIMPSON (ed.), *The Symphony:* vol. 1, *Haydn to Dvořák* (1966) and vol. 2, *Elgar to the Present Day* (1967), a collection of essays surveying the production of important symphonists; NICHOLAS TEMPERLEY, "The *Symphonie fantastique* and Its Program," *MQ*, 57:593–608 (1971), on Berlioz' music as related to its extramusical "plot"; DONALD F. TOVEY, *Essays in Musical Analysis*, vol. 1, *Symphonies* (1935), old but perceptive and well-written discussions of chief works; HOMER ULRICH, *Symphonic Music: Its Evolution Since the Renaissance* (1961), one of the few wide-ranging histories available in English.

(La.L.)

Syracuse

Syracuse (Italian Siracusa) is a city on the east coast of Sicily; it was the island's chief Greek city in ancient times. Its nucleus is the southward-projecting island of Ortygia, half-enclosing a bay (the Great Harbour), defined on the south by the eastward-curving coastline at Plemmyrium.

History. Settled *c.* 734 BC by colonists from Corinth led by the aristocrat Archias, Syracuse soon dominated the coastal plain and hill country beyond. The original Greek settlers formed an elite (*gamoroi*); the Sicel natives (Siculi) worked the land as an oppressed class. Colonies at Acrae and Casmenae (about 12 miles [20 kilometres] to the west) and Camarina (about 43 miles [70 kilometres] to the south) illustrate the extent of Syracusan influence and prosperity.

In the early 5th century BC, the Syracusans were defeated by Hippocrates of Gela, a city to the west. The power of the *gamoroi* was subsequently ended by a democratic revolution, and in exile they turned to Hippocrates' successor, Gelon, who captured Syracuse and transferred his government there. His defeat of a great Carthaginian invasion in 480 at Himera confirmed his supremacy; and under him and his brother Hieron, Syracuse reached a high point of power and cultural brilliance. A revolution in 466 overthrew the third tyrant, Thrasybulus, and under a democratic constitution the Syracusans, though abandoning the territorial empire Gelon had acquired, survived wars against the city of Acragas and the Siculi. Most importantly, they also survived a long siege by the Athenians (415–413) during the Peloponnesian War, ultimately destroying the Athenian forces and weakening Athenian power in Greece itself. A few years later Sicily faced a Carthaginian resurgence. Syracuse was saved from the fate that overtook Acragas and other cities by its general, Dionysius, who obtained autocratic power in 405 and reigned as "tyrant" until 367. He fought three wars against the Carthaginians, confining their territorial dominions to the western part of Sicily, and he extended Syracusan control to most of the "foot" of Italy. His son

Dionysius II enjoyed a decade of peace before his autocracy was challenged. The brief rule of Dion preceded a period of anarchy ended by the Corinthian Timoleon, who defeated Carthage and reordered Sicilian affairs (345–336), introducing a moderate oligarchy in Syracuse. In 317 this was overthrown by the adventurer Agathocles, who became "tyrant" and later king. He again established a Syracusan empire, which broke up at his death in 289.

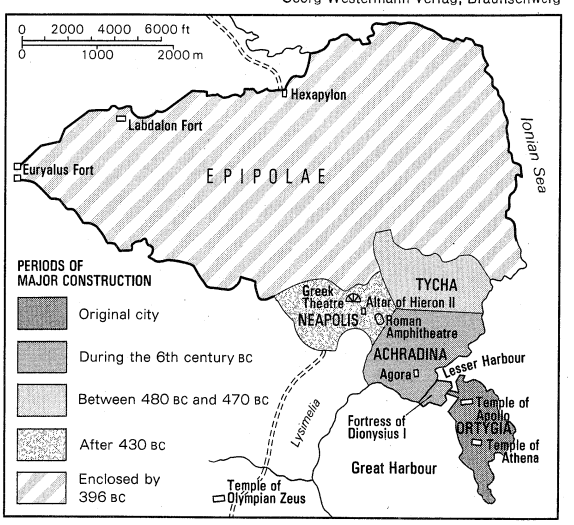
In the ensuing chaotic conditions Sicily was rescued by Pyrrhus of Epirus from further Carthaginian encroachments, but his mercenaries later seized Messana. Their defeat by Syracuse under a new leader, Hieron II, caused the intervention of Rome, with whom Hieron came to terms. After Hieron's death in 215, the Syracusans became allies of Carthage and were besieged by the Romans in 213. After its fall in 211, Syracuse became the seat of the Roman governor, sharing in the general prosperity that Roman rule brought, despite the depredations of the infamous governor, Verres, slave revolts in the hinterland (135 and 104), and the impact of the civil wars at the end of the republican period.

Roman rule

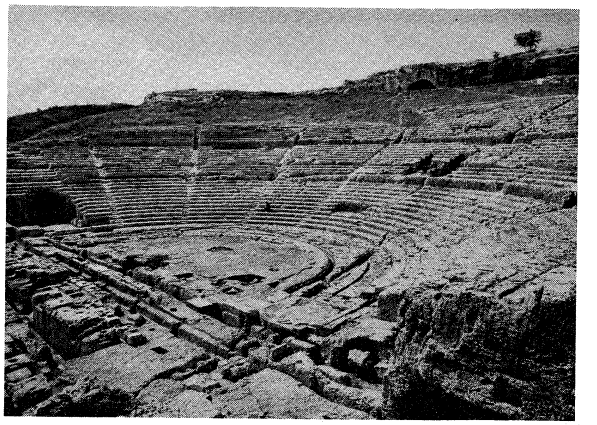
In AD 280 Syracuse was ravaged by Frankish invaders but soon recovered a prosperity that lasted until its capture by the Arabs in 878. During the late Roman and Byzantine periods and under Norman, Swabian, and Spanish rule, it shared the vicissitudes of the rest of the island, surviving as a not unimportant appanage of the Sicilian realm, especially in the Aragonese period (14th–16th centuries), when it formed a special enclave. Shattered by an earthquake in 1693, it owes some of its finest architecture to the reconstruction period. The Allied invasion of 1943 in World War II caused it some damage, swiftly repaired in a new and greater postwar prosperity. The present city, capital of Siracusa Province (*provincia*), is a centre for processing local agricultural produce and has several other light industries. The harbour, with its commerce and fishery, and tourism are further sources of prosperity.

Archaeological remains. Only in modern times has Syracuse begun to cover as wide an area as its ancient predecessor. Remains of antiquity on the continuously inhabited Ortygia are less striking than those in Neapolis, long a country district. Here the Greek theatre, Roman amphitheatre, and altar of Hieron II, pillaged in 1526 to build defensive walls, were archaeologically studied in 1839 by the Duca di Serradifalco but owe most to the systematic investigations of Paolo Orsi between 1890 and 1935. The theatre now visible is that of Hieron II. It had a capacity of 15,000 and was built on the site of two earlier theatres. Above it is a nymphaeum (fountain) that was one of the ancient city's sources of water. Hieron II's great altar, over 200 yards long on which 450 oxen could be sacrificed simultaneously, flanked a shady piazza developed in the Roman period. The amphitheatre was built late in the 2nd century AD as an arena for gladiatorial shows. The quarries in the same district, the source of

Adapted from *Westermann Grosser Atlas zur Weltgeschichte*, Georg Westermann Verlag, Braunschweig



Ancient Syracuse.



Ruins of the Greek theatre of Hieron II (3rd century BC) at Syracuse, Sicily.
Leonard von Matt—EB Inc.

the city's building stone, have been turned into luxuriant gardens; here the Athenian prisoners were quartered in 413. The most imposing evidence of Syracusan power consists in the fortifications of the Epipolae Plateau, culminating in the Euryalus fort at their western end. Begun by Dionysius I, they were probably strengthened by Agathocles but in their final form are the work of the scientist and engineer Archimedes.

The cathedral on Ortygia, with a fine Baroque facade, incorporates the Doric columns of the temple of Athena built as a thank offering for the victory of Himera. Architectural reliefs in painted terra-cotta from earlier buildings in this area are preserved in the museum. The remains of the temple of Apollo (*c.* 565 BC) stand near the bridge to the mainland (another temple of the same period, that of Olympian Zeus, lies on the west side of the Great Harbour).

Ortygia

The streets of Ortygia give frequent witness of the medieval and Renaissance contribution to the charm of Syracuse. The best 14th-century facade is that of the Palazzo Montalto. The Palazzi Bellomo and Parisio span the 13th- to 15th-century Gothic styles. Giovanni Vermexio's Palazzo del Municipio (1628) and Luciano Alì's Palazzo Beneventano del Bosco (1775) represent the best among many notable survivals of the 17th- and 18th-century city.

BIBLIOGRAPHY. The history of Syracuse is chiefly dealt with in general histories of Sicily. D. MACK SMITH, *History of Sicily*, 2 vol. (1968), is the standard history of postclassical Sicily. A good, well-informed, pocket guide is M. GUIDO, *Syracuse: A Handbook to its History and Principal Monuments* (1958). See also M.P. LOICQ-BERGER, *Syracuse, historie culturelle d'une cité grecque* (1967). The archaeological remains are most accessibly treated in general works on Greek and Roman architecture. For the most recent excavations, see the successive volumes of *Fasti Archaeologici*. On the Euryalus fort see F.E. WINTER, "The Chronology of the Euryalus Fortress at Syracuse," *American Journal of Archaeology*, 67:363–387 (1963).

(A.G.W.)

Syria

Syria, officially al-Jumhūrīyah al-ʿArabīyah as-Sūrīyah (the Syrian Arab Republic), is a country situated on the east coast of the Mediterranean Sea on the southwestern fringe of the Asian continent. It has an area of 71,498 square miles (185,180 square kilometres), and, in the early 1970s, a population estimated at about 6,300,000. It is bounded by Turkey to the north, by Iraq to the east and southeast, by Jordan and Israel to the south, and by Lebanon to the southwest. The capital is Damascus (Dimashq) on the Baradā River in an oasis at the foot of Jabal (mount) Qāsiyūn. The present territory does not coincide with ancient Syria, which was the fertile strip between the eastern Mediterranean coast and the desert of northern Arabia.

The persistent Middle East crisis between the Arab states and Israel has resulted in political instability in Syria. Since the Arab-Israeli War of 1967, Israel has occupied the 444 square miles (1,150 square kilometres)

of the Golan Heights region in southwestern Syria. Severe strains are placed on the Syrian economy by the demands of national security and defense, which exhaust more than 40 percent of the national budget, leaving little for development. Syria is linked with Libya and Egypt in the Confederation of Arab Republics. (For an associated physical feature, see TIGRIS-EUPHRATES RIVER SYSTEM; see also the city article DAMASCUS; for historical background, see BABYLON; CALIPHATE, EMPIRE OF THE; NEAR EAST, ANCIENT; OTTOMAN EMPIRE AND TURKEY, HISTORY OF THE; and SYRIA AND PALESTINE, HISTORY OF.)

THE LANDSCAPE

Relief features. *The coast.* The relatively short coast stretches for about 110 miles along the Mediterranean between Turkey and Lebanon. Its sandy bays alternate with rocky headlands and low cliffs. North of Ṭarṭūs, the narrow coastal strip is interrupted by spurs of Jabal Alawite (an-Nuṣayrīyah mountains) immediately to the east. It then widens into the Plain of ʿAkkār, which continues south across the Lebanon border.

The mountains. The Jabal Alawite range borders the coastal plain and runs through the country from north to south. The mountains have an average width of 20 miles, and their average height declines from 3,000 feet in the north to 2,000 feet in the south. Their highest point, 5,125 feet (1,562 metres), occurs east of Latakia. Directly to the east of the mountains is the Ghāb Depression, a 40-mile longitudinal trench that contains the valley of the Orontes River.

The Anti-Lebanon Range, or Qalamun Sinnir, marks Syria's border with Lebanon. The main ridge rises to a maximum height of 8,625 feet (2,629 metres) near al-Nabk while the mean height is between 6,000 and 7,000 feet. Mr. Hermon, a single ridge, is connected farther west to the main range by the Zabdānī Saddle. Syria's highest point, it rises to 9,232 feet (2,814 metres).

Smaller mountains are scattered about the country. Among these are Jabal ad-Durūz that rises to 5,905 feet (1,800 metres) in the extreme south; the Palmyra Range that includes the Jabal ash-Sharqī and stretches northeastward across the central part of the country; Jabal az-Zāwīyah that rises to 3,067 feet (935 metres) east of the Ghāb Depression; and Jabal ʿAbd al-ʿAzīz that attains 3,018 feet (920 metres) in the east.

The Syrian Desert. The undulating plains occupying the rest of the country are known as the Syrian Desert. In general their elevation lies between 984 and 1,640 feet; they are seldom less than 820 feet above sea level. The area is not a sand desert but comprises rock and gravel steppe; a mountainous region in the south central area is known as al-Ḥamād.

Drainage. *Underground water.* Most of the country's drainage flows underground. On the surface, impervious rocks consisting of clay, marl (clay, sand, or silt), and green sands, cover a relatively small area. Porous rocks cover about half of the country and are mainly sandstone or chalk. Highly porous rocks consist of basalt and limestone. Water penetrates the porous rocks, forming underground springs, rivers, or subterranean water sheets close to the surface. Although the springs are profuse, the water sheets are quickly exhausted and may turn saline in areas of low precipitation.

Principal rivers

Surface water. The Euphrates River is the most important water source and the only navigable river in Syria. It flows southeastward across the eastern part of the country; the Khābūr, Jaghjaghah, Balīkh, and Sājūr rivers are its tributaries. The coastal rivers are generally short; they include the Qindīl, the northern Kabīr, the Sinn, the Bāniyās, and the southern Kabīr.

The Orontes is the principal river of the mountainous region. It rises in Lebanon, flows northward through the mountains and the Ghāb Depression, and enters the Mediterranean near Antioch, Turkey. The Yarmūk River, a tributary of the Jordan River, drains the Jabal ad-Durūz and Ḥawrān region and forms part of the border with Jordan in the southwest.

Several scattered lakes are to be found. The largest is Mamlaḥat (Sabkhat) al-Jabbūl southeast of Aleppo; a seasonal saline lake, it permanently covers a minimum area of about 60 square miles. Other major salt lakes are Jayrūd to the northeast of Damascus and Khātūnīyah to the northeast of al-Ḥasakah. Lake Muzayrīb is a small body of fresh water to the northwest of Darʿā; the larger Baḥrat Ḥims is a reservoir to the west of Homs.

Soils. Because of aridity, vegetation plays only a secondary role in soil composition. With the exception of the black soil in the northeastern region of al-Jazīrah, soils are deficient in phosphorous and organic matter.

The most common soils are various clays and loams (mixtures of clay, sand, and silt). Some are calcareous (chalky); others, especially in the area of the Euphrates Valley, contain gypsum. Alluvial soils occur mainly in the valleys of the Euphrates and its tributaries and in the Ghāb Depression. (A.-R.H.)

Climate. *Temperature and precipitation.* The coast and the western mountains have a Mediterranean climate with a long dry season from May to October. In the extreme northwest there is some light summer rain. On the coast, summers are hot, with mean daily maxima of 84° F (29° C), while the mild winters have mean daily minima of 50° F (10° C). Only above 5,000 feet are the summers relatively cool. Inland the climate becomes arid, with colder winters and hotter summers. Damascus and Aleppo have average daily maxima of 91 to 99° F (33 to 37° C) in summer and average daily minima of 34 to 40° F (1 to 4° C) in winter. In the desert, at Palmyra (Arabic, Tudmur) and Dayr az-Zawr, average daily maxima in summer reach 99 to 104° F (37 to 40° C), with extremes of up to 114° F (46° C). Snow may occur in winter away from the coast, and frosts are common.

The coast and western mountains receive 30 to 50 inches of rainfall annually. Rainfall decreases rapidly eastward; the steppe receives ten to 20 inches a year. Jabal ad-Durūz receives more than eight inches. In the desert, or al-Ḥamād, rainfall decreases to less than five inches a year. Rainfall is variable from year to year, particularly in the spring and autumn months. (C.G.S.)

The winds. In winter the prevailing winds blow from the east, the north, and the west. In summer, the prevailing winds are either northerly or westerly. During the summer, the coastal region is subject to westerly winds during the day and easterly at night. Once or twice a year sand-bearing winds, or khamsin, raise a wall of dust almost 5,000 feet high, which darkens the sky.

Vegetation and animal life. *Plants.* Yew, lime, and fir trees grow on the mountain slopes. The date palm is found in the Euphrates Valley. In both coastal and inland regions, plants include grains, olives, vines, apricot trees, oaks, and poplars. Lemon and orange trees grow along the coast. Garigue, a degenerate Mediterranean scrub, and maquis, a thick scrubby underbrush, cover many slopes.

Forests

Forests are found in the mountains, especially in the Jabal Alawite. Glossy-leaved and thorny drought-resistant shrubs such as the myrtle, the boxwood, the turpentine, the broom, the arbutus, and the wild olive, abound to the south. Excessive exploitation of the forests for their wood has largely turned them into scrub. Some forests are, however, protected by the government. The steppe is characterized by the absence of natural tree cover, except for some sparsely-distributed hawthorns. All other trees—such as those in the orchards of Damascus and Aleppo and along the banks of the Orontes and Euphrates rivers—are cultivated. For a brief period before June the land is covered with a great variety of flowering and grassy plants such as the asphodel, lily, tulip, hyacinth, daffodil, and cyclamen. Under the implacable sun of June, however, the plants soon wither, casting off their seeds onto the dry ground.

Animals. Wild animal life is sparse. Wolves, hyenas, foxes, badgers, wild boar, and jackals can still be found in remote areas. Deer, bears, squirrels and such small carnivores as the marten and the polecat are also found, while desert animals include the gazelle and the jerboa (a nocturnal jumping rodent). Vipers, lizards, and chameleons are common in the desert. Eagles, buzzards, kites, and falcons frequent the mountains. Among the insects, mos-

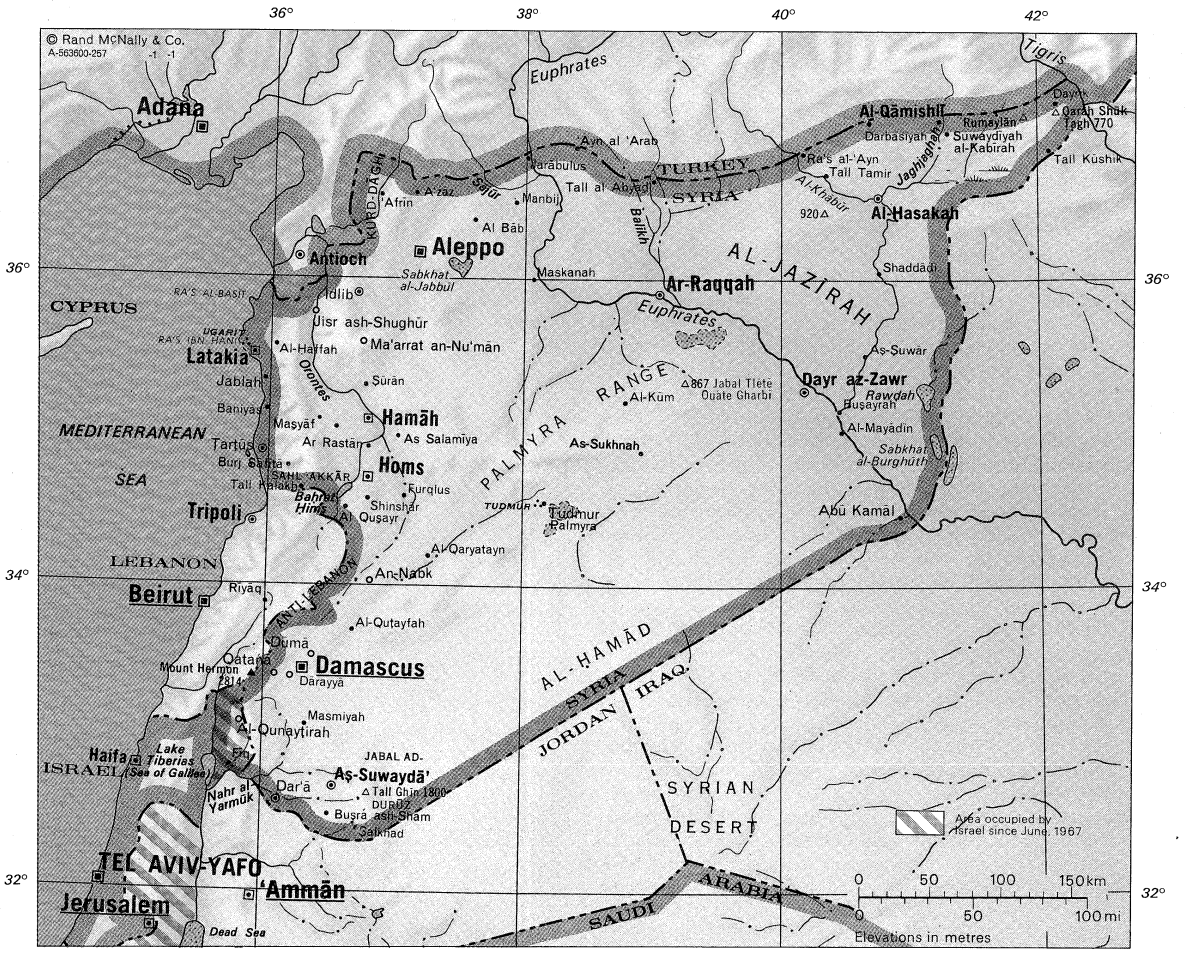
quitoes, sandflies, grasshoppers, and occasionally locusts are harmful.

The mule is the beast of burden in the mountains and the camel on the steppe. Other domesticated animals include horses, donkeys, cattle, sheep, goats, and chickens. Bees are also kept.

Traditional regions. There are four traditional regions —the coastal strip, the mountains, the cultivated steppe, and the desert steppe.

The coast. On the coast, the fertile alluvial plains are intensively cultivated in both summer and winter. The region is the site of Syria's two principal ports of Latakia (al-Lādhiqīyah) and Ṭarṭūs.

The mountains. Jabal Alawite is the only densely forested region. It is the ancient stronghold of the Nuṣayrīs, or ʿAlawites, who form a sect of Shīʿite Islām. The economic resources of the mountains are too meagre to meet the needs of the evergrowing population, and there is a migration to the Ghāb Depression and coastal towns.



SYRIA

MAP INDEX

Cities and towns

Abū Kamāl....34·27n 40·55e
'Afrīn....36·31n 36·52e
al-Bāb....36·22n 37·31e
Aleppo (Ḥalab)...36·12n 37·10e
al-Ḥaffah....35·35n 36·02e
al-Ḥasakah....36·29n 40·45e
al-Kūm....35·11n 38·48e
al-Lādhiqīyah, see Latakia
al-Mayādīn....35·01n 40·27e
al-Qāmishlī....37·02n 41·14e
al-Qaryatayn....34·14n 37·14e
al-Qunayṭirah....33·07n 35·49e
al-Quṣayr....34·31n 36·35e
al-Quṭayfah....33·44n 36·36e
'Āmūdā....37·05n 40·54e
an-Nabk....34·01n 36·44e
ar-Raqqah....35·56n 39·01e
ar-Rastān....34·55n 36·44e
as-Salamīyah....35·01n 37·03e
aṣ-Ṣuwār....35·30n 40·39e
aṣ-Suwaydā'....32·42n 36·34e
A'zāz....36·35n 37·03e
Bāniyās....35·11n 35·57e
Burj Ṣāfitā....34·49n 36·07e
Buṣayrah....35·09n 40·26e
Buṣrá ash-Shām....32·31n 36·29e
Damascus (Dimashq)....33·30n 36·18e
Dar'ā....32·37n 36·06e
Dārayyā....33·27n 36·15e
Darbāsīyah....37·04n 40·39e
Dayr az-Zawr....35·20n 40·09e
Dayrik....37·10n 42·08e
Dimashq, see Damascus
Dūmā....33·35n 36·24e
Fīq....32·47n 35·42e
Furqlus....34·36n 37·05e
Ḥalab, see Aleppo
Ḥamāh....35·08n 36·45e
Ḥoms (Ḥimṣ)....34·44n 36·43e
Idlib....35·55n 36·38e
Jablah....35·21n 35·55e
Jarābulus....36·49n 38·01e
Jisr ash-Shughūr....35·48n 36·19e
Latakia (al-Lādhiqīyah)....35·31n 35·47e
Ma'arrat an-Nu'mān....35·38n 36·40e
Manbij....36·31n 37·57e
Maskanah....36·01n 38·05e
Masmiyah....33·03n 36·20e
Maṣyāf....35·03n 36·21e
Palmyra, see Tudmur
Qaṭanā....33·26n 36·05e
Ra's al-'Ayn....36·51n 40·04e
Salkhad....32·29n 36·43e
Shaddādī....36·02n 40·45e
Shinshār....34·36n 36·44e
Suknah....34·53n 38·55e
Sūrān....35·17n 36·45e
Suwaydīyah al-Kabīrah....36·55n 41·19e
Tall al-Abyad....36·41n 38·57e
Tall Kalakh....34·40n 36·15e
Tall Kūshik....36·48n 42·04e
Tall Tamir....36·39n 40·22e
Ṭarṭūs....34·53n 35·53e
Tudmur (Palmyra)....34·33n 38·17e

Physical features and points of interest

'Akkār, Sahl, *plain*....34·39n 36·03e
al-'Āṣī, Nahr, see Orontes
al-Furāt, see Euphrates
al-Ḥamād, *plain*..33·25n 38·30e
al-Jazīrah, *physical region*..35·45n 40·30e
al-Khābūr, *river*..35·08n 40·26e
Anti-Lebanon (al-Jabal ash-Sharqī), *mountains*....34·00n 36·30e
Balīkh, *river*....35·53n 39·10e
Basīt, Ra's al-, *point*....35·51n 35·48e
Burghūth, Sabkhat al-, *salt lake*....34·58n 41·06e
Dijlah, see Tigris
Durūz, Jabal ad-, *mountain*....32·40n 36·44e
Euphrates (al-Furāt), *river*....34·25n 41·00e
Ghīn, Tall, *mountain*....32·39n 36·43e
Hermon, Mount..33·26n 35·51e
Ḥimṣ, Baḥrat, *lake*....34·39n 36·34e
Ibn Hānī'....35·35n 35·43e
Jabbūl, Sabkhat al-, *salt lake*....36·03n 37·39e
Jaghjaghah, *river*....36·29n 40·45e
Kurd Dāgh, *mountains*....36·40n 36·46e
Mediterranean Sea....35·00n 35·00e
Orontes (Nahr al-'Āṣī), *river*....35·30n 36·18e
Palmyra Range, *mountain range*..34·50n 38·20e
Qarah Shūk Ṭāgh, *mountain*..37·03n 42·10e
Rasm al-Arwām, Sabkhat, *salt lake*....35·53n 37·40e
Rawdah, *salt lake*..35·15n 41·05e
Rumaylān, *mountain*....37·01n 41·58e
Sājūr, *river*....36·40n 38·05e
Sharqī, al-Jabal ash-, see Anti-Lebanon
Tigris (Dijlah), *river*....37·12n 42·20e
Tlētē Ouāte Gharbī, Jabal, *mountain*....35·20n 39·13e
Tudmur, *historic site*....34·33n 38·17e
Ugarit, *historic site*....35·35n 35·45e
Yarmūk, Nahr al-, *river*....32·38n 35·34e

The cultivated steppe. This region constitutes the principal wheat zone; agriculture is fairly intensively pursued along the banks of the rivers. The villages are characterized by domelike huts which are built of mud because of the scarcity of wood and stone; the few stone and cement houses are owned by the wealthy. Some of Syria's most important cities—Damascus, Aleppo, Homs, Ḥamāh, and al-Qāmishlī—are situated here.

The desert steppe. The arid country is the natural domain of the nomads and seminomads. Sheep graze until the beginning of summer when water becomes scarce, after which the shepherds lead their flocks either westward into the cultivated steppe or to the hills. The area once contained oases which served as caravan towns on the trade route joining Mesopotamia and the Indian Ocean with the countries of the Mediterranean. The most famous of these towns is Palmyra at the northern edge of the Syrian Desert. The most important feature of the region is the Euphrates River.

The nomadic domain

Patterns of settlement. *Rural settlement.* The choice of a Syrian village site is usually determined by the availability of water. Some of the Alawite Mountain villages, however, have given priority to the requirements of defense and fortification. Village dwellings are built close together, and village streets are extremely narrow. Usually, there is a small central common overlooked by a minaret (a tall tower attached to a mosque from which the populace is called to prayer). There are usually a few small shops, containing articles manufactured in the cities or towns. Travellers are accommodated by the village headman according to local custom.

Work takes place according to the seasonal rhythm of agriculture. Women generally share in all agricultural labour except plowing. Agricultural machinery, introduced on a large scale after World War II, caused unemployment and drove numbers of the villagers to the cities.

Urban settlement. Ten centuries of Greek and Roman rule left an urban mark still to be seen in the towns of Latakia, Palmyra, and Buṣrā. The urban tradition of Islām is represented by such cities as Damascus and Aleppo. The continuation of old commercial and religious interests has enabled the cities to maintain their economic and cultural supremacy under the four centuries of Ottoman rule. The great abyss between the rural and urban areas seems to preclude any economic or social balance. Today, however, the cities are inundated by rural immigrants, and unemployment is rife.

The cities

The national capital of Damascus, situated in the southeast on the banks of the Baradā River, is the largest city in Syria with an estimated population in the early 1970s of more than 835,000. It is the national headquarters of government and the diplomatic community, and is the main centre of education, culture, and industry. It also serves as the marketing centre for mid-Syria and produces traditional handicraft products such as brocades, engraved wood, gold and silver ornaments, and carpets. It is well served by transport facilities and public utilities. The old houses in Damascus are built of soft unbaked bricks, wood, and stone. Contemporary buildings are built of concrete, while hewn stone is reserved for official buildings, mosques, and churches.

Located between the Orontes and Euphrates rivers, Aleppo is a trade and light industry centre of almost 639,000 inhabitants. Its limestone buildings have given the city its name of Halab ash-Shahbā' ("Aleppo the White"). The city is well served by roads and railroads and is surrounded by an area that specializes in the production of sheep for market in Damascus and other countries. The Mediterranean port of Latakia has a population of about 126,000. It is surrounded by a rich agricultural region and contains some industry. Because of its seaside location, the city is a major tourist centre.

Homs is located in the midst of a fertile plain east of the Orontes River. With a population of almost 216,000, it is a hub of the country's road and railway systems. Ḥamāh, to the northeast of Homs, is bisected by the Orontes River and has a population of about 137,000. It contains irrigated orchards and is an agricultural trade centre. There is also some light industry.

PEOPLE AND POPULATION

The Syrian people evolved from several origins over a long period of time. The Greek and Roman ethnic influence was negligible in comparison with that of the Semitic peoples of Arabia and Mesopotamia—Aramaeans, Assyrians, Chaldeans, and Canaanites. Later the Turks, like the Greeks and Romans before them, influenced political and economic structures, but failed to produce any noticeable change in the dominant Arab character of the Syrian people.

Population groups. *Linguistic groups.* The mother tongue of about 80 percent of the population is Arabic. According to their numerical importance the other languages are: Kurdish, spoken in the extreme northeast and northwest; Turkish, spoken in villages east of the Euphrates and along the border with Turkey; Armenian, spoken in Aleppo and other major cities; Circassian, spoken in a number of Golan district villages, east of Homs and Aleppo, and in the town of Manbij; Syriac, spoken in two villages near Damascus and in the northeastern town of al-Qāmishlī; and Hebrew, used as the liturgical language of the small Jewish community. English is taught as a foreign language to more than 70 percent of the school children; others are taught French, German, or Russian.

Ethnic groups. There is a rough correspondence between ethnic and linguistic groupings, although some ethnic groups have been partially assimilated by the Arab majority that includes the Bedouin. Second in number to the Arabs, the Kurds have partially lost their mother tongue. The Turkmen intermingle freely with the Kurds and Arabs, but have lost none of their ethnic identity in some northern villages. The Armenians may be divided into two groups—the early settlers, who have been more or less Arabized; and the later immigrants, who came after World War I and retained their identity and language. The Assyrians are quickly disappearing as a group because of intermarriage and migration to the cities. Most of the Jews have emigrated to Israel; the remaining few thousand live in Damascus or Aleppo.

Religious groups. Islām, the predominant religion, is practiced by about 88 percent of the population. The Muslims are divided into several sects. The Sunnites number about 72 percent of the Muslim population and are in the majority everywhere except in as-Suwaydā' *muḥāfaẓah* (governorate) in the south and the Latakia *muḥāfaẓah* in the north. The ʿAlawites number about 600,000, three-fourths of whom live in the Latakia *muḥāfaẓah*, the rest residing mainly in the *muḥāfaẓāt* of Homs and Ḥamāh. The Druzes (*q.v.*) number about 190,000, of whom some 8 percent live in as-Suwaydā' *muḥāfaẓah* and the rest in Damascus, Aleppo, and al-Qunayṭirah.

The Muslim sects

The Ismāʿīlites number about 60,000; over 80 percent live in as-Salamīyah and around Ḥamāh, and the rest in the Alawite mountains. The Shīʿites number about 25,000 and are found mainly in Damascus, Aleppo, and Idlib. The Yezidis are a small group, and many of them have adopted the Sunnite rite.

Christians constitute about 12 percent of the Syrian people. They are divided into 11 churches, listed here according to their numerical importance: Greek Orthodox, Armenian Orthodox, Greek Catholic, Syrian Orthodox, Armenian Catholic, Syrian Catholic, Maronite, Protestant, Nestorian, Latin, and Chaldean. Jews numbered about 20,000 in the early 1970s.

Demography. *Population growth.* The population was estimated at about 6,500,000 in 1971. The figure included some 400,000 Syrians who live abroad but retain their citizenship. The average annual rate of growth in population was 32 per 1,000 in the period 1965–70. At this rate the population will double in 22 years.

More males are born than females. About 10 percent of the infants in rural areas die in their first year; this rate drops considerably in the towns. Individual life expectancy varies on the average between 50 and 55 years.

Immigration and emigration. The last wave of immigration consisted of 80,000 Palestinian Arabs who fled to Syria in 1948. Increasing numbers of workers and profes-

sionals have been leaving the country since the 1950s for other Arab nations, the United States, and western Europe. This movement has caused an alarming drain of Syrian manpower, and has reduced the number of doctors and engineers to one-third of the original total.

Distribution of the population. The overall population density is 90 per square mile. Only about half the country's land can support population, however, and in addition about 44 percent of the population lives in the cities; the density in the arable areas is consequently about 98 per square mile. The desert steppe contains 150,000 Bedouins and 25,000 oasis inhabitants; the area's population density is less than five persons per square mile. In the rest of the country the density, which varies considerably, is highest in the west and northeast, and is fairly high along the banks of the major rivers.

Syria's rate of population increase is out of all proportion to the rate of economic expansion. Family planning and birth control are becoming popular among the educated, despite an official policy of awarding bonuses and allowances to expanding families.

THE NATIONAL ECONOMY

Syria is favourably located on the regional trade routes running between the Mediterranean and the countries of Mesopotamia and the Persian Gulf, and between Asia Minor and the Arabian Peninsula. Commercial aviation routes and oil pipelines cross Syrian territory.

Oil and natural gas reserves

The extent and distribution of resources. *Mineral resources.* Petroleum in commercial quantities was first discovered in the northeast in 1956. Workable reserves are estimated at 260,000,000 tons and potential reserves at 1,200,000,000 tons. The most important oil field is that of Suwaydīyah, followed by those of Qarah Shūk and Rumaylān. The fields are a natural extension of the Iraqi fields of Mosul and Kirkūk. Natural gas was discovered at the field of Jbessa in 1940. Its reserves are estimated at 115,284,000,000 cubic feet.

Raw phosphates were discovered in 1962; reserves are estimated at 240,000,000 tons in the four richest beds at Khunayfis, Ghadīr al-Jamal, Ṣawwānah, and Wādī ar-Rakhim. About 500,000,000 tons of rock salt lie to the west of Dayr az-Zawr at depths ranging from about 300 to 600 feet.

Iron ore at Kurd Dāgh in the northeast is estimated at 62,000,000 tons. Asphalt has been found northeast of Latakia, and west of Dayr az-Zawr, and table salt is provided by the salt lakes of al-Jabbūl, Palmyra, and al-Buwarah. Syria is rich in limestone, basalt, and marl, and has scattered reserves of chrome and manganese.

Biological resources. Forests occupy about 5 percent of the country's total area. Systematic economic exploitation is hampered by the wide variety of trees and the fact that no one kind is widespread enough to be utilized on an industrial scale. The commercially important forest plants include sumac, which is used as a spice; the wild pistachio, which is important for its oil-rich fruit; laurel, which is used in the production of cosmetics; and the mulberry, whose leaves are fed to silkworms. Pine tree branches are used for smoking tobacco leaves. Other useful plants are such winter vegetables as *khubbayzah*, a kind of spinach; *ʿakkūb*, a flowering plant; and the truffle. Licorice is widely exploited for its root, which is used in the pharmaceutical industry.

Game animals are scarce except for the wild boar and the hare. Game fowl are available in fairly large numbers, and include wild ducks and pigeons.

Power resources. Local oil supplies constitute a major power resource. With the exception of the Euphrates River, which has a considerable hydroelectric potential, other rivers flowing through Syria are capable of producing only small amounts of hydroelectric power.

Sources of national income. *Agriculture, forestry, and fishing.* Agriculture constitutes the major source of income. It provides work for nearly half the population, including a large proportion of townspeople. In 1969, there were about 3,000 fishermen using small and medium-sized boats. The fish catch was over 1,200 metric tons in 1969. Forestry workers did not exceed a few hundred in the early 1970s.

The dominance of agriculture

Wheat is the most important food crop, although its production is constantly subject to great fluctuations in rainfall. Barley, maize (corn), and millet are the other important grains. Cotton is the largest and most reliable export crop. Lentils are a major domestic food, but are also exported. Other vegetables include onions, potatoes, and sugar beets. Olives, grapes, and apples are grown at high altitudes, while citrus fruits are cultivated along the coast. High-grade tobacco is grown in the area around Latakia.

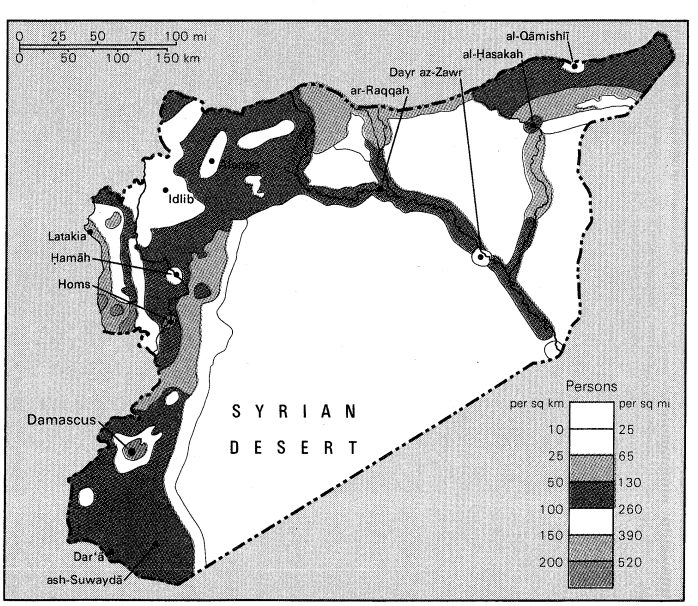
Numbers of livestock are raised, including sheep, cattle, camels, and poultry. The mountains north of Latakia are being reforested. Elsewhere, poplar trees are grown along irrigation ditches and are used for building construction, although most of the nation's timber has to be imported. The annual fish catch ranging from 600 to 1,200 metric tons includes sardines, tuna, groupers, tunny, and both red and grey mullet, but the lack of refrigerated transport and storage facilities have prevented possible development of inland markets.

Mining and quarrying. The principal limestone quarries are located north and west of Damascus, especially near the city of Aleppo, which itself is built of limestone. Basalt, used for road pavement, is quarried in Jabal ad-Durūz and near Homs and Aleppo. Marl is used in the cement industry; the main quarries are in the vicinity of Damascus and Aleppo, and at ar-Rastān. Sandstone suitable for glass manufacture is mined in the Palmyra mountains. Phosphate ore is mined in Jabal ash-Sharqī and rock salt is extracted from the mid-Euphrates region. Asphalt and gypsum are also mined, and table salt is produced from the salt lakes.

Manufacturing. Most of the traditional handmade manufactures—damask steel, swords and blades, brass and copper work, wood engravings, gold and silver ornaments, mother-of-pearl inlays, silk brocades—have disappeared since the introduction of industrial processing.

Wool, cotton, and nylon textiles are the most important manufactures. The mills are mainly in Aleppo, Damascus, Homs, and Ḥamāh. Next in importance are the technical engineering industries, most of which are in Damascus. The chemical and industrial engineering products include cement; glass panes, bottles, and utensils; pharmaceuticals; plywood; and batteries. The food processing industry produces salt, vegetable oils, cotton cake, canned fruit and vegetables, tobacco, and dairy products. There is also an oil refinery at Homs that supplies local needs. Other industries include a fertilizer plant, and distilleries.

Energy. Muscular energy—both animal and human—is still used in operating oil-extracting mills and in the



Population density of Syria.

production of pottery. Modern factories, however, are run by electricity, supplied either by state-owned generating plants or by factory generators. Gas is used on a minor scale in small industries such as goldsmithing, but its supply is limited. Coal is imported and is also used on a limited scale, mainly by blacksmiths.

Electric power

Electricity, largely generated by thermal stations, is used mainly for lighting. Thermoelectric generation depends upon the fuel oil produced by the oil refinery at Homs. There are small hydroelectric stations such as those on the Orontes and Baradā rivers. A large increase in production is anticipated when the Euphrates River project is completed in 1974. The dam will have a generating power of 800,000 kilowatts, and a potential capacity of 1,100,000 kilowatts.

Financial services. The Central Bank of Syria issues the national currency and exercises control over all other banks that operate in the country. The Commercial Bank of Syria finances trade, markets agricultural products, and carries out foreign exchange operations. The Real Estate Bank finances the building industry and carries out all ordinary banking operations. An Industrial Bank finances the private industrial sector, while an Agricultural Bank extends loans to farmers and agricultural cooperatives. A Popular Discount Bank makes loans to small manufacturers, artisans, and production cooperatives. There is one semi-official insurance company and a stock exchange connected with the central banks.

Foreign trade. Syria has an unfavourable balance of trade. The deficit is made good from tourism, transit trade returns, oil pipeline royalties, and overseas earnings.

Imports include industrial and agricultural machinery, motor vehicles and accessories, drugs, food, and fabric. Exports consist of ginned cotton, petroleum, cotton seeds, barley, lentils, cotton and woolen fabrics, dried fruit, olive oil, vegetables, skins, and raw wool. Foreign trade is controlled through a special state central agency, Syrian Import-Export (Simex).

Management of the economy. Socialism became the official economic policy in 1958. Since then, the trend has been toward Socialist transformation and industrialization. In commerce, state control is mainly restricted to foreign exchange operations.

The private sector. Small private businesses and cooperatives are still in operation, and the retail trade is still part of the private sector, despite competition from consumer cooperatives in the large cities.

The public sector. The government controls the most vital sections of the country's economy, and largely regulates private business. The state operates the oil refinery, the large electricity plants, the railways, and various manufacturing plants.

Taxation. Indirect taxes, which produce the most tax revenue, are levied on industrial products, customs, exports, and state domains. Direct taxes are levied on wages, circulating capital, livestock, and the transfer of property.

The General Federation of Workers

Trade unions and employer associations. The General Federation of Workers, with which almost 200 unions are affiliated, was founded in 1938, and in the early 1960s grew tremendously in power and scope. Composed only of industrial employees, it is represented on all industrial boards and is responsible for a wide range of social services and benefits for its members and all workers in general. There is also a federation for artisans and vocational workers, and there are also associations for the professions and a General Federation of Farmers.

Employers are not permitted to run businesses employing more than ten workers. An employers' association is virtually nonexistent.

Contemporary economic policies. The government encourages private savings by paying higher rates of interest on deposits, and by guaranteeing investment by citizens of other Arab countries. There are severe restrictions on all luxury imports, including automobiles. At the same time strenuous efforts are made to mobilize economic potential, combat unemployment, and discourage emigration. Financial and technical assistance, which is offered free or at minimum interest rates, is usually sought from Socialist countries such as China, Bulgaria, and the Soviet Union. Aid, at higher interest rates, is also sought from France and other capitalist countries.

Problems and prospects. The most serious national problem is the financial drain caused by the demands of national security and defense arising from the Arab-Israeli conflict. There is also a severe loss of professional, technical, and manual labour to Lebanon, to the oil-rich Arab countries, to the United States, and to western Europe. Purchasing power, generally low, discourages prospective investors. It is anticipated that the formation of an Arab common market will provide buyers for Syria's produce and stimulate economic expansion. The Euphrates River Project is expected to promote both industrial and agricultural growth.

TRANSPORTATION

Railways. Three major railways were built before World War I under the Ottoman Empire. A northern line runs northeastward from Aleppo into Turkey and then east along the border to al-Qāmishlī, where it crosses the northeastern extremity of Syria en route to Baghdad (Iraq). The second (the Hejaz Railway) runs from Damascus to Maʿān (Jordan), and a third from Aleppo to Tripoli (Lebanon) and to other towns on the Lebanese coast. Smaller lines run between Homs and Rīyāq (Lebanon), and between Beirut (Lebanon) and Damascus. In 1967, the Syrian government began the construction of a railway from Latakia to Idlib, Aleppo, ar-Raqqah, Dayr az-Zawr, and al-Ḥasakah, passing by the Euphrates River Dam.

Roads. Paved roads were first constructed for military use before World War I. The road network has continued to expand and improve everywhere except in the arid region in the east. By 1971 there were over 7,100 miles of roads, of which about 5,100 miles were paved or gravel. The best road is a 225-mile highway between Damascus and Aleppo. A paved road is under construction between Damascus and Baghdad, Iraq.

Ports. The two ports, Latakia and Ṭarṭūs, were built after independence. Construction at Latakia was begun in 1952; it has two main jetties, wharves, and warehouses. The port of Ṭarṭūs is also equipped with two jetties and wharves. Its basin is larger than those of Latakia and Beirut, Lebanon, combined. Baniyas is used to ship oil from Iraq.

Air transport. Syria has international airports at Damascus and Aleppo and domestic airports at al-Qāmishlī, Latakia, Dayr az-Zawr, and Palmyra (Tudmur). Only the facilities at Damascus and Aleppo can handle jet aircraft. International services connect Syria with the Arab and east European countries, France, Italy, Pakistan, India, and the United Kingdom. Domestic services are provided by Syrian Arab Airways.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *Constitutional framework.* The constitution declares that Syria constitutes an integral part of the Arab homeland, that all legislative power lies with the people, and that freedom of expression and equality before the law is guaranteed. The enforcement of these principles has not been thorough, however, because of constant political and military tensions.

The regional (Syrian) leadership of the Baʿth Arab Socialist Party (see *The political process* below) elects the head of state and appoints the cabinet, which exercises legislative as well as executive powers. Suffrage is not universal, and elections take place within each of the various popular and labour unions.

Local government. Syria is divided into 13 *muḥafaẓāt* (governorates), one province-level city, 47 *manāṭiq* (districts), and 120 *nawāḥī* (subdistricts). The governors, or *muḥāfiẓīn*, enjoy some power within their administrative divisions, but local government is centralized and is dependent upon the minister of the interior.

The Baʿth Party

The political process. Most authority is wielded by the ruling Baʿth ("Renaissance") Arab Socialist Party, the only legally recognized political party in the country.

Syria, Area and Population

	area		population	
	sq mi	sq km	1960 census	1970 census*
Governates (*muḥāfaẓāt*)				
al-Ḥasakah	9,024	23,371	353,000	467,000
al-Lādhiqīyah	941	2,437	292,000	389,000
al-Qunayṭirah	660	1,710	106,000	16,000
ar-Raqqah	8,494	22,000	178,000	243,000
as-Suwaydā'	2,143	5,550	100,000	139,000
Dar'ā	1,440	3,730	140,000	233,000
Dayr az-Zawr	12,764	33,060	221,000	292,000
Dimashq	7,061†	18,287†	368,000	621,000
Ḥalab	6,232	16,142	957,000	1,316,000
Ḥamāh	3,415	8,844	370,000	512,000
Ḥimṣ	16,303	42,226	401,000	546,000
Idlib	2,291	5,933	301,000	383,000
Tartous	730	1,890	221,000	302,000
Municipality (*madīnat*)				
Dimashq (Damascus)	†	†	557,000	835,000
Total Syria	71,498‡	185,180	4,565,000§	6,294,000§

*Preliminary figures. †Area for Madīnat Dimashq is included in the area for Muḥāfaẓah Dimashq. ‡Converted area figures do not add to total given because of rounding. §Including Palestinian refugees.
Source: Official government figures.

Since its foundation in 1941, the party has undergone radical internal changes as a result of successive coups d'état and internal power struggles. The party has branch organizations in all Arab countries, each headed by its own regional leadership. The organs of administration are the National Command, the Regional Command, and the People's Council; the latter operates as a legislature. The supreme national leadership is composed of representatives from each regional branch, who are elected by their own party congresses. The regional leadership for Syria is the highest authority in the country, but is subordinate to the national leadership.

Though all other political parties are officially banned, several manage to operate under varying measures of toleration. These include the Communists, the Arab Nationalists, the Arab Socialists, the Muslim Brotherhood, and the Syrian Nationalists.

All participation in the political life of the country is channelled through the particular popular organization to which a citizen may belong. In addition to the unions mentioned earlier, there is a General Union of Students, a Revolutionary Youth Union, and a General Federation of Women.

Justice. The principles of Syrian law and equity derive basically from Islāmic jurisprudence, and secondarily from the French civil code. The headquarters of each administrative district has a First Instance Court for criminal cases and another for private disputes. The capital city of each governorate also has a criminal court, a court of labour relations, a court of common pleas, a juvenile court, and a court of summary jurisdiction. Damascus houses a state tribunal with jurisdiction over claims raised against governmental agencies, as well as a military tribunal, and the *muftī*'s court for the maintenance of Islāmic law. The principal provincial towns also have courts of agrarian relations to settle litigation arising between landowners and labourers.

The armed forces. Military service is compulsory for all adult males over 18 years of age; only sons are exempted, and college students receive deferments. Military service lasts for 30 months, during which time general and technical—as well as military—education and training are given. The army is responsible for defense, public works, road construction, public health, and controlling the mass media. The air force, equipped with Soviet assistance, has a substantial number of jet combat aircraft. A small navy consists of a few minesweepers, patrol vessels, and torpedo boats, and surface to surface missile patrol boats. Arab guerrilla organizations operate from Syria and receive training facilities, but do not come under the control of the Syrian government.

Social conditions and services. *Educational services.* The literacy rate is about 50 percent. Schooling is compulsory for all children above six years of age, and all school books are free. In the early 1970s there were over 5,000 elementary schools; elementary education lasts for six years. In 1970 it was announced that all preparatory and secondary schools would be open to all elementary students who wished to continue their education. Within this framework, increased attention is being given to technical education, which is given to about 15 percent of those who finish the three years of preparatory school. There are about 38,000 university students, most of whom attend the universities of Damascus and Aleppo. A university was founded in Latakia in 1971, and about 10,000 students attend universities abroad. There is a shortage of facilities for scientific training; about three-fourths of Syrian university students study the arts, the humanities, law, and similar disciplines. The majority of those who go abroad specialize in medicine, engineering, and various branches of technology.

Higher education

Health and welfare services. Except for typhoid, malaria, and dysentery, there are no serious endemic diseases in Syria. Excluding the hospitals and out-patient clinics of the armed forces, some 32 state and 49 private hospitals and sanatariums provide a combined ratio of about one bed for every 1,000 citizens. There are also a number of public and private out-patient clinics, maternal and child-care centres, antituberculosis centres, malaria eradication centres, and rural health centres.

Child mortality, relatively high, is mostly caused by measles and diseases of the digestive and respiratory systems. Tuberculosis is widespread, particularly among the Bedouin, the peasants, and residents of city slums.

The Ministry of Social Welfare and Labour is empowered to find work for, and distribute cash allowances to, the unemployed. The ministry also encourages such youth activities as sports, athletics, scouting, literacy campaigns, and the organization of cooperatives. The government gives substantial grants to private welfare societies.

Housing. The shortage of housing is acute because of the pace of change from an agricultural to an industrial economy, and the accompanying migration to the cities. To further aggravate the shortage, young adult males are breaking with tradition and leaving their parental homes for their own houses.

The Ministry of Municipal and Rural Affairs undertakes the construction of blocks of low-income flats in the cities.

Wages and the cost of living. Labour legislation establishes minimum wage limits, prohibits child labour, and organizes relations between workers and employers. But economic and social conditions as well as the extent of unemployment make rigorous enforcement impractical. Employees in heavy industry receive the highest industrial wages, textile workers the lowest. State employees have more job security. Up to 75 percent of the average salary is generally spent on housing and food, leaving little for clothing, medicine, and recreation.

Health conditions. Health conditions and sanitation in the cities, towns, and larger villages are generally satisfactory. Running water is supplied to almost all houses, buildings, and public places. Each municipality maintains its streets and collects refuse regularly.

Medical services are badly distributed; three-quarters of the doctors are concentrated in the large cities. In the rural districts, health conditions are unsatisfactory because of polluted water and the lack of drainage and sewage systems.

Social and economic divisions. The upper-income group is composed of the owners of small industrial businesses, important merchants, and senior officials. The middle-income group comprises the great majority of shopkeepers, professionals, and farmers. Workers, peasants (farmers who do not necessarily own the land they cultivate) and most employees—the remaining 75 percent of the population—belong to the lower-income group.

The process of Socialist transformation has caused much social, political, and economic confusion, and has led to emigration among the rich. The rising birth rate has caused family lands to be broken up into ever smaller

lots, and has reduced the standard of living of the poor, forcing some of them also to emigrate.

CULTURAL LIFE AND INSTITUTIONS

The cultural milieu. The long periods of foreign, and particularly Ottoman, rule contributed little to the development of a national culture. Education was confined to a few, who maintained the country's ancient libraries and manuscripts. After independence in 1946, a new cultural life began to stir on a humble scale, mostly through the medium of a small number of clubs that received some financial assistance from the state.

In 1958 the Ministry of Culture and National Guidance was created, and has since been active in directing and promoting the nation's cultural life. An important objective has been the affirmation of the Arab national character in the face of foreign cultural influences.

The state of the various arts. Until World War I, the artistic representation of animal or human life was proscribed by Islām. Figurative art was restricted to geometric designs as manifest in the arts of arabesque and calligraphy, which decorates most palaces and mosques.

The fine arts

Following World War I drawing was taught in the schools and talented artists began to emerge. A faculty of fine arts was instituted at the University of Damascus in 1963. Some of the teachers, such as Fateh Moudarres, are well-known artists. Syrian artists are concerned more with content than with form or style.

Sculpture is mainly confined to decorations hewn in white marble. Damascus is particularly famous for this type of sculpture, and beautiful examples of it can be seen in its mosques, palaces, and public buildings.

National folk traditions have been emphasized by the state, which has established a wax-work museum for national folk traditions in Damascus. One of the oldest forms of folklore—the epic—continued in the form of recitations about the exploits of legendary characters. The National Theatre and other theatrical and folk-dance companies give regular performances.

Cultural institutions. The Ministry of Culture has established an Arab institute of music and opened numerous courses in the figurative and applied arts, as well as centres for teaching the domestic arts. The Arabic Language Academy, founded in 1919 and the oldest such academy in the Arab world, runs the Ẓāhirīyah Library, well-known for its rare manuscripts and old book editions. The University of Damascus, however, is the most important of Syria's cultural centres.

Press and broadcasting. The history of Syria's newspapers has been one marked by suppression. There are now four daily papers, two of which are published in Damascus. In addition, there are two weeklies run by the federations of workers and of peasants. All newspapers are under direct government control, and the magazines and journals are run mostly by official or semi-official bodies.

Syrian broadcasting began in 1947 and has since become the most powerful rival of the local press. Arabic transmission by the Syrian Broadcasting Service extends over 17 hours a day. Short programs are also transmitted in English, French, Turkish, and German. Broadcasts—many of which are devoted to instruction—reach almost every Syrian home.

The Syrian Television Service, started in 1960, reaches a large audience throughout the country. Programs are transmitted for seven hours a day, with additional hours at weekends and during holidays. Broadcasting includes educational and cultural programs, drama, music, news, and sports.

PROSPECTS FOR THE FUTURE

Syria's mild Mediterranean climate and its favourable position on regional trade routes should enable its people to prosper. Complex political circumstances, however, and impoverishment have retarded its progress. In recent years the state of war, declared or undeclared, with Israel has had profound repercussions. Political and military tensions have checked the development of the democratic process, and have diverted resources into channels that have little relationship to its economic and social problems. The most pressing of these problems is the rapid increase in the population, which is out of proportion to the rate of economic growth. In consequence it is estimated that one-third of men of working age and almost all women have no paid employment and are driven to emigrate in ever increasing numbers. Syria is nevertheless making determined efforts to plan and carry out a number of industrial, agricultural, and social projects, some of which are vital to its future growth.

BIBLIOGRAPHY

General works: P. BIROT and J. DRESCH, *La Méditerranée et la Moyen-Orient*, 2 vol. (1953–56); W.B. FISHER, *The Middle East*, 4th ed. rev. (1961); D. WARRINER, *Land and Poverty in the Middle East* (1948); J. WEULERSSE, *Paysans de Syrie et du Proche-Orient* (1946); EUGEN WIRTH, *Syrien* (1971).

Physical geography: C. COMBIER, *Aperçu sur les climats de la Syrie et du Liban* (1945); N. MOUSSLY, *Le Problème de l'eau en Syrie* (1951).

Regional geography: A. BAGH, *Le Djolan* (1958); A.R. HAMIDE, *La Région d'Alep* (1959); J. WEULERSSE, *Le Pays des Alaouites* (1940).

(A.-R.H.)

Syria and Palestine, History of

The geographical area covered by this article extends from the foothills of the Anatolian Plateau in the north to the Sinai Desert in the south and from the Mediterranean coast on the west to the Euphrates River on the east (including the present states of Israel, Jordan, Syria, and Lebanon). The greater part of this large area, however, consists of the Syrian Desert and plays little part in history except as a reservoir of nomads who, at intervals, erupted into the surrounding settled lands. The settlements that provide the units for a historical picture of ancient times form a semicircle called the Fertile Crescent, sweeping around from the southeastern Mediterranean coast to the river valleys of Mesopotamia. This area and its fringes are of crucial importance, for it is here that the first steps were taken that led to the development of civilization.

The Fertile Crescent

This article is divided into the following sections:

I. Early history, to *c.* 1600
- Stone Age cultures
  - Paleolithic and Mesolithic
  - Pre-pottery Neolithic
  - Pottery Neolithic
- The end of the Neolithic and the dawn of civilization
  - The Chalcolithic
  - The beginnings of towns and city-states
- The Early and Middle Bronze Ages
  - Early Bronze Age cities
  - The Amorite invasion
  - Middle Bronze Age

II. Syria and Palestine, *c.* 1550 BC–AD 634
- The era of the Egyptian and Hittite empires, *c.* 1550–1200 BC
  - The struggle of the Great Powers for Syria and Palestine
  - Organization and culture
  - The coming of the Hebrews
  - The Sea Peoples and the advent of the Philistines
- Syria and Palestine, *c.* 1200–*c.* 330 BC
  - The Philistines
  - The Syro-Hittite states and the Aramaeans
  - The Phoenicians
  - The Hebrews
  - The era of the Assyrian, Neo-Babylonian, and Achaemenian empires
- Syria and Palestine in the Hellenistic and Roman periods, *c.* 330 BC–AD 634
  - The Macedonian conquest
  - The Hasmonean state in Palestine
  - The Romans in Syria and Palestine
  - Syria and Palestine in the late Roman and Byzantine periods, *c.* 300–634

III. From the Islāmic conquest to 1920
- Medieval period
  - The conquest and its sequence
  - From the 8th to the 12th century
  - The Ayyūbids and Mamlūks
- Ottoman period

Ottoman government, 16th–17th centuries
Decline of Ottoman authority
Egyptian domination
Ottoman rule restored
World War I
IV. Syria and Palestine since 1920
Syria and Lebanon under French mandate, 1920–41
The Druze revolt
Establishment of the Lebanese Republic
Constitutional developments in Syria
Problems and crises in the Lebanese Republic
Palestine and Transjordan under British mandate, 1920–48
The Balfour Declaration
The Palestine revolt and the Peel Commission and later commissions
The achievement of Syrian and Lebanese independence
Syria, 1941–43
Lebanon, 1941–43
The partition of Palestine and the emergence of Israel
Pressures for a Jewish state
The Palestine war
Syria after independence
The colonels
The union with Egypt, 1958–61
The "secessionist" regime, 1961–63
Ba'thist Syria after 1963
Lebanon after independence
The Khuri regime, 1943–52
The Chamoun regime and the 1958 crisis
The Chehab and Hélou regimes, 1958–70
Israel and the Arab states
Jordan and the Palestine question after 1949
Israel after 1948
The Arab–Israeli War of 1967 and after
Militancy of the Palestinian refugees
The war of 1973
Civil war in Lebanon, 1974–76
Background to the war
Beginning of the war
Final phases of the war
Consequences of the war

I. Early history, to c. 1600

STONE AGE CULTURES

Paleolithic and Mesolithic. For hundreds of thousands of years, during the Paleolithic Period, or Old Stone Age, primitive men lived on natural resources, both animal and vegetable. Paleolithic man was differentiated from other animals by little more than the fact that his lesser physical strength and natural weapons were compensated for by the tools and weapons that his greater mental development enabled him to provide. During this long Paleolithic Period, men in Syro-Palestine were culturally linked with men in the Eurasian landmass. Good records of their occupation have been found, especially in the caves of Mt. Carmel, where a long succession of levels has been excavated.

The contribution of Syro-Palestine to history does not lie here but begins at a time when the northward recession of the ice cap at the end of the last glaciation brought a different climate to the region at approximately 10,000 BC. The challenge of this change in environment resulted in an enormous step forward, in that man began to seek to control his environment—that is, he began to cultivate plants and domesticate animals. Having taken this step, he was no longer forced to follow the seasonal migration of animals or growth of grains but could produce his own food supply within reach of his home. Settlement is dependent on a food supply controlled by means of agriculture and stock herding, and from the first village settlements developed the first towns and, ultimately, civilization.

Development of agriculture

Archaeologists have proved that it was in Syria and Palestine, along with adjacent areas in northern Iraq and Anatolia, that agriculture and the domestication of animals began, and experts in the environmental sciences have shown why this is so. Clearly, the earliest steps in the domestication of plants or animals must have taken place in those areas in which these plants and animals are found wild. The habitat of the wild species has been extensively studied, and archaeological excavations have provided paleobotanists and paleozoologists with specimens from early sites that have enabled them to begin the study of the process of domestication.

The two grains important to early man in the western part of the Old World were barley and wheat. Wild barley grew throughout the whole of Anatolia and from there spread in a belt past the south end of the Caspian Sea toward India as well as down the Mediterranean coast to southern Palestine. Cultivated barley is found in early sites but never without wheat, and it was the presence of wild wheat that was the controlling factor in incipient cultivation. Two kinds of wild wheat prove to have been especially important: *Triticum boeticum*, which was the ancestor of *Triticum monococcum*, or einkorn; and *Triticum dicoccoides*, the ancestor of *Triticum dicoccum*, or emmer. The habitat of the ancestor of einkorn covers the whole of Asia Minor across to northern Iraq and also down the Mediterranean coast into Syria. That of emmer was much more restricted, covering an area from northeastern Iraq westward to the Mediterranean coast as far south as the hills of Galilee. Emmer also proved to be more important than einkorn, as it was more adaptable to different environments. The habitat area of wheat, and especially that of emmer, provided the environment in which all the known earliest transitions took place from man as a hunter and gatherer of wild foods to man as a settled agriculturalist. The natural habitat of the earliest domesticated animals—goats, sheep, cattle, and pigs—lay in the same general area.

This transformation was gradual. At the end of the Ice Age and of the equivalent cool and wet period in the Mediterranean area, man had to adapt himself to a new environment and to new sources of food. This stage, beginning *c.* 10,000 BC, is usually known as the Mesolithic Period. Man was still basically a food gatherer, but in western Asia the changing natural resources turned his attention more to vegetable foods. There is archaeological evidence of this Mesolithic stage in most of the area in which the subsequent Neolithic was to develop. Within the geographical limits of this article, sites providing evidence of the Mesolithic are confined to the southern part of Syria (and here and below the modern political division of Syria and Lebanon is ignored) and to Palestine (again ignoring modern political divisions). No Mesolithic sites are known in the northern part of Syria, and the development in Palestine of the flourishing Mesolithic culture known as the Natufian provided the basis for the genesis of the earliest known Neolithic stage.

Natufian culture

The classic site for the Palestine Mesolithic is found in the caves on the western side of Mt. Carmel, facing the Mediterranean, where a very long period of Paleolithic occupation was followed by occupation by Mesolithic man. These Natufians lived in the caves much as their predecessors had lived, burying their dead, with elaborate equipment, on the terraces in front. They still depended on hunting for their livelihood, mainly on gazelle. They certainly also harvested grain, for many sickle blades and finely carved bone handles in which they were set have been found, as well as implements for grinding grain—a grain, however, that was probably of the wild variety. Evidence has been found at 'Ayn al-Mālihah, near Lake Hula in the Jordan Valley, of another Natufian group with similar equipment and burial customs, but this group lived very differently. Here there was an incipient village with the inhabitants living in a group of round huts. Tall Murāba'āt in inland Syria is a similar site. A considerable number of Natufian sites are known, stretching as far south as Bayḍā' (Beidha) near Petra, indicating a vigorous and thriving population at this time.

The present evidence is that the Mesolithic Natufians used wild grains, and their sites are found in the habitat of the wild ancestors of einkorn and emmer wheat and of barley. This practice was in itself conducive to the development of fixed settlements. The ripening period of the grain lasted only two to three weeks, and during this time the year's supply of perhaps a ton per family had to be harvested. If grain was to be eaten the year round, a central base is implied. As far as is known the Natufians had not begun the practice of cultivating the plants, involving both preparing the land and the storing

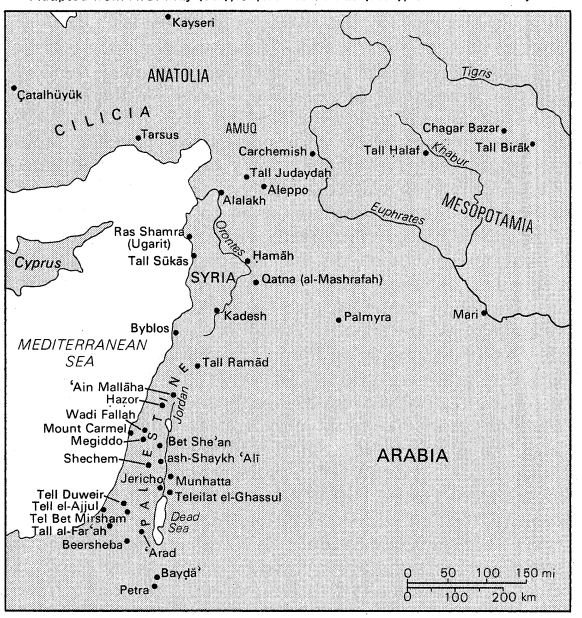
of seed grain. Wild grain is differentiated from cultivated grain by well-defined characteristics that can be identified in the impressions found in the course of excavations. Wild grain is dependent for survival on its ability to scatter freely as it ripens. When man started to reap grain, he would have preferred specimens with a tougher spike axis, which would not scatter as he cut it. Inevitably, a greater proportion of the tougher spike-axis grains was successfully harvested and was included in the next year's seed corn. Thus, man developed a type of grain that had lost the characteristics required for survival in a wild state but was much more useful when deliberately sown. When grain impressions are found showing the characteristics of the tough spike axis, it can be taken as certain that it was cultivated by man.

The physical characteristics of wild animals were also affected by domestication, usually in the direction of reduction in size. This process is rather more difficult to chronicle, partly because the bones recovered in excavations are usually fragmentary, partly because the criteria of domestication are not yet fully established, and partly because it requires many generations to bring about changes that are reflected in the skeletal remains.

The vigorous Natufian stock of Palestine and southern Syria was ancestral to the people who evolved the earliest Neolithic culture, fully settled and practicing a developed agriculture. The site of Jericho in the Jordan Valley beside the spring ʿAyn as-Sultān (ʿAin es-Sultan) has provided evidence of the full sequence. The earliest evidence of occupation there consists of a structure that is probably a sanctuary. The flint and bone equipment of the people who constructed this sanctuary is Mesolithic Natufian, closely similar to that of the lowest Mesolithic level of the Mt. Carmel caves. It may be deduced that Mesolithic hunters established their sanctuary beside the perennial spring that today still waters the wide-stretching Jericho Oasis in evidence of its life-giving importance. The date of this first appearance of man at Jericho is given as *c.* 9000 BC by carbon-14 (a radioactive carbon isotope, the rate of disintegration of which is used to calculate ancient dates).

Jericho

Adapted from H.G. May (ed.), *Oxford Bible Atlas* (1962); Oxford University Press



Principal sites in Syria-Palestine from the Mesolithic Period to the 18th century BC.

Pre-Pottery Neolithic. *Pre-Pottery Neolithic A.* The site proved so attractive to the Mesolithic hunters that eventually they settled there permanently. For a long time they lived in flimsy huts of a type suitable to a nomadic way of life, but they were clearly no longer nomads, for the accumulation of successive huts built up a deposit of 13 feet (4 metres). Eventually the huts were succeeded by one-room round houses, solidly built of brick, in plan clearly derived from the preceding huts. At this stage, occupation spread to the full ten acres of the subsequent town, and the first Neolithic stage of Jericho, now designated Pre-Pottery Neolithic A, was established. The qualification pre-pottery has been included in the title, since until about 1950, when it was accepted that a true Neolithic site was marked by the appearance of pottery. Jericho was the first major site with Neolithic levels earlier than the appearance of pottery, though now this stage in the Neolithic of western Asia is well-known from similar evidence elsewhere.

Soon after the spread of these early houses over the site, there was the significant development of the construction of a town wall of imposing dimensions, associated with at least one massive stone tower 30 feet (9 metres) in diameter, and enclosing the whole ten acres. Such an enterprise implies a fully developed communal organization. The carbon-14 dates show that this stage was reached *c.* 8000 BC.

Jericho was now a town housing perhaps 2,000 to 3,000 persons; this in itself implies an agriculture sufficiently developed to support a large population in one place, though hunting still formed part of the economy, with foxes apparently the favoured meat diet. Evidence has been found, however, of cultivated emmer wheat and cultivated barley. Grains of the wild ancestors have not been found, and they would not have been native to the Jordan Valley with its Sindo-Saharan climate, so it is plausible to suggest that the emergence of the cultivated types took place in the hills, perhaps in the neighbourhood of the Mt. Carmel caves, where the settlement at Wadi Fallah provides evidence of early agriculture. It is also plausible to suggest that to secure a sufficient area for agriculture it was necessary to use irrigation. The communal organization required for such a project could be the origin of the organization implied by the construction of the defenses.

Irrigation

Jericho Pre-Pottery Neolithic A lasted about a millennium. Knowledge of similar settlements, though none on this scale, is growing. It shows that groups with a similar equipment, which at Jericho had developed without a break from the Mesolithic Natufian, and with similar houses are to be found all over Palestine—*e.g.*, at Wadi Fallah on the Mediterranean coast, at Shaykh ʿAli (Sheikh Ali) and Munhatta in the Jordan Valley, and at Bayḍā' to the south, where the earliest Neolithic levels are dated to *c.* 7000 BC.

Palestine thus provides evidence of the development of a fully settled Neolithic culture based on agriculture. It was not the only region in which this took place. The process also certainly took place in the uplands of north-eastern Iraq, of which the type site is Jarmo, dated to *c.* 6750, and in southern Anatolia at Hacilar, *c.* 7000. Both sites were Pre-Pottery, but their equipment and their architecture were different, both from each other and from Jericho. Each area was probably the centre of independent development of agriculture. On the present evidence, however, the first development took place in Palestine, as suggested by the date of 8000 BC for the fully grown walled town of Jericho.

In the northern part of Syro-Palestine, there is evidence of Pre-Pottery stages at Ras Shamra (Level VC) in the north and at Ramād (Level III) in the south, dated at Ras Shamra to *c.* 6600 BC. The equipment and architecture are different from those of Jericho Pre-Pottery Neolithic A, especially in the houses with rectilinear walls and floors of lime plaster. It seems at present that occupation was sparse in this northern area, perhaps because of the absence of a preceding Mesolithic population, though observations may be modified by fuller exploration. It seems probable, too, that this first Neolithic occupation in northern Syria had its roots in Anatolia, where from 6500 BC there flourished a Neolithic culture that is brilliantly exemplified at Çatalhüyük, where rectangular houses with beautifully polished plastered floors have been uncovered (see ANATOLIA, ANCIENT).

Pre-Pottery Neolithic B. It is very probable that the next stage in Palestine is derived from the north Syrian

Neolithic. At Jericho and a number of other sites in Palestine, such as Wadi Fallah and Shaykh ʿAlī, there is an abrupt break. The round houses of Pre-Pottery Neolithic A are succeeded by rectangular houses with floors and walls covered with polished plaster. At Jericho there is clear evidence of the break, with erosion and with erosion and torrent channels cutting into the earlier levels. Above these erosion levels, the new style of houses appears fully developed. In addition to the change from curved to rectilinear walls, the buildings are very different in character, being large and with multiple rooms grouped around courtyards. The flint industry is new, as are most of the types of stone utensils, such as querns for grinding grain. There is still no pottery. The carbon-14 dates at Jericho indicate that the new stage began late in the 8th millennium and may have lasted through the greater part of the 7th.

Architecture of Pre-Pottery Neolithic B

It is clear that newcomers, designated Pre-Pottery Neolithic B, had arrived and had displaced the earlier inhabitants. A Pre-Pottery Neolithic culture that had developed elsewhere had displaced the Pre-Pottery Neolithic culture indigenous to Palestine. The style of architecture indicates that the newcomers came from north Syria; reference has already been made to the rectilinear architecture and plastered floors of Ras Shamra and Tall Ramād (Tell Ramad). The link is made clearer by the fact that at Tall Ramād there were plastered portrait skulls resembling those of Pre-Pottery Neolithic B of Jericho.

Skull decoration

The Jericho skulls constitute one of the most remarkable achievements of Neolithic art. The skulls, usually the cranium only, were removed from bodies buried beneath the floors of the houses after decay had begun but was not complete, and the lower part of each was covered with plaster modelled into the features of the face with extraordinary fidelity and artistic skill. The surface was often painted to represent the complexion; and, in one case, a moustache was depicted. These plastered skulls can, with some assurance, be related to a cult of ancestor worship.

Further evidence of the northern origin of these newcomers is provided by the appearance of einkorn wheat. The wild ancestor of this wheat is not indigenous to Palestine but is found in Syria and Anatolia, and it seems likely that it was brought south by the newcomers.

The new culture spread widely in Palestine, but other sites there do not always present the clear-cut division that exists at Jericho. At some sites, especially those in more arid surroundings, there appears to have been a divergent development from the preceding Mesolithic, which only at this stage became fully Neolithic. In others, it can probably be inferred that the Pre-Pottery Neolithic A people retreated before the advance of the Pre-Pottery B people. This may account for the fact that the earliest Neolithic levels at Bayḍā', about 100 miles (160 kilometres) south of Jericho, predominately Pre-Pottery Neolithic A in character, date only from about the time of the arrival of the Pre-Pottery Neolithic B people at Jericho. The subsequent levels at Bayḍā' in the first half of the 7th millennium show the arrival of the later culture but with more mingling of characteristics than at Jericho; additional groups in contact with the newcomers had probably arrived.

There is thus now fairly adequate documentation for the development of the Neolithic in the Syro-Palestinian area. By *c.* 8000 Palestine saw the full flowering of settlement derived from the indigenous Mesolithic and based on an economy in which agriculture played an appreciable part. This is the first known development of a full Neolithic economy. By *c.* 7000 BC a separate culture had developed in Anatolia and, shortly afterward, another in northeastern Iraq; but the earliest date so far established for northern Syria is mid-7th millennium. It is probable, however, that in due course the north-Syrian culture became preponderant over the whole area of Syria and Palestine, which formed a relatively unitary cultural province by the second half of the 7th millennium.

Pottery Neolithic. The next stage of cultural development certainly came from the north, with the addition of pottery to domestic technology. Pottery was long thought to have come into use with the earliest villages, but it is now clear that the earliest vessels of molded clay came more than 1,000 years later, and the earliest true fired pottery in north Syria came nearly 1,000 years after that.

The earliest known vessels that may be considered as ancestral to pottery have been found at Çatalhüyük in Anatolia, first in Level VIII, about 6750 BC and fully developed, in Level VII, about 6200 BC. It is reasonable to accept that inventions in this area would in due course spread to the south, for in it is an important source of obsidian, which was widely used at an early date. Even as early as Pre-Pottery Neolithic A at Jericho, tools were made of obsidian imported from near Kayseri in central Anatolia; how obsidian was transmitted is still an unsolved problem. The early ware at Çatalhüyük may have been the progenitor of the earliest molded vessels in Syria. In Syria it is known as white ware, or pozzolanic ware, and is not true pottery, for it is not fired clay but rather molded plaster. Pozzolanic ware appears in Ras Shamra (VB) at Tall Sūkās (Tell Soukas) on the coast south of Ras Shamra, in the classification known as Amuq A, based on soundings in tells in Amuq Plain, and Tall Ramād II, dating to *c.* 6000 BC. It clearly represents first experiments in making molded vessels, as opposed to stone or wood-carved vessels. At Tall Ramād II only these vessels are found. In Ras Shamra and the Amuq A classification, with less closely defined stratification and possibly covering a longer period, it is accompanied by the first true ceramic vessels fired in some form of kiln. This constituted a considerable technological advance; man had now begun to alter the natural properties of materials by applying to them extraneous influences such as fire.

Pozzolanic ware

The first pottery to appear in north Syria is found over such a wide area that it can be used both to establish chronology and to define cultural areas. It is known by the clumsy but descriptive phrase dark-faced burnished ware. Its ancestry is certainly northern and can be suggested in finds at Çatalhüyük VIII dated before 6200 BC. It appears at Ras Shamra *c.* 6000 BC and at Amuq A. The ultimate spread of this pottery was enormous, covering Cilicia and Syria as far south as Byblos, where it is found in the early Neolithic levels in which the plastered floors reflect the continuation of the Syrian Pre-Pottery province, and at Tell Ramad, where it appears in Level III.

Dark-faced burnished ware

The finds of pottery of this type indicate a wide range of a technological development and also a great impetus toward settlement, for in many sites the pottery appears in the earliest levels. The widespread occurrence of fired pottery most likely resulted from the transmission of techniques rather than the movement of peoples.

Most of the excavations that have penetrated the levels where pottery first appears have at that depth become soundings of relatively limited extent. Not a great deal is known, therefore, about the characteristics of the settlements. At Ras Shamra the houses seem to have been single-roomed with rectilinear stone walls and plastered floors. The houses were scattered haphazardly and widely spaced over the whole mound. In the soundings at Tall Judaydah (Tell Judeideh), the houses of the Amuq A level had walls covered with finely polished wall plaster. Tall Ramād II had rectilinear walls of stone foundations and a superstructure of mud bricks forming rooms of somewhat irregular shape. There is evidence for the cultivation of emmer wheat in the Amuq A stage. At Ras Shamra there is an increase in the remains of domesticated animals—the bones of pigs, oxen, and sheep being found.

It is possible that this evidence of an increase in stock breeding is associated with a significant opening up of the countryside. The early postglacial vegetation of coastal Syria consisted of coniferous forest on the Lebanon Mountains and of oak forest in the al-Buqayʿah (Bekaʿa) Plain to the east, both dense and inhibiting to early occupation. The open oak forest–savanna area east of the Anti-Lebanon and in Palestine was more suitable country for early settlement; this is probably the reason for the Mesolithic or early Neolithic development in Palestine. In the 6th millennium BC, the forest cover of Syria was

Opening of the countryside

gradually much reduced, very possibly in connection with stock breeding. As a result, the culture or cultures of northern Syria spread steadily, though probably slowly to the south, using both the coastal and the inland routes.

The spread of the culture associated with the dark-faced burnished pottery probably spanned most of the 6th millennium, for it appears at Ras Shamra and in the Amuq area about 6000 BC, but, if the carbon-14 dating is correct, it did not reach Byblos until *c.* 5000 BC. It later penetrated into Palestine, where it is found mainly in the coastal area in the first half of the 5th millennium.

THE END OF THE NEOLITHIC AND THE DAWN OF CIVILIZATION

The Chalcolithic. The ultimate extension into Palestine of pottery-using Neolithic people was preceded by a development of great significance in northern Syria. About 5300 the Chalcolithic Ḥalafian culture appears in Mesopotamia, with its centre in the region of the Khabur (modern Nahr al-Khābūr) River. It was probably brought by newcomers of nomadic origin, perhaps from the foothills of the southern Taurus Mountains. Pottery of the Ḥalaf type, extremely sophisticated and elaborately decorated, appears at Ras Shamra, in Amuq C, and at Ḥamāh on the Orontes River. The spread of true Ḥalaf pottery was very wide and that of derived types even wider. This is significant evidence of a new cultural pattern, with trade links over the wide plains area of the Fertile Crescent from east of the Tigris to the Mediterranean coast, in the east extending into southern Mesopotamia, in the west as far south as Ḥamāh and Tell Tabbut el-Hamman, and in the north into the Anatolian Plateau. The stage is important as indicating the breakup of the tradition of largely (though never completely) self-sufficient village economy. It is also important for the introduction of the use of metal, which was in itself a stimulus to the development of trade, for the sources of copper are limited. Moreover, the introduction of metal tools and weapons led to the growth of a class of specialized craftsmen who had to be supported by food surpluses produced by the rest of the population.

While this technological and economic revolution was taking place in north Syria, the preceding stage of the widespread pottery-using Neolithic of that area was spreading south into Palestine. The highly developed Pre-Pottery Neolithic B culture of Palestine came to an abrupt end, and a number of sites show a gap in occupation and a clear change in culture. The most striking evidence comes from Jericho. The latest carbon-14 dating for Pre-Pottery Neolithic B at Jericho is *c.* 6600 BC; this is not from the latest level, and the settlement may have lasted until the end of the 7th millennium. It must then have been empty and abandoned for a millennium or more, as must have been the other contemporary sites. For this strange gap there is at present no explanation. The next occupation was of an utterly different character. Into the ruins of the preceding settlement at Jericho were dug large numbers of pits, some of them very deep, in which were the huts of the newcomers. Their flint techniques were completely new, and they used pottery, the finer examples of which were characterized by a decoration of burnished red chevrons. This pottery has not been found in Syria, though it is fairly widespread in Palestine; it may be a relatively local development, technologically inspired by the pottery-using province to the north. Its chronological position is suggested by the occurrence of this Pottery Neolithic A ware of Jericho at Tel Batash (Tell Batashi) on the coastal area, down to which the dark-faced burnished ware had spread. At Tel Batash, again in pit dwellings, Pottery Neolithic A ware was found in association with dark-faced burnished ware of the type with "scratched" decoration found at Byblos. It probably belongs to the stage of Ras Shamra IV C, Amuq C, and Itamāh L, at those sites to be dated toward the end of the 6th millennium, though chronologically it may well be somewhat later in Palestine. The way of life of these newcomers was probably that of pastoralists, perhaps migrating seasonally with their flocks.

The cultural hiatus at Jericho

In the next stage in the north, the long-lived dark-faced burnished ware gradually disappeared, and its place was taken by a red-washed ware whose origins are obscure. The pottery forms in this ware, first appearing in the mid-5th millennium, are readily recognizable, especially the type of jar with what is known as a bow rim, and decoration in bands of incised chevrons. It can be readily recognized in the next arrivals at Jericho, designated Pottery Neolithic B, and at a large number of other sites in Palestine. At Jericho, the new arrivals continued to live in pit dwellings, and their way of life in Palestine was still Neolithic, though contemporary with the Chalcolithic Tall Ḥalaf (Tell Halaf) culture in the north.

By *c.* 4500 BC, Mesopotamian influence had died out in the north, and Ḥalaf ware is no longer found. It is, in fact, a period of cultural recession and at some sites is followed by a gap.

The beginnings of towns and city-states. The last centuries of the 5th millennium saw the beginnings of the development of town life in Mesopotamia that were to lead to the city-states in that area of the 3rd millennium. This culture in Mesopotamia is called the Ubaid. Ubaidi pottery penetrated into Syria. It was found on the Mediterranean coast—for instance, in Ras Shamra III B, which was dated by carbon-14 to *c.* 4180 BC, and in the Amuq E stage, which probably had a long and ill-defined duration. Syria did not at this stage reflect the urbanizing tendencies of Mesopotamia in the Ubaid period. At sites such as Ras Shamra and Tall Judaydah, where the pottery associated with this phase is found, buildings are irregularly arranged, though rectangular in plan. At Tall Birāk (Brak), in the Khabur Plain, the earliest levels belong to the final stage of Ubaid, but the successive temples—clear evidence of an urban civilization with strong Mesopotamian contacts—belong only to the second half of the 4th millennium, to the Uruk–Jemdat Nasr stages. It would in fact appear that the spread of urbanization from Mesopotamia into Syria was patchy and sporadic. At some sites, the Chalcolithic Ḥalafian culture may have continued alongside the newer developments. At some sites (*e.g.*, Chagar Bazar in the Khabur) there is no trace after the final Ḥalaf stage, mixed with some Ubaid, of the Uruk and Jemdat Nasr stages; there may, therefore, have been a gap in occupation from *c.* 4000 to *c.* 2800 BC. At Mari, on the Middle Euphrates, a site that was to become very important in the 3rd and 2nd millennia, there is scant evidence for the Ubaid–Uruk period. In the Orontes Valley and Amuq Plain, there was also a very considerable decrease in settlements of appreciable size during the course of the 4th millennium. On present evidence Syria was a backward area for much of the 4th millennium, and the nascent city-states of Mesopotamia were not yet much interested in the area.

The Ubaid period

In Palestine there is growing archaeological evidence for the 4th millennium BC. The second Pottery Neolithic stage, the Pottery Neolithic B already referred to, overlaps chronologically with a stage that is certainly Chalcolithic, for there is a linking of distinctive pottery forms. The new stage is known as Ghassulian, from Teleilat el-Ghassul (Tululeit Ghassul) in the Jordan Valley where it was first identified. Though the Ghassulian has links with Pottery Neolithic B, it certainly represents the arrival of newcomers, probably from the northeast. Closely linked with the Ghassulian is the Beersheba culture of southern Palestine, and here its full character emerges. The settlements were villages with primitive often semi-subterranean dwelling places, such as those of the Pottery Neolithic B. But the inhabitants were artisans and skilled workers in copper and other materials. The carbon-14 method at first gave dates for these settlements in the second half of the 4th millennium, but more refined dating techniques suggest dates in the first half of the millennium. Palestine at this stage was therefore absorbing the technological advances that had reached Mesopotamia and Syria in the Ḥalaf period.

The next stage in Palestine is marked by the arrival of fresh and successive waves of newcomers in what can best be called the Proto-Urban phase, as a prelude to the Early Bronze Age. This terminology is not universally accepted. Some experts prefer to classify some elements

Proto-Urban phase

as Late Chalcolithic and some as Early Bronze Age. The main group, Proto-Urban A, originated to the north, probably in the Amuq Valley. A small group of the same people even penetrated to the north, to Tarsus in Cilicia. The newcomers, when they arrived in Palestine, were still villagers. They introduced new pottery, new burial methods, with the first rock-cut tombs containing multiple burials, and they were certainly sedentary. Their settlements constituted the genesis of the Early Bronze Age towns of Palestine. At Tell Gath in southern Palestine, five levels preceded the intervention of the Egyptians at the beginning of the 1st dynasty (*c.* 3100 BC), and at Jericho the period is represented by at least five and possibly seven occupation stages. Not all sites show this same development toward urbanization, but it was clearly taking place. During this period other groups, Proto-Urban B and Proto-Urban C, were arriving, again from a northern direction, and combining with the Proto-Urban A group.

In the last centuries of the 4th millennium, therefore, Syria and Palestine were occupied by village populations, with connecting links from east to west in Syria and from Syria in the north to Palestine in the south. The succeeding millennium shows the development of these villages into towns and probably city-states and also the influence upon them of the Euphrates Valley empires to the east and of Egypt to the southwest.

THE EARLY AND MIDDLE BRONZE AGES

At this stage, written records become available in addition to archaeological evidence. For Syria, historians tend to regard periods or sites for which there is no written evidence as *terra incognita*. Archaeologists work out their sequence and gratefully accept evidence that texts supply to fix points of chronology. Textual evidence from the direction of Mesopotamia is greater for Syria, but negligible for Palestine. For the coastal area, Egyptian contacts provide valuable fixed points for the archaeological evidence.

The reign of Sargon I

The epigraphical evidence concerning Syria is explicit only in the second half of the 3rd millennium. Sargon I of Akkad (2334–2279) marched across to the Mediterranean, and his records deal with the Syrian states with which he had relations, warlike or diplomatic. He needed the wood, stone, and metal that the west could provide. Records, however, carry the history of Syrian states to a pre-Sargonid stage. At Mari there were rulers of importance before the middle of the 3rd millennium. Mari was certainly one of the most important Syrian towns of the period. In culture, it was strongly influenced by Mesopotamia, but its kings had Semitic names. Epigraphical evidence is scrappy, but the general picture is similar for most of northern Syria: a growth of towns with strong Mesopotamian influence, the rulers mainly Semitic, but a large proportion of the population non-Semitic, representing the indigenous population prior to incursions of a ruling stratum from the Syrian Desert.

The towns of the Mediterranean coast at this stage assume a new importance based on the growth of Egypt. It is possible that down to the 4th millennium, Egypt was the recipient of progress rather than its instigator. The Neolithic stage was late in Egypt, and the Ghassul–Beersheba people of Palestine may have instigated the development of copper working there. From the beginning of the 1st dynasty (*c.* 3100 BC) Egypt became an expanding power. The Early Bronze Age towns of Palestine and coastal Syria show Egypt's impact, sometimes that of trade but sometimes implying political control.

'Arad

Early Bronze Age cities. The places in Syria and Palestine with which Egypt had contact grew rapidly to have the status of towns. One of the most remarkable of the new towns is 'Arad in southern Palestine. 'Arad sprang up in an area with few agricultural advantages but nevertheless had become a large walled city by Early Bronze I (*c.* 3000 BC). Its importance was certainly connected with Egyptian trade, possibly in bitumen (or asphalt) from the Dead Sea. Other walled cities dating from the beginning of the Early Bronze Age include Megiddo, Tall al-Fāri'ah (Tell el-Far'ah), and Jericho. A similar walled city appeared at Byblos, and urban development is suggested at Ras Shamra, though the evidence from excavations there is limited. It is clear that the culture of the Mediterranean coast is closely related to the southernmost part of Palestine. The pottery evidence shows that on the combined cultures of the Proto-Urban stage in Palestine a new influence supervened, to which perhaps is to be ascribed the full development of urbanization. This is linked by pottery with northern Syria, whence the preceding people had also come.

The more ample archaeological evidence in Palestine enables the various stages of the Early Bronze Age to be more closely linked with Egyptian chronology than that of the corresponding period in Syria, the Egyptian chronology being more exact than that of Mesopotamia. In Palestine, Early Bronze I overlaps with the 1st dynasty of Egypt, whose first king, Narmer, destroyed the Proto-Urban town of Tell Gath. Early Bronze II extends to the beginning of the 3rd dynasty (*c.* 2686 BC), and Early Bronze III in Palestine covers the 3rd, 4th, and 5th dynasties. The correlations are based on Egyptian imports into Palestine, sometimes bearing the names of pharaohs, or at least fairly closely dated.

Egyptian imports into Palestine imply close contacts, at least in the area of trade, and Egyptian records suggest that Egypt may have exerted some political control over the cities of Palestine. But it is difficult to assess how much this was real or how much traditional bombast; certainly the basic culture remained Palestinian–Syrian.

Byblos

Probably the relations between Egypt and Byblos typify Egyptian contact with the Syrian coast. Egypt lacked stone, metal, and, above all, wood. As Egypt of the 1st dynasty developed an urban civilization and an empire, its need for these materials grew. They were to be found in Syria, where the Early Bronze Age culture just described covered the length of the coast. Byblos was the most important port, with its hinterland the best source of lumber and other materials Egypt needed. The Byblos of the Early Bronze Age was a very different place than that in the early period. It was a walled city, with houses closely built up, regularly orientated, and with a degree of uniformity of plan. Contacts between Egypt and Byblos probably began at the beginning of the Early Bronze Age, but they are epigraphically confirmed by gifts to the temples of Byblos bearing the name of Khasekhemui, the last king of the 2nd dynasty (*c.* 2890–*c.* 2686 BC), associated with finds belonging to the beginning of Early Bronze III.

Early Bronze III marks the full urban development of Palestine, and it is probable that further archaeological exploration in Syria will show that period, *c.* 2700–*c.* 2300, to be the most prosperous for the 3rd-millennium towns of the entire Syro-Palestine area. During the period, a new people certainly arrived, as evidenced by a new type of pottery known as Khirbet Kerak ware. Its origins are northeastern, possibly Caucasian. It is much more in evidence in northern than in southern sites. It must have been brought by newcomers, but the evidence in Palestine is clear, and is probable in Syria, that there was no disruption but rather an infiltration of newcomers. It can be suggested—but very tentatively—that the recrudescence of copper working, which is a certain fact in Palestine, where for the earlier part of the Early Bronze Age metal weapons are scarce, was due to newcomers from the metal-bearing area to the northeast.

Inland Syria

Early Bronze Age Palestine and coastal Syria culminate in a flourishing urban civilization in the third quarter of the 2nd millennium. In inland Syria, the picture is probably similar. Mesopotamian textual references for this period can seldom be related to the states that emerge in the 2nd millennium. Nevertheless, the references show that embryo cities existed. The economies of the inland states were linked with Mesopotamia, and during the 2nd millennium the desire of the Mesopotamian states for the products of the Syrian coast and the Mediterranean increased. The coastal states were reached by tendrils from Mesopotamia, but their main development came from coastal trade with Egypt. The interlocking trade between Mesopotamia, the Mediterranean, and Egypt was to

reach a fuller development in the 2nd millennium, but its Early Bronze Age development was clear.

It is certain from the survival of certain personal names that the population of Syro-Palestine in this period was Semitic. It can be suggested that these people were Canaanites, but the name Canaan appears only in the 2nd millennium, when coastal Syria and Palestine again shared a common material culture. This culture was not directly derived from the Early Bronze Age culture but emerged as an amalgam after the nomadic incursions described below. It seems preferable, therefore, to reserve the term Canaanite for the 2nd millennium stage, but it is not a distinction where dogmatism is necessary.

The Amorite invasion. In the last centuries of the 3rd millennium the urban civilization of Syria and Palestine suffered a series of setbacks. The inhabitants of the fringes of the Syrian Desert, nomadic or seminomadic pastoralists, warlike in character, succeeded in breaking into the town-centred states of northern Syria and the coastal areas from the Euphrates to the Mediterranean. In the Sumerian language, the invaders were known as Martu and in Akkadian as Amurru—the Amorites of the Old Testament. Their language was West Semitic. By about 2050 they seriously menaced much of Mesopotamia of the 3rd dynasty of Ur (*q.v.*), but their penetration into the settled states of Syria and northern Mesopotamia may have begun as early as the end of the 4th millennium. During the 3rd millennium Amurru kings, recognizable by their West Semitic names, appear in Mesopotamian records, and it is clear that some of the newcomers had settled down and formed small principalities. There remained, nevertheless, a great wandering population not interested in town life; it was these Amurru elements that broke into the settled lands of Syro-Palestine in the last centuries of the 3rd millennium.

Dating the invasion of the Amorites

The precise date of their invasion is difficult to fix and may, indeed, have varied appreciably according to the strength of the different towns and principalities. Pressure on the empire ruled from Ur was serious by the middle of the 21st century. At the other extremity of the Fertile Crescent there was a break at Byblos when trade with Egypt ceased at the beginning of the First Intermediate Period (*c.* 2160), a break probably due both to disruptions at Byblos and to incursions of nomadic Asiatics into Egypt, probably connected with the same population movements. The fact that Asiatic incursions into Egypt had reached this scale by this date supports the Palestinian evidence that the break came about the 24th century. It is reasonable to suppose that the new invaders from the desert fringes penetrated more easily into the weaker states of Palestine and perhaps some of those of north Syria before they affected major Syrian sites, such as Byblos, or endangered the great empires of Ur and Egypt.

The archaeological evidence of the nomad incursions is dramatic. In Syria, great centres such as Ugarit (Ras Shamra) and Qatna (al-Mashrafah [Meshrife']) cease to yield evidence of any built-up town. The evidence of the arrival of the newcomers is provided by graves dug into the slopes of the mounds that had built up over the centuries-long growth of the pre-existing towns. At Byblos, however, the evidence is somewhat different; houses still stood on the town site, but these were different in architecture and scattered irregularly over the earlier built-up area. The most striking point at Byblos is the sudden appearance of a flourishing metallurgic industry that manufactured tools, weapons, vessels, as well as an enormous quantity of presumably votive figurines in copper or bronze. The greater part of these objects were found in votive deposits in sacred areas that may have existed earlier and that certainly continued in existence afterward; the votive deposits, however, belong to this period. The most famous deposit, the Montet jar, after its discoverer, Pierre Montet, is probably to be dated between 2130 and 2040 BC.

Nomadic "tinkers"

Deposits of tin in the neighbourhood of Byblos may have attracted there bands of "tinkers," nomadic groups with no ethnic roots who, through the millennia, have wandered over Asia and Europe. Nevertheless, the archaeological links, in the form of weapons and ornaments, make it clear that this stage at Byblos can be related to the break brought about by the newcomers at such sites as Ugarit and Qatna. Tinkers may have taken advantage of the disruption caused by other populations on the move.

In Palestine, the evidence of the break is even clearer, for the archaeological evidence is more abundant. The new stage is best described as the Intermediate Early Bronze–Middle Bronze period, though others prefer Middle Bronze I, a label given when the significance of the period was not yet clear. The Early Bronze III period ends, as has been seen, *c.* 2300 BC, soon after the end of the 5th dynasty of Egypt. At Jericho there are clear signs of a complete destruction by fire of the final Early Bronze III town. The next houses are preceded by a period when there was only camping occupation, and houses, when they appear, are quite different from those of the Early Bronze Age; the town, moreover, was not walled, and the pottery and other equipment of its occupants were entirely different from that of their predecessors. The presence of the newcomers is, in fact, attested far more by their tombs than their houses, for some 400 have been excavated at Jericho, each containing a single or (rarely) a double burial, in marked contrast to the multiple burials of the Early Bronze Age.

Throughout southern and central Palestine the position is the same—almost no traces of buildings or villages, certainly no towns, but a multitude of tombs. The finds in the tombs and the method of burial fall into a number of groups. At Jericho itself, seven groups have been differentiated. The evidence suggests tribal groups only loosely attached to local centres, a suggestion supported by the practice of some groups of burying—in elaborately prepared rock-cut tombs—disordered and desiccated bones, evidence that the bodies of those who had died during the course of seasonal migrations were brought back in due course to the tribal centre for burial. An urban population had been succeeded by a pastoral, seminomadic population, loosely organized in tribal groups.

Burial rituals

The position is somewhat different at Megiddo and probably Bet She'an (Beth-shan) in the north of Palestine. The break with the Early Bronze Age is equally sharp. The pottery of the newcomers shows that they are linked with, and chronologically contemporaneous with, the groups further south. There seems, however, to be more in the way of town life. At Megiddo at least one of a succession of temples belongs to this period and probably some other buildings of importance. Both the pottery and the metal objects show contacts with Syria. It is probable that the newcomers reached Bet She'an and Megiddo after having passed through the settled areas of Syria and were thus more sophisticated, whereas those who appear further south came more directly from the desert fringes.

The picture is thus similar over the whole of northern Syria and Palestine. The preceding urban civilization was disrupted by wide-scale incursions of seminomadic Amorites. In Syria, the break may have lasted from *c.* 2200 to *c.* 2000. In Palestine, with poorer and weaker urban development, it may have lasted from *c.* 2300 to *c.* 1900.

Middle Bronze Age. In Palestine there is again a sharp break between this period and the Middle Bronze Age, when urban civilization reappears. This is because newcomers again appear, bringing with them a more civilized way of life developed elsewhere. The evidence suggests that the Middle Bronze Age of Palestine is closely linked with that of coastal Syria as far north as Ras Shamra. The evidence suggests that in the region around Byblos, with its urban tradition, the invading Amorites were assimilated into the existing patterns of town life and that the new urban civilization developed that, at least on the coast, can certainly be called Canaanite; it lasted until at least 1200 BC.

Literary evidence. At this stage Syria and, to a much lesser extent, Palestine enter into the light of written history. There is no evidence that the towns of Palestine that developed in the 19th and 18th centuries themselves kept written records, though this may be the accident of sur-

vival or discovery. The historical evidence for this area comes from Egypt. Trading relations with Byblos were resumed under Amenemhet I (1991–1962 BC), first king of the 12th dynasty, who also re-asserted control over the turquoise mines of Sinai and sent an expedition as far north as Shechem. Relations thereafter with Byblos were close. Gifts from Egyptian kings are found in the tombs of the rulers of Byblos, whose Semitic names are written in hieroglyphics. During the 12th dynasty (1991–1786 BC), Egypt maintained close relations with towns in coastal Syria and Palestine by means of representatives whose influence may have varied between that of ambassador and that of a governor.

Evidence from Egyptian records

Egypt thus provides the literary evidence for the coastal and southern areas. For the inland and northern areas there is much fuller evidence based on Mesopotamian contacts and local records of Mesopotamian type, which were kept on baked-clay tablets, thus surviving much better than records on other materials; this evidence interlocks on some of the northern coastal sites with the Egyptian evidence. The main sources from Syria itself come from the archives found at Mari on the Euphrates, which cover a period of about half a century, centred on the year 1800, and the archives of Alalakh ('Atshanah [Atchana]), which cover the period *c.* 1730 to 1655. The picture is one of a unified culture or at least way of life, Mesopotamian in inspiration, stretching from the Euphrates to the coast. Within this broad band, bounded by the Syrian Desert on the south and the plateau of Anatolia on the north, were a number of states of fluctuating importance, to which lesser states were loosely attached. The tie connecting the greater states together was trade. Mari, Carchemish, Aleppo, Qatna, Alalakh, and Ugarit (Ras Shamra) all formed links in a chain between Mesopotamia and the Mediterranean, while Byblos, to the south, played its part both in trade with Mesopotamia and with Egypt.

The princes and ruling classes of most of the states in the first half of the 2nd millennium had a not far-distant nomadic background, having established themselves in the period of disturbance just described following the breakup of the 3rd dynasty of Ur. Iakhdunlim, whose rule began *c.* 1830 BC, and who was the founder of Mari's greatness, was of West Semitic (Amorite) origin. Though Mari was on the right bank of the Euphrates, its territory lay mainly to the east, especially in the triangle formed by the River Khabur and the Euphrates. For a time, Mari fell under the control of Shamshi-Adad I of Syria but on his death was recovered by Zimrilim, son of Iakhdunlim. Though Zimrilim was constantly engaged in warfare against other states and against the ever-present menace of desert nomads, such as the Benjaminites, the records found in his palace show the wealth of trade that passed through his capital and the ambassadors that he maintained at foreign courts, while the remains of his enormous palace are one of the most impressive pieces of evidence recovered by archaeology as an illustration of the advanced urban civilization of the period.

Mari

One of the other most important states of the time was Iamkhad, with its centre at Aleppo, which rose to importance under Iarimlim, *c.* 1775 BC, also Amorite in origin. The kingdom of Iamkhad stretched from the Euphrates to the Mediterranean coast, with Alalakh and Ugarit as vassal territories. All trade from the Mediterranean countries to Babylonia that did not go by the desert route via Palmyra passed through Aleppo. Qatna, on the Orontes, the rulers of which again were Amorite, bordered Iamkhad to the south and was another independent and important state. Qatna was the southernmost of the major Syrian states. To the south was the general area of the Amurru, where petty Amorite princes ruled small areas.

From place-names and the names of their rulers, the states of north Syria can be identified as Amorite (West Semitic). There exists, however, a substratum of names of persons and places of which the ethnic significance is not clear; presumably this represents the indigenous population of the Neolithic and Chalcolithic periods for which the history, described above, can be based on archaeology alone. In addition, during the first half of the 2nd millennium a new element becomes increasingly important—that of the Hurrians. The Hurrians appear in Mesopotamia east of the Tigris, probably coming from the northeast, in the mid-3rd millennium. Their spread, from *c.* 1800 onward, is marked by an increasing number of names, and they seem to have secured positions of importance in the administrations of a number of states. They remain, however, a minority until the second half of the millennium.

Rise of the Hurrians

Archaeological evidence. The wealth of literary evidence for this period in Syria means that archaeology has not the importance there that it has, as will be seen, in Palestine. Only four sites have been excavated on a large scale: Mari, Ugarit, Alalakh, and Byblos. Mention has already been made of the great palace of Mari. All the sites excavated show royal buildings and town life on an imposing scale, though lesser than that of Mari. There are two aspects on which archaeology throws important light. The first is that there was a clear difference in culture between coastal and inland Syria. Evidence for this is provided by the pottery and other objects of common domestic use. The disruptions caused by Amorite movements at the end of the 3rd millennium brought about a complete break in the ceramic traditions. When urban life was re-established, a completely new repertory of pottery forms came into use, which probably developed in the neighbourhood of Byblos. Vessels closely similar in form came into use, from Ugarit in the north to the extremities of the Mediterranean. In inland Syria (*e.g.*, at Alalakh, Qatna, and Ḥamāh) the forms are for the most part quite different, being more closely allied to those of Mesopotamia. The coastal area where this new pottery appears is subsequently known as Canaan (though it includes an area to the north), and it is reasonable to give the new culture there the name Canaanite.

The second aspect upon which archaeology throws light is that of the defenses of a number of towns. It will be convenient to deal with this aspect in connection with the evidence from Palestine.

The history of Palestine at this time, in contrast to that of Syria, is based almost entirely on archaeology. The towns of Palestine were not of sufficient importance to receive much attention in the Mari archives, and the only site to be mentioned is Ḥazor. The evidence from Egyptian sources has already been mentioned, but it is slender in comparison with that for the Syrian states. On the other hand, the archaeological evidence is far more copious than in Syria.

Sometime after 1900 BC there is a break as sharp as that which brought the Early Bronze Age to an end. A completely new culture appears, that of the Middle Bronze Age. The direction from which it came is clear, for the new pottery's appearance in coastal Syria has just been described. The best example for the pottery in Syria is at Kafr Djarra. The newcomers were probably at first few in number, but they fairly rapidly stimulated a revival of town life. Almost all the old sites of the Early Bronze Age that had been deserted or sparsely occupied during the Intermediate period became walled towns again. The earliest city to regain its importance was Megiddo, where town life had not been completely interrupted. Excavations there have shown a long sequence of levels of well-built houses and fortifications, and mention of connections with Egypt stress its outstanding position.

The full urban development of the Middle Bronze Age seems to come only in the late 18th century. The most striking evidence of this is at Ḥazor, in the Jordan Valley north of Lake Huleh. The first Middle Bronze Age occupation there was confined to the area of the mound that marked the site of the Early Bronze Age town. To this town was added, probably in the second half of the 18th century, a vast new area, enlarging the site to about 182 acres. This area was built up with houses and temples and was defended by an entirely new type of fortification, consisting of a great earth bank with a wall on the top. Defenses of the same sort, an earth bank usually with a plaster facing, are found at many sites, such as Jericho, where it can be dated to the late 18th century BC, Megiddo, Tel Bet Mirsham (Tell Beit Mirsim), and Tel

New defensive works

Lakhish (Tell Dnweir), though at these sites the bank is for the most part backed against the earlier mound. Two important sites in the south Tell el-Ajjul and Tall al-Farʿah, seem first to have been occupied only when this new type of fortification was coming into use.

There is a particular interest attaching to this type of fortification, since it is found over a wide area in Syria and Palestine. At Carchemish on the Euphrates, at Ugarit on the coast, and at Qatna on the Orontes, a similar great expansion outside the area of the original town takes place, defended by similar massive ramparts. It will be noted that these sites include both the cultural areas (inland and coastal Syria) described above. The defenses are similar, but the culture is different. The clear suggestion is that when the important states of the 19th and 18th centuries were beginning to weaken, a new warrior aristocracy was acquiring influence and introducing new methods of warfare. There is a strong probability that these warriors came from the nomadic tribes of the Syrian Desert, the Semitic Benjamites, and the Ḫapiru (Habiru, ʿApiru) who may have been nontribal warrior bands, against whom Zimrilim of Mari, for instance, was continually at war.

The invasion of the Hyksos

This is, in fact, the period at which the 12th dynasty of Egypt succumbed to pressure of Asiatics, first perhaps infiltrators, then invaders—the so-called Hyksos, whose capital was established in the Delta by *c.* 1720. It is thus very significant that a site defended by earth ramparts appears at Tall al-Yahūdiyah (Tell el-Yahudiyah) near Cairo, together with pottery partly native Egyptian but partly linked with Palestine.

The flourishing period of states ruling large areas in Syria thus seems to come to an end in the 17th century. The kings and princes known from the written records are much less imposing. Town life goes on, and the towns of the end of the Middle Bronze Age in Palestine are certainly flourishing. It is reasonable, however, to conclude that it is a period when the warriors from the desert had succeeded in imposing their rule in many places, with effects extending from the Euphrates to Egypt.

(K.M.K.)

II. Syria and Palestine, c. 1550 BC–AD 634

THE ERA OF THE EGYPTIAN AND HITTITE EMPIRES, C. 1550–1200 BC

The struggle of the Great Powers for Syria and Palestine. When Kamose of Thebes raised the standard of revolt against the Hyksos—the hated foreign rulers of Egypt—in the second quarter of the 16th century BC, he could not have known the profound effect his action would ultimately have on the history and civilization of western Asia. For Kamose's successors, the first pharaohs of the 18th dynasty, pursued their erstwhile masters into Palestine and at the same time began to evince imperial ambitions of a totally new scale. Although close trade relations had existed between the Nile Valley and the countries of the Levant since about 3000 BC, and although during the Middle Kingdom (*c.* 2040–1786 BC) these relations had been reinforced by something approaching political control over such coastal cities as Byblos, it was not until the 18th dynasty that the Egyptian armies regularly treaded the roads of Syria and that Syrian taxes regularly swelled the Egyptian treasury.

Egyptian hegemony over Palestine and Syria

The Egyptians. The growth of Egyptian hegemony over Palestine and Syria is relatively well documented in the inscriptions with which the pharaohs embellished their temples, and the main stages of that growth are clear, though many problems of chronology and historical geography remain. To the first stage belong the activities of the first four rulers of the dynasty—Ahmose (reigned 1570–1546), Amenhotep I (1546–1526), Thutmose I (1525–*c.* 1512), and Thutmose II (1512–1504). The initial campaigns of Ahmose into Asia seem to have been on a relatively small scale, and progress was slow, despite the Egyptian chariotry. An indication of the Asians' resistance to the Egyptians is the fact that the town of Sharuhen in southern Palestine (probably Tall al-Farʿah near Gaza) withstood a siege of three years before it was taken, and it was probably not until the end of his reign that Ahmose reached northern Palestine. The pace of conquest, however, seems then to have accelerated. No information concerning the campaigns of Amenhotep I has survived, but the Egyptian position in Palestine must have been consolidated and some advance made into Syria, for, as early as the second year of Thutmose I, the Pharaoh claims to have made the Euphrates his frontier and is said to have erected a victory stela on the bank of that river. Thutmose's claim was clearly exaggerated and premature; Egyptian authority was by no means yet permanently established anywhere in Asia, and the records show that Thutmose's son, the second of that name, campaigned against rebellious native rulers in areas that his father had already tried to pacify. Egypt could not yet maintain permanent armies of occupation or establish an imperial administration in Syria and Palestine. Nevertheless, its achievement in Asia by 1500 was spectacular, reflecting the determination and dynamism that characterized the new royal house of Thebes.

The second stage in the establishment of Egyptian empire in Asia began with the assumption of power by Thutmose III in 1482 (after a coregency with his stepmother Hatshepsut) and covered the rest of the 15th century. Before turning to this, however, it is necessary to describe briefly the conditions found by the Egyptians in Palestine and Syria during the initial phase of conquest. Evidence for this is quite inadequate: the surviving written records are meagre and difficult to interpret, while the archaeological evidence is uneven and equally doubtful. It would appear that, although the whole area of Palestine and Syria consisted of more or less independent principalities centring upon strongly fortified cities, there were striking differences in ethnic composition and political organization between the north and south. In Palestine, the population was still almost exclusively West Semitic, speaking a language that scholars variously designate as either Amorite or Canaanite, and there is as yet little sign of the non-Semitic speaking peoples—Hurrians and Indo-Aryans—who were already well established farther north. Archaeology shows that the Palestinian towns were strongly fortified (witness the defense of Sharuhen, mentioned above), prosperous if not luxurious, and ruggedly independent in both their political life and their culture. Above all, there is no sign of external interference in their affairs prior to the Egyptian invasion, though lack of documentary evidence prevents scholars from knowing how much interstate rivalry and warfare existed. Many excavated towns show signs of having been destroyed about the middle of the 16th century and abandoned for a time thereafter. At some (*e.g.*, Tall al-Farʿah) this is clearly the work of the Egyptian armies, but at others, and especially those off the main routes (*e.g.*, Jericho, Tel Bet Mirsham [Tell Beit Mirsim]) the devastation could have been caused by marauding bands of Asians taking advantage of the troubled times.

Conditions during the initial phase of conquest

Mitanni. In Syria, the population on the eve of the Egyptian invasion was much more mixed, and political conditions were much less stable. From early in the 2nd millennium, people speaking a non-Semitic language known as Hurrian had been spreading from their homeland—probably in eastern Anatolia—and had established themselves alongside the older Semitic population in the northern Mesopotamian and Syrian cities. By the 16th century, the principalities of northern Mesopotamia seem to have coalesced to form a unified state, known as Mitanni, the rulers of which (judging from their names) belonged to yet another ethnic group, the Indo-Aryan. Mitanni soon became a powerful force, and it is generally supposed that one element of its strength was the progress made by the Hurrians in developing, if indeed not inventing, the light horse-drawn chariot as a weapon of war. The affairs of Mitanni belong properly to Mesopotamian history; but it is relevant to note that the 16th century in northern and central Syria was a period of political turmoil, in which the kings of Mitanni tried to assert their authority, by intrigue or by war, over such Syrian states as Aleppo, Mukish (in the Plain of Antioch), and Qatna (modern al-Mashrafah). It was no doubt because of the

Political turmoil

instability of the area that Thutmose I and II were able to bring Egyptian arms to the Euphrates with such apparent ease; but it was also because of the growing might of Mitanni that these same pharaohs were unable, by 1500, to consolidate their position satisfactorily in the north.

Dynastic quarrels within Egypt at the beginning of the 15th century were largely responsible for a hiatus in its Asian adventure, but, with the accession to full authority of Thutmose III in 1482, the situation changed. The Asian princes were aware of the challenge presented them by the young soldier-pharaoh, and a coalition was organized to oppose him. No fewer than 330 Palestinian and Syrian rulers were involved in the revolt, led by the Prince of Kadesh (modern Tall an-Nabī Mind, near Homs) and including many of the Hurrian and Indo-Aryan *mariyanna*, or warrior aristocracy, that by now were prevalent throughout the region. Such was the might of Egypt, however, that Thutmose had little difficulty in smashing the confederates at Megiddo in northern Palestine (1480). The ferocity of Thutmose's punitive measures is indicated by archaeological evidence that this site was probably completely deserted for most of the 15th century. With this battle it may be said that Egyptian authority in Palestine was firmly established, and it was to remain unchallenged for more than a century; this was the Golden Age of empire. The same was not true, however, of the northern marches. There the power of Mitanni continued to thwart the Egyptian design, and, although Thutmose III was able, in the years immediately following the fall of Megiddo, to defeat "the wretched Mitannian foe" at Aleppo and Carchemish, the situation in Syria remained unresolved, with the major cities changing hands almost annually. While Thutmose was alive, the balance of power was mostly in Egypt's favour, and it is possible that a treaty of peace between the Pharaoh and Saustatar of Mitanni at this time led to an uneasy relaxation of tension. But the old Pharaoh died in 1450, and this was the signal for a revolt by the cities of the Orontes Valley (*e.g.*, Kadesh, Tunip). The first campaign of Amenhotep II (1488) repressed this with savagery: the Pharaoh executed the rebel princes with his own hand, and their bodies were exposed in the Egyptian capital as part of the paean of victory. Such ferocity, however, had no lasting effect. In the seventh year of his reign, Amenhotep was obliged to march again against many of the cities that he had reduced only a few years previously; and, two years later, even Palestine seems to have been in revolt, and 70,000 captives are reported to have been sent to Egypt. These revolts were no doubt instigated and fostered by the Mitannian kings.

Revolts against the Egyptians

The Mitannian opposition to the Egyptian imperial design was ending, however, though through no wish of the Mitannian kings. A new power, that of the Anatolian state of Khatti, was rising on the northern horizon, and its threat to both Mitanni and Egypt induced the rulers of these empires—Thutmose IV and Artatama—to conclude (*c.* 1425) another treaty of peace, later sealed by the Pharaoh's marriage to a Mitannian princess. The empires agreed on respective spheres of influence; slightly later documents indicate that Egyptian control stretched as far as Ugarit on the Syrian coast and inland as far as Tunip (probably west of modern Ḥamāh) and Damascus. North and northeast of this boundary the Mitannian held sway, in such cities as Qatna, Aleppo, and Carchemish.

The Hittites. The history of the Egyptian Empire in Asia during the century following the détente with Mitanni was marked by the same kind of boundary fluctuations and internal revolts that had characterized its earlier development. There were, however, two important differences: the pharoahs of the 14th century were weak, whereas those of the 15th had been strong, and the enemy in the north was no longer Mitanni but the Hittite state. Before 1400 the Hittites had hardly impinged upon Syrian history, though as early as *c.* 1595 Mursilis I of Khatti had destroyed the cities of Aleppo and Alalakh and had marched as far southward as Babylon. These events had left little permanent trace on Syrian affairs, however, and the Hittites in the 16th and 15th centuries, preoccupied with domestic matters, played only a minor role in Syria compared with that of the Mitannians. The fact that Idrimi, king of Mukish, was able to attack with impunity a number of Hittite towns toward the end of the 16th century is illustrative of the prevailing situation. Despite some temporary success in the 15th century in imposing their authority over some of the most northerly Syrian towns, the Hittites at that time could hardly be considered a Syrian power.

The event that, more than any other, can be said to mark an end to this situation was the accession (1380) of Suppiluliumas to the Hittite throne. Suppiluliumas was a young, ambitious soldier who saw his imperial destiny in Syria. The immediate enemy was Mitanni, and the Hittite king took pains to preserve formally correct, though undoubtedly cool, diplomatic relations with Egypt. Having reorganized his homeland, he embarked, in about 1365, on his first Syrian campaign. Tushratta of Mitanni was unable to oppose him; most of the country from the Euphrates to the Mediterranean as far south as Lebanon fell. And, although the Mitannian king was the loser and the Egyptian vassal cities such as Ugarit and Kadesh seem to have remained loyal to Pharaoh, the potential threat posed to the Egyptian Empire could not be ignored.

The accession of Suppiluliumas

The threat, however, was not yet a direct one. The result of the Hittite victory in Syria was to upset the balance of power; it removed Mitannian control but did not replace it with equally strong Hittite rule. Although some of the Syrian principalities became Hittite vassals, others, particularly those in the south, took the opportunity to assert their independence and follow their own ambitions. The state that most clearly shows this development is Amurru, which had existed, apparently quietly, for many centuries between the southern Orontes and the Mediterranean, in what is now largely Lebanon. Since the time of Thutmose III, the princes of Amurru had been content with their vassalage to Egypt, but, at the beginning of the 14th century, the new ruler, Abdi-Ashirta, had begun to exert some independence and to extend his territories. The Hittite successes gave him further opportunities, and a number of cities loyal to Egypt were taken by him.

Abdi-Ashirta was also helped by internal difficulties in Egypt. In 1379 Akhenaton (Amenhotep IV) had come to the throne; he had soon shown himself to be untypical of the rulers of the 18th dynasty—he was a mystic rather than a soldier and a thinker rather than a doer. This was a time of religious revolution in Egypt, and the Pharaoh had to attend to things in the new capital Akhenaton (modern Tell el-Amarna) that were more important to him than the defense of his Asian provinces. Much detailed information is known of this period from an archive of diplomatic letters found at Amarna, many of which are either from or to the princes and the Egyptian officials in Palestine and Syria. In them are seen the continued appeals of loyalists such as Rib-Adda of Byblos to the Pharaoh for aid against rebels such as Abdi-Ashirta; rarely is there evidence of any such aid being given. In the circumstances it is not surprising that the spirit of independence and rebellion that had manifested itself in Amurru should spread farther southward also, into Palestine. The Amarna Letters show Labaya of Shechem and Milkilu of Gezer seeking to aggrandize themselves at the expense of their neighbours, while Shuwardata of Hebron is in conflict with Abdi-Kheba of Jerusalem, to the consternation of the Egyptian governor.

The Amarna Letters

This same Labaya is seen in the Amarna Letters to belong to a group called the Hapiru. There are many other references to this group. The same word, or an equivalent, is found in many documents, in many different languages, throughout the Near East, ranging from the 20th to the 12th centuries BC; and, although its precise meaning in any particular context is a matter for scholarly debate, there is general agreement that the word basically denotes a social class rather than an ethnic group and refers to people who were stateless or outside the law —*e.g.*, bandits, mercenaries, vagabonds, or peddlers. Clearly, the majority of such people can have had little influence on the course of history, though some few of

them—Labaya, for example—played an important role. But the word Hapiru is probably related to the word Hebrew (though there is no agreement on this, despite voluminous scholarly discussion), and it is possible that the troublemaking groups who appear under this name in the Amarna Letters are related in some way to the people who later settled in Palestine, as their Promised Land.

The troubles that disturbed Egyptian rule in Palestine were not resolved until the accession of Seti I at the very end of the 14th century; but meanwhile there had been developments adverse to Egypt in the north. Although Suppiluliumas had achieved a spectacular victory in his First Syrian War, it was not a permanent one; Mitannian influence was still strong and was able to foment a revolt of the major cities—*e.g.*, Aleppo, Alalakh, Qatna, and Kadesh—against the Hittites. The revolt was easily crushed by Suppiluliumas (*c.* 1360), and the triumphant Hittite progress, as far southward as Damascus, marked the final end of Mitannian aspirations in Syria, and indeed the end of the Mitannian empire. All Syria fell under Hittite sway, including Ugarit, which hitherto had been friendly if not wholly subservient to Egypt. In the south, the Hittite troops entered Damascus—a clear infringement of Egyptian suzerainty—while Suppiluliumas concluded a treaty with Aziru, Abdi-Ashirta's successor in Amurru, and encouraged him to wrest from Egypt the remaining coastal cities, as far south as Sidon and including Byblos, which had been the jewel of Pharaoh's Asian crown. Just before midcentury, a series of rather half-hearted attempts was made by the new pharaoh, Tutankhamen (reigned 1361–52), to reassert his authority in southern and central Syria. But the Hittite enemy was too strong, and Suppiluliumas consolidated his position on all sides. When Suppiluliumas died in *c.* 1336, all Syria was under firm Hittite control.

The end of the Mitannian empire

So affairs stood for another generation. Seti I, the second pharaoh of the 19th dynasty, shook off the torpor that had afflicted his predecessors and set about re-establishing Egyptian authority in Asia. His first concern was Palestine, which, although technically still an Egyptian province, was so unstable as to be more a liability than an asset. The fact that Seti was able to reach the northern part of the country during his first campaign, however, shows that many Palestinian cities had remained loyal to Egypt. Victory stelae erected at Bet She'an show that the Egyptian army relieved that city, under attack from an alliance of two neighbouring towns, and repulsed some bands of troublesome Hapiru from the mountains of Jordan. This campaign seems to have established Egypt's position once more in Palestine, but Seti's main purpose was to restore its position in the north. In this he achieved a certain measure of success, and, when he died, the effective boundary between the Egyptian and Hittite spheres of authority was probably as far north as Kadesh, while the King of Amurru had been forced to abrogate his treaty with the Hittites. Meanwhile, the Hittite king, Muwatallis, was planning to meet the renewed challenge to his empire. When the new pharaoh, Ramses II (reigned 1304–1237), advanced into Syria to carry on the work of his father, the Hittite forces were massed to meet him at Kadesh, and there they fought (*c.* 1300) perhaps the most famous battle of preclassical antiquity. Although militarily a draw, the battle represented a strategic defeat for the Egyptians, who retired to Palestine; and, although the Egyptians made extended raids back into northern Syria in subsequent years, the Battle of Kadesh had finally killed all chance of a northern Asian province being reactivated by them. Syria remained solidly under Hittite control.

The Battle of Kadesh

The final episode in the history of the Egyptian and Hittite rivalry in Asia was a peaceful one. The threat provided by a rising power, Assyria, led Ramses II and the Hittite ruler Hattusilis III to settle their Syrian dispute once and for all with a treaty of friendship (*c.* 1284) and subsequent dynastic marriages. The text of the treaty, which has survived, makes no mention of a frontier, but it can be safely assumed that the status quo was endorsed and that the effective boundary lay along a line from about Damascus to somewhere north of Byblos. The treaty ended the long series of imperial contests that had been waged over Syria; and, although the Egyptians were left to put down many a revolt in their Palestinian province, and the Hittites to defend their Syrian vassals from the nascent ambitions of Assyria, the Levant was to remain at relative peace until late in the century.

Organization and culture. The cultural history of the Levant during the 16th–13th centuries BC—the Late Bronze Age of the archaeologists—seems to disprove the theory that civilization requires peaceful conditions in which to flourish. Although Syria and Palestine were torn by imperial contests, the period witnessed remarkable achievements in art, architecture, and literature, and the general level of material culture was perhaps higher than it would be again in the region for more than 1,000 years. Before looking at this culture, however, it is necessary to consider the political, economic, and social organization of the region.

Political organization. At the beginning of the period, the basic political unit in both Syria and Palestine was the city-state. In some cases, particularly in the south, these comprised no more than the city itself and its immediate surroundings; but others, mainly in the richer, more fertile north, had already grown into sizable kingdoms, of which Iamkhad, with its capital at Aleppo, and Kinza, with its capital at Kadesh, were among the most important. As the Egyptians and Hittites extended their authority over the area, both allowed this political fragmentation to persist, while at the same time attempting to bring about some sort of centralized control into their own hands. Because of the greater volume of extant texts from Egypt, more is known about the administration of the Egyptian area than about that of the Hittite, and it is difficult to compare the two systems in detail. It seems that in both areas the local rulers became vassals of pharaoh or king, as the case might be, bound by written treaty to him. The local rulers had to help their overlord in time of war, pay tribute and organize forced labour, extradite fugitives and refugees, keep the central government informed of local events, and renounce all rights to an independent external policy. In return, the great king or the pharaoh confirmed them in their kingship and protected them from their enemies. If the local rulers fulfilled their obligations, renewed their oaths of allegiance, and did homage on the appointed occasions, they seem to have been left a considerable measure of independence in the conduct of their own affairs.

This essentially feudal system seems to have suited the Hittite kings; their Anatolian homeland was organized in much the same way—as a confederation of vassal states—and the administrative system of the Hittite Empire was feudalistic throughout. In the Egyptian territories the Asian lands were, in theory, an integral part of the pharaoh's realm, and, although it was clearly impossible in practice to integrate the African and Asian domains under one monolithic system of government, a much greater effort does seem to have been made by the Egyptian rulers than by the Hittite to keep the vassal princes under strict military and administrative control. To this end viceroys and governors were appointed, directly responsible to the pharaoh, and Egyptian Asia was divided into three provinces. In the north was the province of Amurru, covering modern Lebanon north of Byblos and the contiguous parts of Syria; its capital was Sumuru (identified with Tall Kazel, near modern Tripoli). East of this was Upi, extending from Kadesh to the northern part of Transjordan; Damascus was its principal city, but the administrative centre was Kumidu (today the village of Kāmid al-Lawz, in southern Lebanon). Finally there was Canaan, comprising all Palestine (though not Transjordan) and the coast as far as Byblos, where its boundary ran with that of Amurru; the capital of Canaan was Gaza (Tell el-Ajjul), in the south of the country. During the vicissitudes of the empire, the exact boundaries of the provinces changed, and eventually the province of Amurru was lost, as seen above, to the native dynasty of Abdi-Ashirta, while, before the end of the period, Upi seems to have been fused with Canaan to form a single province, sometimes called Huru. General names, impos-

Egyptian Asia

sible to locate precisely, used by the Egyptians for the whole of their Asian lands, were Retenu and Djahy.

The people of the Levant. During this time the population of the Levant began to assume the heterogeneous character it has had ever since. The Hurrian and Indo-Aryan elements that were already well established in the north in the 16th century, side by side with the older West Semitic stock, had, 100 years later, spread throughout Syria and appeared in Palestine. At the time of the Amarna Letters, many of the Palestinian rulers had Indo-Aryan names, while the name of the king of Jerusalem, Abdi-Kheba, is compounded of the West Semitic word for "servant" and the name of the Hurrian earth goddess Khebat. In general it seems that, just as the Indo-Aryans had established themselves centuries before in Mitanni as the ruling aristocracy of the Hurrian people, so now in southern Syria and Palestine they had achieved a similar position vis-à-vis the Semitic population. Many of the names used in the Old Testament to denote the pre-Israelite inhabitants of Palestine—*e.g.*, Hivites, Jebusites, Girgashites—can be assumed to refer to Indo-Aryan groups, while the Horites are almost certainly the Hurrians. Palestine never became so permeated with non-Semitic peoples, however, as did the region north of Damascus, as is proved by the language of the Amarna Letters. This is Akkadian, the diplomatic language of the time; but the letters written from Palestine are full of West Semitic (Canaanite) words, expressions, and syntax, reflecting the native tongue of the scribes. Far more Hurrian influence is discernible in the letters and other documents from farther north.

Agriculture and stock raising

The economy. The basis of the Levantine economy was agriculture and stock rearing. Although not as fertile as the Egyptian and Mesopotamian alluvial valleys, parts of Syria and Palestine were very favourably suited to the production of certain commodities, particularly grapes and olives on the seaward-facing hill terraces, barley and wheat on the plains of central Syria, and sheep and goats on the poorer steppelands of the interior. Cattle and (after their introduction early in the 2nd millennium by, probably, the Hurrians) horses were also raised. These products were traded widely. The northern part of the region was generally richer agriculturally than the southern, a fact reflected in the greater size and importance of the Syrian city-states as compared with those of Palestine. But this inequality should not be exaggerated. The vast booty taken by Thutmose III at the Battle of Megiddo included great numbers of cattle, sheep, and goats, and, even when allowance is made for Egyptian hyperbole, this suggests that the larger cities of Palestine were by no means economically backward.

The prosperity of the Levantine cities depended mainly upon trade, not only in agricultural products but also in raw materials, especially the timber of the Syrian and Lebanese mountains. The Syrian and Palestinian merchants were well placed to handle—or to hinder, if they so wished—the caravan traffic (which at this period still relied on donkeys) between Anatolia, Mesopotamia, and Egypt, and the relative status and prosperity of the cities varied according to the extent to which they controlled the major routes. Trading was nothing new to the Levant, and archaeology reveals evidence of such commerce from earliest times. In the 3rd millennium a flourishing sea traffic had developed along the Levantine coast, blessed with many natural harbours; and a new dimension was added to this maritime activity in the 2nd millennium, when contact was made with the Aegean world. Already as early as 2000 BC some exceptionally favoured towns on the Syrian coast—*e.g.*, Byblos and Ugarit—had had relations with Crete. But it was not until just before the beginning of the period just described that trade with the West became important, with progress in marine architecture and navigational science making longer voyages possible. As evidence for this, from about 1600 BC Cypriot pottery is found in increasing quantities on all excavated sites; it is accompanied in the 14th and 13th centuries by beautiful Mycenaean ware, probably exported from Greece itself. The importance of the Western trade is shown by the fact that colonies of Minoan, and later Mycenaean, merchants were to be found at Ugarit and probably elsewhere. Egyptian tomb paintings show that Palestinian and Syrian merchants commonly visited, and perhaps resided in, Egypt, and it is likely that some were also found in the ports and marts of Cyprus and the Aegean.

The expansionist policies of both the Hittite and the Egyptian states were largely the result of their desire to control the Levantine trade and to secure the natural wealth of the region for themselves. At the same time, the increasing needs and demands of the great empires produced very favourable economic conditions in Syria and Palestine, and in general terms the Late Bronze Age in the Levant can be considered a time of economic boom. It has been suggested, it is true, that the oppressiveness of the Egyptian administration, in particular, resulted in the impoverishment of Palestine during the latter part of the period; but, although there were undoubtedly economic fluctuations and although certain cities clearly suffered from the punitive measures of the pharaohs, it is likely that this suggestion is based on inadequate archaeological and textual evidence and that the general level of culture and prosperity at the end of the period in Palestine was higher than it had been at the beginning, as was clearly the case in Syria.

Economic development

Culture. The Levant of the Late Bronze Age was thus a region of prosperous commercial and agricultural city-states, normally in some sort of vassalage to either the Hittite or Egyptian empires, populated by a mixture of Semitic, Hurrian, and Indo-Aryan peoples, and constantly visited by foreign soldiers, administrators, and merchants. The hallmarks of both its material and its higher culture were therefore cosmopolitanism, eclecticism, and diversity. This last should be stressed first, for it explains why any attempt to generalize about the appearance of the Levantine towns of the period, or about the paraphernalia of life within these towns, is doomed to failure. From one end of the region to the other there was a bewildering variety of architectural styles, building materials, fashions in pottery and other utensils, burial customs, and the like, which defy categorization within the limits of a short article and which can only be studied in the archaeological handbooks. Temples, for example, varied in plan from the small rectangular structure outside the walls of Lachish (Tel Lakhish, in central Palestine), through the amorphous, extensive "courtyard" temple of Bet She'an and the massive monocellular "Fortress Temples" of Megiddo and Shechem, to the tripartite sanctuaries of Hazor and Alalakh, on which the Temple of Solomon in Jerusalem was later modelled. Even the siting of the temples varied from place to place: on the acropolis at Ugarit, for example, but in the centre of the flat "lower town" at Hazor. At Jericho the dead were commonly buried in rock-cut family vaults outside the city; at Ugarit, in beautifully constructed masonry chambers, perhaps of Aegean inspiration; at Megiddo, in small individual graves, beneath and between the houses. There was, perhaps, greater uniformity in the styles of such utilitarian objects as pottery, tools, and weapons, though even here it is possible to discern regional differences.

In face of such diversity, any brief review of the material culture of the Late Bronze Age Levant must confine itself to an indication of some of the more important surviving monuments and objects. Pride of place must go to the remains excavated at Ugarit, for these show what heights the Canaanite artists and architects could attain.

The royal palace at Ugarit

The royal palace at Ugarit—the residence of the king and the administrative centre of the state—was, in the 14th and 13th centuries BC, one of the most imposing and famous in all western Asia; the King of Byblos wrote his Egyptian masters that only the palace of Tyre could rival it. At its greatest extent it covered an area of some two and a half acres (one hectare); but it originated as a much smaller structure. In time the palace building was repeatedly enlarged; the final building was amorphous and irregular but of great charm and luxury. Many of its features—*e.g.*, courtyards, burial chambers, piped water, and drainage—were to be found also in the houses of the well-to-do inhabitants of the city, who lived around

the palace; their houses confirm the impression that Ugarit was a very wealthy city.

Nowhere else in the contemporary Levant is such luxury encountered, but this must be partly due to the chances of discovery: Tyre, at least, must have been as rich, if the King of Byblos is to be believed. At nearby Alalakh the slightly earlier palace of King Niqmepa was a much smaller and compact edifice, though based on the same basic plan of portico, main room, and courtyard. Only the lower parts of the walls were of stone—rubble faced with basalt slabs—and the superstructure was of mud-brick, strengthened by wooden beams. This style of construction may have been of Anatolian origin, though it had appeared in Syria, at Alalakh, as early as the 18th century, where King Yarim-Lim's palace is constructed in like fashion. It is found as far south as northern Palestine —*e.g.*, at Hazor and Megiddo—where the quality of the masonry in the temples and fortifications is little inferior to that of the Syrian towns. Farther south, the greater prevalence of simple mud-brick construction, with or without stone foundations, may suggest an indigenous tradition of building or perhaps one inspired by Egyptian contacts. Certainly the architectural forms and devices of Egypt were well-known in Palestine, particularly in the Egyptian garrison and government centres such as Bet She'an, where fragments of Egyptian cornices and papyrus capitals have been found in the ruins of the Late Bronze Age temples.

In Palestine, only at Megiddo has a building been discovered whose plan is at all comparable to the palaces of Ugarit and Alalakh, though it is much smaller than the former of these. Otherwise, the chief surviving monuments of Late Bronze Age Palestine are the temples, of which (as already indicated) there is a wide variety of types, reflecting the diversity of cults that existed.

In other fields of artistic expression, the Levantine peoples of the Late Bronze Age produced work of uneven quality. Sculpture in the round does not seem to have been a favoured art form; little survives, and what does is poor. Greater success was achieved in relief sculpture, though here, too, examples of first-class work are few. The stela of a weather god from Ugarit deserves mention, as does one with two scenes of combat between a lion and a dog from Bet She'an, showing that good quality work was not a prerogative of Syria.

Crafts

If carving in stone was not a Levantine forte, carving in ivory was. The origins of this craft must be sought long before the period being described, perhaps in an indigenous school of wood carving, perhaps stimulated by the ancient bone- and ivory-carving tradition of Egypt, going back to predynastic times. In the early 2nd millennium, bone inlay work was a flourishing craft in Syria and Palestine; the increased prosperity of the Late Bronze Age fostered the use of ivory, which was readily available in the herds of elephants that still roamed the valleys of the Euphrates and the Orontes, where the pharaohs were wont to hunt them for recreation. Ivory was used for ornaments, toilet articles, and furniture inlay, and a variety of techniques—*e.g.*, engraving, high- and low-relief carving, openwork—was developed. Most of the finest examples came from the 13th-century palace of Ugarit, and the actual amount of ivory used there (one openwork tabletop, for example, measures more than three feet [one metre] in diameter) testifies again to the extraordinary wealth of this city. But other specimens have been discovered—*e.g.*, at Megiddo, in a cache hidden under the floor of the palace already mentioned, indicating that such *objets d'art* were also prized in at least the more important Palestinian cities. Many pieces are first-rate works of art, worthy to stand beside anything the ancient world produced. But they are also the most striking evidence for the eclectic nature of Canaanite art, for often, in a single piece, are seen motifs and conventions derived from the art of the surrounding countries, combined with such skill and assurance as to produce a design that is homogeneous and uniquely Canaanite. Much of the originality and charm of Canaanite art lies, indeed, in the way in which Egyptian motifs were "misunderstood" in this way and used in manners that the convention-bound artists of the Nile Valley would have thought impossible (*e.g.*, the constant reproduction in Syro-Palestinian art of the sphinx—always male in Egypt—as female). The habit of looking at things in a different way is a characteristic of Canaanite art and gives it a freshness and unexpectedness unknown in Egyptian art before the 14th century, when, because of the Levantine influence that then began to stream back into the Nile Valley, Egyptian art also shook off some of the rigidity that it had acquired during the long centuries of its existence.

Fine metalwork and the engraving of cylinder seals were also art forms in which the people of Syria and Palestine excelled. The former craft arose, probably under Egyptian inspiration, at the beginning of the 2nd millennium. From the very beginning of the Egyptian–Hittite period, several hoards of fine gold jewelry from Gaza (Tell el-Ajjul) are distinguished by many fine specimens in granulation, cloisonné, and repoussé techniques. The designs, particularly pendants with female figures or heads on them, are strongly Egyptian in effect, though many of the details are un-Egyptian, and there is no doubt that they were made by local craftsmen. Similar ornaments come from Ugarit, but the most famous examples of goldwork from this site are decorated bowls of the 13th century. One of these shows a common western Asian theme—the chariot hunt—but treats it more realistically than did Egyptian and more decoratively than did Assyrian artists; it comes closest to the Mycenaean artists' conception of the hunt, though it is perhaps not quite so fine. Another bowl has a frieze in which animals, natural and unnatural, humans, plants, and magical symbols—all the iconographic gleanings of ancient western Asia—are loosely combined into an irregular, unbalanced design, the very uncanonicity of which defines it as Canaanite. The same could be said of most of the cylinder seals, though, because the vast majority of them come from the northern part of the region (the cylinder seal being a Mesopotamian form rather than an Egyptian, its equivalent in the south being the scarab seal), the influence of Assyrian and Mitannian styles and motifs can more clearly be discerned.

Religion

The religion of Syria and Palestine in the second part of the 2nd millennium also exhibits great diversity of beliefs, mythologies, and practices. From personal names and place-names, from representations of deities and cult scenes on cylinder seals and the like, and from cult objects found in excavations, it is clear that Egyptian, Babylonian, Assyrian, Hurrian, and Indo-European gods were familiar to the Levantine countries. Side by side with the foreign religions, however, there was also a body of indigenous religious beliefs and forms that, while they had much in common with those of neighbouring peoples, were distinct enough to merit the term Canaanite. This Canaanite religion is known primarily from two sources —the Old Testament, where it arouses the ire of the devout Hebrews, and a collection of clay tablets discovered in the excavation of Ugarit, containing part of its mythological and ritualistic literature. This literature, written in the local Ugaritic dialect of the Canaanite language, enables scholars to define the Canaanite pantheon, headed by El, who, however, played a rather passive and withdrawn role in the myths. The most active god was Baal, "the Lord of Earth," "the Victor," a deity associated in the myths with storms and rain and thus with vegetation and fertility. Baal's antagonist in the myths is Mot, the god of drought and death. Other important figures are Horon, god of the underworld, Resheph, the god of pestilence, and Koshar, patron of all kinds of craftsmanship. Of the goddesses, the three most prominent are Asherah, Astarte, and Anath; the first was the consort of El, and the two others apparently were sisters and wives of Baal. The roles played by these and other deities and their relationships with each other vary from myth to myth, thus complicating the task of defining their origins and functions. But the dominant part played by Baal and by the three goddesses is sufficient to establish the fact that Canaanite religion was primarily concerned with fertility. To a much greater extent than the inhabitants of either the Nile or Euphrates valleys,

where agricultural abundance was a product of never-failing annual floods, the Canaanites were at the mercy of an unreliable and fickle climate, whereby the tardy arrival of the winter rains or the unseasonal blowing of desert winds could ruin a crop and therefore a livelihood. The Canaanites desired a regularity of the seasons and of the natural forces of fertility, and their myths—whether stories to explain observed phenomena or the scripts of ritual dramas, enacted periodically to produce desired results (the two most widely held interpretations of the texts)—are overwhelmingly concerned with these matters. It was this emphasis on fertility, stronger than in either Egyptian or Mesopotamian religion, with its concomitant cultic practices of sacred prostitution, orgiastic rites, and perhaps human sacrifice, that provoked the fulminations of the Old Testament prophets.

Yet, although the Old Testament writers reacted strongly against the concepts and practices of Canaanite religion, the Canaanite mythological texts exerted a strong and positive influence upon Israelite literature, and its influence continues to be felt today wherever the Bible is read. The language of Canaanite religious literature is full of imagery and metaphor, which reappears time and again in the Old Testament, as (to cite but one specific example) the likening of Yahweh to a storm god, "The Rider on the Clouds" (Psalms 68:4; Isaiah 19:1), as is Baal in the Ugaritic tablets. Stylistic conventions well-known in the Bible are also derived from Canaanite poetry; most striking is the use of repetitive parallelism, in which statements are made two or three times for greater emphasis:

Canaanite influence on the Old Testament

> Thy right hand, O Lord, glorious in power,
> Thy right hand, O Lord, shatters the enemy.
>
> (Ex. 15:6–7)

So close is the poetic diction of some of the early Hebrew literature to that of the Ugaritic texts that, were it not for its religious content, it could properly be classed as Canaanite. Similar stylistic conventions and imagery also occur in the Homeric poems, and these may well have been inspired by a Mycenaean tradition that itself owed something to Canaanite literature. Thus, both because of its own intrinsic quality and its far-reaching influences, the literature of Syria in the 2nd millennium BC must be considered one of the most important legacies of the ancient world.

But the Levant's most important single contribution to civilization lay in writing—the art that made literature possible. By the beginning of the Egyptian–Hittite period, most writing for business, administrative, and diplomatic purposes was carried on, in the Levant, in the Babylonian cuneiform script, although the Egyptian hieroglyphic script was also undoubtedly known in the areas where Egyptian contacts were strong. Both scripts were cumbersome, however, with their hundreds of complicated signs, many of which often represented several different objects, concepts, and, sometimes, sounds. The busy Canaanite merchants would have preferred a simpler method of writing down their language, and several attempts seem to have been made in the 2nd millennium to invent one, based on the phonetic principle, whereby each symbol represented one distinct sound and nothing else. Such a principle marks a great step forward in the history of writing and stands at the beginning of the true alphabetic system, even though (because of the nature of the Semitic languages, where vowels play a less important role than consonants) the Canaanite systems should strictly be termed syllabaries because each sign stood for a syllable of a consonant and vowel.

Alphabetic system

The Ugaritic mythological texts were themselves written in a syllabary of 30 signs, based on cuneiform symbols; because these still proved tedious to write, this script, although probably widespread in the region, did not gain much popularity. In Byblos, about the middle of the millennium, a script based on the Egyptian hieroglyphs was evolved, but this, too, failed to prosper. Credit must go to the inhabitants of Palestine for inventing a syllabary that was to succeed. Sometime before the middle of the 2nd millennium, cursive characters, at least in part derived from Egyptian, were adopted to represent about 30 of the basic sounds of the Canaanite tongue. The earliest examples of this Proto-Canaanite alphabet (its most common designation) are short words, mostly names, on a few objects such as potsherds, a dagger blade, and a gold ring, and a series of dedicatory inscriptions roughly carved on the rocks of the Sinai Peninsula, in the vicinity of the Egyptian mines. Originally the symbols were still pictograms, though simplified and standing for sounds, but, by the end of the millennium, they had evolved into a linear script, easier to write and to read. By 1000 BC they were in use, with minor regional variations, among the Phoenicians on the Syrian coast and the Hebrews in Palestine. From this linear Canaanite alphabet were to develop the true Phoenician, Aramaic, and Early Hebrew alphabets of the 1st millennium BC; and from these all the alphabets in use throughout the world are derived, directly or indirectly. No other invention in the history of mankind, unless perhaps the invention of the wheel or the discovery of fire, has had so lasting an effect upon the human race.

The coming of the Hebrews. During the course of the events described above, there occurred in Palestine a different series of events, which find no mention in the multifarious historical documents surviving from the great Near Eastern empires but which, without doubt, had a greater effect on the history of the area than did any action of the Egyptian, Mitannian, or Hittite kings. These events were the arrival and settlement of the Israelites; they are recorded only in Hebrew folklore and are deducible only with difficulty from archaeological data. Were it not for the Old Testament, the process by which Israel first became associated with its Promised Land would have passed unnoticed by ancient historians.

Evidence of early Israelite history

The books of the Old Testament that deal with the early history of Israel are not historical narratives in the usual sense of the term. Rather, they are collections of reminiscences, of greater or lesser reliability, written down several centuries after the events they describe and rearranged and edited many times since. As they stand, therefore, they contain many anachronisms, contradictions, and obscurities; and, although modern scholarship has made progress in solving some of the more important problems of interpretation, there are probably as many interpretations of detail as there are Old Testament scholars. Even on the main outlines and chronology of Israel's early history there is not complete agreement, and, although the summary that follows seems to reflect a certain consensus, it should nevertheless be borne in mind that individual scholars of great repute have very different views on major points, of which limitations of space prevent discussion.

In the Old Testament, the earliest representatives of the Hebrews are the patriarchs, who are said to have been of Aramaean stock and to have come to Palestine from the East and who are pictured as leading a wandering life, searching for pasture for their flocks, and occasionally travelling southward as far as Egypt. Some of these statements are open to criticism. The reference to the Aramaeans, for example, is an anachronism, because that people did not emerge until toward the end of the 2nd millennium, when Israel was already settled in Palestine. Nor is there any certainty that the town of Ur, from which Abraham's father is said to have come, can be identified with the Babylonian city of that name. There is no proof, even, that the patriarchs are historical figures at all; many scholars regard them as personifications of tribal or other groups (such as the 12 sons of Jacob almost certainly are), while others see them as former Canaanite gods, shorn of their divinity and incorporated into a new religious mythology. Yet whatever their true nature, there is enough authenticity about the patriarchs—*e.g.*, their legal and social customs and their semi-nomadic way of life, as revealed in the narratives pertaining to them—to make it reasonably certain that the narratives reflect a genuine memory of tribal groups that were present in Palestine long before the invasion of the historical Israelite tribes and whom these later Israelites considered to be their true ancestors. It is not known when these early patriarchal groups first arrived. Most scholars

would place them in the 19th and 18th centuries BC, at a time when it is postulated from other evidence that West Semitic peoples were moving en masse into the Levant from the northeast. But the infiltration of semi-nomadic pastoralists into Palestine is a constant feature of its history and cannot be related to any one period, while the analogies between patriarchal customs and those of other Near Eastern peoples, for example the Hurrians or the Amorites, although they indicate a date before about 1400 BC, are not sufficiently close to establish anything more precise. In short, the first Hebrews in Palestine, as evidenced by the patriarchal stories, could have appeared at any time between about the 20th and the 14th centuries BC.

It is in the 14th century that, in the search for Hebrew origins, one treads firmer ground, though it is only a relative firmness. As mentioned above, there are reasons, mainly philological, for equating the word Hapiru of the Amarna Letters ('Apiru in Egyptian) with the biblical word 'Ivri (Hebrew). Not all scholars accept this, however, and the question must be left open until more evidence is available. Yet there are other reasons for supposing that there were Hebrews in Palestine by the 14th century, closely related to those who eventually occupied the whole country. To understand these reasons requires a glance forward to consider the nature of the event or events that, in the Old Testament, are subsumed under the terms Exodus and Conquest.

Biblical accounts

For reasons described below, it is virtually certain that the major phase of the Israelite occupation of Palestine, with which the Bible associates the name of Joshua, had been completed by the end of the 13th century. A careful reading of the biblical narrative reveals that, although an attempt has been made by later editors to arrange all the events and incidents relating to the Exodus, the wanderings in the desert, and the conquest and settlement of Palestine into a logical and coherent sequence, the attempt is far from successful, and inconsistencies, duplications, and contradictions abound. There are, for example, two conflicting accounts of the route the Israelites took through Transjordan, one (Num. 33) suggesting a direct route through Edom and Moab, the other (Num. 20:14 ff. and Judg. 11:17) telling of a long detour through the desert to the east of these countries. Similarly, in Num. 21 and Judg. 1 the conquest of southern Palestine is attributed to the tribes of Judah and Simeon, while elsewhere it is ascribed to Joshua. Examples could be multiplied. Given the fact that, when studied in detail, most of the various narrative strands have a definite ring of truth about them, it must be inferred that, far from being a single sequence of events under one or two successive leaders, the actual process by which the Israelites reached Palestine and settled therein was complex and piecemeal, extending over a long period of time and involving many different groups, acting independently and with varying degrees of success.

One part of this process could have involved the Hapiru of the Amarna Age, and there is evidence that makes this hypothesis attractive. For example, in the Bible there is no mention of the capture of Shechem in central Palestine during Joshua's campaigns of the late 13th century, yet Shechem was, from the very beginning of Israelite settlement, one of their most important centres. In the Amarna Letters Shechem and its vicinity is shown to be in the control of Labaya, a Hapiru, and it is tempting to assume that this region was already in "Hebrew" hands long before Joshua's time. The archaeological evidence from Jericho may also point to "Hebrew" conquests in Palestine in the 14th century. This site has always been a problem for Old Testament historians because, although the Bible makes much of Joshua's capture of the town, it was almost certainly destroyed not later than 1350 BC and then abandoned, more than a century before the date usually, and on good grounds, given to Joshua. It is true that some scholars, largely on the Jericho evidence, would date Joshua's activities and the activities of his predecessor as Israelite leader, Moses, to the 14th century, while a very few others would consider that the archaeological evidence from Jericho is incomplete or unsound. Yet these solutions fly in the face of other evidence, and a more likely solution might be that Jericho was in fact destroyed by Hapiru in the troubled 14th century but that, because of the strategic importance of the site on the main ford over the river Jordan and because of the symbolic importance to the Israelites of the crossing of that river into the Promised Land, its capture was erroneously attributed to Joshua, who became a semimythological hero to the later Israelites.

The capture of Jericho

If the Hebrews were already conquering and settling parts of Palestine as early as the 14th century BC, as many scholars now believe, then the Exodus and Conquest traditions associated with Moses and Joshua in the Old Testament must be revised. Yet, although scholars can no longer think in terms of one Conquest (and probably not of one Exodus), there is no good reason to doubt the basic historicity of the activities attributed to Moses and Joshua and their companions by the biblical authors. To begin with, although the Egyptian records never mention the presence of Israelites by name in the Nile Valley, there is ample documentary evidence to show that Semites from Asia were constantly entering and settling in Egypt, not only in the Hyksos period (when the rulers of Egypt themselves were Semitic) but at all other times as well. Nothing is more likely than that these Semites included, in the mid-2nd millennium, groups identical with or related to the Hapiru and the later Hebrews and that some of them were in a state of slavery. Nor is there anything unlikely about the Mosaic Exodus, and a date for this event can, with great probability, be given. The Bible tells (Ex. 1:11) that the Hebrews were forced to labour at the construction of the royal Egyptian cities of Pithom and Raamses (Pi-Ramesse). The latter was the former Hyksos capital Avaris, in the eastern Delta, rebuilt by the 19th-dynasty pharaohs, Seti I (reigned 1318–1304) and Ramses II (1304–1237), and called by the latter the House of Ramses. For such large-scale building works the pharaohs would have used slave labour and very likely foreign slaves. It seems virtually certain, therefore, that Exodus 1:11 refers to these activities at Avaris, and it can be reasonably postulated that Ramses II was the "pharaoh of the oppression" and that the Exodus under Moses took place in his reign. A *terminus ante quem* for the event can also be deduced from an extant document from the reign of Ramses' successor Merneptah; an inscription commemorating a successful raid by the Egyptian army into Palestine during the fifth year of the reign (*c.* 1231) states that, as a result of the Egyptian action, "Israel is laid waste and has no seed." The Egyptian word used refers to Israel as a people, not as a country, and the inference is that, at this time, the Israelite tribes were present in Palestine in sufficient strength to be considered a national group but had not yet wrested political control from the existing inhabitants. It is a reasonable assumption that such conditions pertained at about the time of Joshua, and the resultant central date of *c.* 1230 for the Conquest would fit perfectly with a date somewhere in the middle of Ramses II's reign, *c.* 1270, for the Exodus, allowing the traditional 40 years for the wanderings between the two events.

Archaeological evidence

The above reconstruction of events relating to the arrival of the Hebrews and Israelites in Palestine is based entirely upon documentary evidence. Corresponding archaeological evidence is practically nonexistent, but what little there is can at least be said not to conflict with this reconstruction, provided that the problem of the fall of Jericho is interpreted as above. This archaeological evidence concerns the destruction of a number of towns in Palestine at about the time of the Israelite invasion. There is clear evidence that many sites were destroyed then, and a number of them are towns specifically mentioned in the Old Testament as having been captured by the Israelites —*e.g.*, Hazor, Debir (Tel Bet Mirsham), and Lachish (Tell Duweir). Moreover, at these sites the archaeological evidence reveals either a period of abandonment or of very poor occupation following the destruction—circumstances that agree with their having been captured by the Israelites, who had no urban traditions. Convincing though this interpretation is, however, the destruction of

these Palestinian cities could, with equal justification, be attributed to armies other than that of the Israelites: the Egyptians were engaged in punitive raids (as Merneptah's inscription proves), while other invaders, notably the Philistines, were responsible for widespread ravaging of Levantine cities about the same time. As long as archaeologists cannot prove any one particular destruction to be the work of the Israelites rather than that of the Egyptians or the Philistines, the bearing of archaeological data upon the problems of the Israelite conquest of Palestine must be seen as minimal.

The Sea Peoples and the advent of the Philistines. Had all been well with the Egyptian Empire of the late 13th century BC, it is inconceivable that the Israelites would have been able to establish themselves in Palestine, and thus the whole future history of Judaism and of the world would have been different. But all was not well, and the ease with which the newcomers captured towns and settled the country was largely due to the fact that Egypt was weakening and was beset on all sides by enemies, compared with whom Israel was seemingly so unimportant as to merit but one single mention in the Egyptian records (in the Merneptah stele already noted). Other inscriptions of the same pharaoh point to the real source of danger to Egypt and indeed to the whole of the Levant: in the fifth year of his reign, Merneptah records that he successfully defended his kingdom from an attack from the west, in which Libyan tribes were allied with certain "Northerners from all lands," peoples "of the Sea." The ethnic names of some of these latter are recorded, and it appears that they were mostly of Indo-European stock and probably came from the islands and coastlands of the Aegean and southern Asia Minor, though the matter is still highly controversial. Merneptah's success held good for a generation, but in the eighth year of Ramses III (1190), these Sea Peoples were threatening again and on a wider front. In the words of the Egyptians: "The foreign countries made a plot in their islands, and the lands were dislodged and scattered by battle all at one time, and no land could stand before their arms—Khatti, Kode, Carchemish, Arzawa, and Alashiya." The Egyptian statement is confirmed by other evidence, especially archaeological. It was just about 1200 BC that the Hittite Empire collapsed, never to rise again, with its cities in ruins. In Cyprus (Alashiya), southern Anatolia (Kode and Arzawa), and throughout the Levant, site after site reveals evidence of catastrophic destruction—Enkomi, Ugarit, Alalakh, Carchemish, Ḥamāh, Tall Abū Hawam, Ashkelon (Ascalon) Tall Jammah, to name but a few. The entire eastern Mediterranean was in turmoil, and, although it is probable that deep-seated environmental factors, still imperfectly understood (*e.g.*, climate deterioration, famine, or even such geophysical calamities as earthquakes), lay at the root of the trouble, there can be little doubt that the cities were actually destroyed by marauding Sea Peoples, as the Egyptian texts specify.

Assault on Egypt

From these texts it is deducible that the assault on Egypt came from two directions—one group of raiders attacking the delta by sea, the other entering Egypt by land from the northeast, from the direction of Palestine. Two battles were therefore fought, and these are depicted in beautiful reliefs that accompany the commemorative inscriptions at Madīnat Habu. The Sea Peoples are clearly distinguished from the Egyptians by their facial characteristics, dress, and equipment; the most distinctive feature is the helmet, of which two types are shown, one with horns and the other with an upstanding circlet of feathers or reeds. The ships and chariots of the Sea Peoples are also distinctive and suggest Anatolian prototypes, confirming the evidence of the ethnic names and pointing to an Anatolian origin for at least some of the raiders. Not all these chariots were vehicles of war; some are heavy carts and were drawn by slow oxen and carried women and children. This was clearly a migratory movement, therefore, and not merely a pirate raid.

The victory that Ramses claimed over the Sea Peoples was a Pyrrhic one. His effort proved too much for Egypt, and from this time can be dated the effective end of the Egyptian Empire in Asia. The real victors seem to have been that section of the raiders whom the Egyptians called Peleset and whom we know from the Bible as the Philistines. They probably formed the bulk of the migrants who had attempted to invade Egypt by land, for, either just before or immediately after the battle with Ramses, they settled on the southern coastal plain of Canaan, a region that was to become known as the Plain of Philistia and from which derives the name Palestine. Although the Egyptian account of these events makes it seem as though Pharaoh deliberately promoted this settlement, he more likely found himself unable to prevent what amounted to a dismemberment of the Asian province, with even the former provincial capital itself, Gaza, in Philistine hands. For perhaps another two centuries, Egypt continued to maintain isolated garrisons in Canaan —*e.g.*, at Bet She'an and Gezer—but they were poor reflections of former imperial glories, anachronisms in a world that had entirely changed.

SYRIA AND PALESTINE, C. 1200–C. 330 BC

The incursions of the Sea Peoples upset the balance of power in the Levant and enabled other groups, such as the Hebrews and the Aramaeans, to establish themselves. Geopolitically, therefore, the region was fragmented, and the following two centuries saw the foundation of the new states and peoples that were to dominate the history and culture of the 1st millennium BC.

The Philistines. Knowledge of the Philistines after their settlement comes both from the Bible and from archaeology. They ruled as a military aristocracy over a predominantly Canaanite population, and their territory comprised a federation of five city-states—the Pentapolis —each ruled by a king or tyrant. The capitals of these five states were Gaza, Ashkelon, and Ashdod, on or near the coast, and Gath and Ekron further inland. The port of Tel Qasila, on the outskirts of modern Tel Aviv-Yafo, seems to have been the northern extent of Philistine territory; between it and the Phoenicians to the north there seems to have been another group of Sea Peoples, the Tjekker, about whom little is known.

The Pentapolis

The Philistines preserved certain distinctive cultural traits for a century or so. These included their attractive pottery, whose designs and shapes are closely related to contemporary Mycenaean pottery from Cyprus and the Aegean and indicate the direction from which the Philistines continued to draw their artistic inspiration.

Little is known of Philistine religion. Biblical references to flies and bees in the cults seem to indicate Greek and Minoan influences. The Philistine gods, however—Dagon, Ashtoreth, and Baal—are Semitic in origin and must have been taken over by the newcomers, perhaps replacing their own deities. Indeed, perhaps the most striking fact about the Philistines is their rapid assimilation to the Canaanite civilization. Only one Philistine innovation had a lasting effect on the civilization of Palestine: the working of iron. The Bible credits the Philistines with introducing this metal into the country and retaining a monopoly, at first, in its use. The resultant superiority of Philistine weaponry and armour helps explain their early success in defeating Israel. It is from *c.* 1200 BC that the archaeologist's Iron Age conventionally begins. Knowledge of this metal must have been acquired by the Philistines in Anatolia. It remained a rare and valuable metal for several generations, however, and not until the 1st millennium did iron weapons and tools become common in Palestine and Syria.

Beginning of the Iron Age

Until about the mid-11th century the Philistines evinced no expansionist tendencies. Not until the growing power of the Israelites seemed to threaten them did they assume an aggressive role, and the second half of the 11th century saw a decisive struggle. At the Battle of Ebenezer (*c.* 1050), the Israelites suffered a severe defeat in which the ark of the Covenant was captured. A number of Israelite towns were destroyed, and Philistine control was extended into the hill country. By the beginning of Saul's reign in Israel (*c.* 1020), the Philistines were clearly in the ascendancy, and Saul could not check them. Following the Battle of Mt. Gilboa, in the Plain of Esdraelon, Saul

took his own life, and Israel was at the mercy of the Philistines. But David, Saul's successor, defeated the Philistines, and by the end of his reign Philistine princes were paying tribute to him. The Pentapolis ended as a political entity; the Philistine city-states survived for almost three centuries, and there was constant border friction between them and the Hebrew states. In the 8th century, when Tiglath-pileser III of Assyria subjugated the northern kingdom of Israel, he first conquered the Philistine cities; his successor, Sargon II, reorganized northern Philistia into the Assyrian province of Ashdod, while the southern cities remained semi-independent Assyrian vassals. This reorganization seems to have survived into the Neo-Babylonian and the Persian periods.

The Syro-Hittite states and the Aramaeans. The destruction of the Hittite Empire in *c.* 1200 BC coincided with a period of Assyrian weakness. Because of a lack of textual evidence and unreliable archaeological evidence, it is possible to discern only the outlines of the history of the next few centuries.

The neo-Hittites. The southeast Hittite provinces were protected by the Taurus and Amanus mountains from the western invaders who, in conjunction with Sea Peoples, overran the central Anatolian plateau; in these districts some old city-states survived and new ones were created, becoming the legatees of Hittite culture and power. Most rulers of these neo-Hittite states were neither Hittites nor descendants of the Hurrians of northern Syria. The language of their earliest monumental inscriptions is related to the Luwian tongue of the southern Anatolian seaboard, and there must have been a migration of people eastward from the coast to the northern Syrian plains. This influx added another element to the racial mixture of northern Syria.

About 20 of these neo-Hittite states are known. Together they formed an important bloc in the Syrian power structure. At the end of the 12th century Tiglath-pileser I brought about a short-lived Assyrian revival in Syria; but the resistance of the Syro-Hittite states helped frustrate Tiglath-pileser's ambitions, and the re-emergence of the Assyrian Empire was delayed another century.

The Aramaeans. A new force now appeared on the scene—the Aramaeans. The first known mention of these people appears in the records of Tiglathpileser I, where it is implied that they are desert nomads. Their language, Aramaic, was a dialect of North West Semitic, closely allied to Canaanite and Hebrew; they may have originated in southern Syria. The biblical account of wars between the Israelites and Aramaean states in southern Lebanon and Syria indicates that these states had been established south of the neo-Hittite area by the 11th century. During the 10th and 9th centuries, many of the neo-Hittite states were taken and renamed, and south of the Anatolian mountains only the plain of Antioch and Carchemish remained Hittite. A brief period of stability and even prosperity then ensued, only to be ended by the Assyrian recovery (*c.* 900).

The Assyrian king Ashurnasirpal II appeared at Carchemish in 876, met no resistance, and led his army right through Syria to the coast. Some years later, Shalmaneser III, following the same route, defeated a coalition of the northern Syrian states. The southern states were more successful against the Assyrian threat. At Qarqar (near Ḥamāh) in 853, the Aramaean kings of Ḥamāh and Damascus, together with 12 subject princes, sustained heavy losses in a battle that, however, forced the Assyrians to halt their advance. But Assyria could not be thwarted. In 804 Adad-nirari III reduced Damascus and claimed mastery over all inland Syria, taking tribute also from Phoenicia, Israel, and the Philistines. Constant revolts forced the Assyrian rulers to introduce a policy of direct annexation, and by 709 all the former Hittite and Aramaean states had become an integral part of the Assyrian Empire.

Battle of Qarqar

Culture. The motivation for Assyrian expansion west of the Euphrates was mainly economic. The routes west to the Mediterranean, north to Anatolia, and south to Egypt ran through the Syrian plains, and along them flowed valuable trade. The Assyrian tribute lists testify to the wealth of Syria—mainly metals (iron, copper, silver, gold) and timber. The economies of the states that provided this tribute were undoubtedly founded on a solid basis of agriculture and stock rearing.

In the 1st millennium the population of inland Syria was even more mixed than before. Linguistic evidence shows that the predominant Hurrian population still survived, as did the Hittite aristocracy in some areas. But the migration of Luwian-speaking people from the west had added a strong new element, at least to the ruling class, while, most important, Semitic Aramaeans had spread from the south. The region also acquired its share, in time, of Assyrian officials and soldiers and of Phoenician merchants and craftsmen.

This mixture of races is paralleled by and explains the mixture of religions. The Hurrian element was apparently strongest: the gods worshipped in the northern states remained for a long time those venerated throughout the Hurrian lands for more than 1,000 years. The Aramaeans, however, introduced a mixture of East and West Semitic deities into their states. Thus, the religious life of the Syrian towns must have comprised a rich and colourful medley of cultic observances, rites, and beliefs.

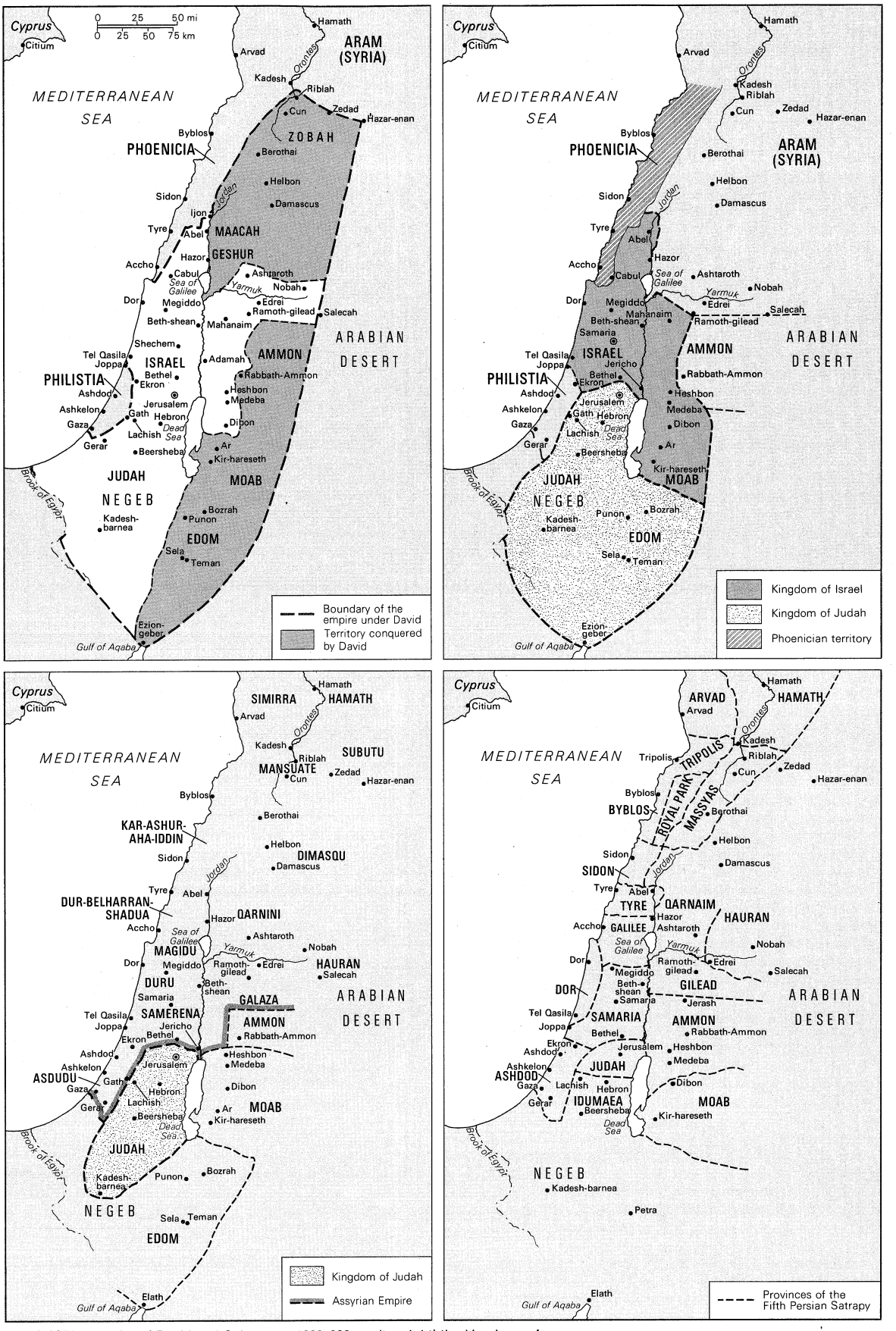
The prosperity of the neo-Hittite and Aramaean states was considerable and gave rise to rich artistic production. A summary of this art is difficult because of the heterogeneity of races and religions and because there is controversy over the dating of much of the evidence. The main points at issue are the date of the earliest neo-Hittite art and the source of its inspiration. Some authorities argue that northern Syrian art arose after the Assyrian expansion in the 9th century and under eastern stimulus; others stress the preservation of the imperial Hittite artistic tradition in some neo-Hittite states and date the earliest work to the 11th and 10th centuries. Whatever the truth, Hittite influence is clearly very strong on some of the artistic productions. In time, however, Assyrian conventions developed, and there thus arose a composite Hittite-Assyrian style. The contribution of the Aramaeans to this art was probably slight; their role was primarily that of patrons (the princes of Damascus and Arslan Tash employed Phoenician ivory carvers, and these and similar artists must have introduced Canaanite and therefore Egyptian motifs into northern Syrian art).

The most important artistic achievement of northern Syria at this time lay in architecture, particularly in the development of a type of building called the *bit hilani* by the Assyrians. Its basic plan seems to have been a Syrian design. It did not become standard in the neo-Hittite period until the mid-9th century, however. From the time of Tiglath-pileser III it became one of the most characteristic elements of Assyrian architecture.

Achievements in architecture

The major cultural achievement of the Aramaeans and their major contribution to world civilization lay in the field of writing. The Aramaeans adopted the Canaanite alphabetic script, which had already developed from a pictographic to a simpler linear form both on the Levantine coast (Phoenician) and in Palestine (Early Hebrew). Their rapid movement throughout Syria and Mesopotamia, as well as their commercial contacts even farther afield, soon spread the new writing and the language that went with it. During the 7th century the Aramaic language and script were widely adopted for international correspondence and even for private communication among the Assyrians. In *c.* 500 BC the Achaemenian kings chose Aramaic as the lingua franca of their vast empire, and it long remained unchallenged over the entire Near East. The Aramaic script is the progenitor of Arabic and of all the alphabetic scripts used in India and the Far East today.

The Phoenicians. Even after the break-up of the Late Bronze Age around 1200, enclaves of Canaanite population retained their independence and culture in northern Palestine and along the coast of Lebanon and Syria. In Palestine the cities of Megiddo and Bet She'an survived for a time; at Bet She'an an Egyptian garrison may still have been stationed during the 12th century, and a local prince may have ruled as late as *c.* 1000, when the town was captured by the Israelites. But the survival of Canaa-



(Top left) The empire of David and Solomon *c.* 1000–930 BC, (top right) the kingdoms of Israel and Judah *c.* 860 BC, (bottom left) the Kingdom of Judah *c.* 700 BC, and (bottom right) Palestine under the Persian Empire *c.* 440 BC.

From G.Wright and F. Filson (eds.), *The Westminster Historical Atlas to the Bible* (rev.); The Westminster Press; Copyright 1956, by W.L. Jenkins, used by permission

nite culture was most significant on the coast, and from it grew the civilization of Phoenicia.

The name Phoenicia derives from the Greek word *phoinix,* meaning "purple"; many scholars believe the name Canaan comes from a Semitic word also meaning purple, because from at least the mid-2nd millennium the Levantine coast was famous for the purple dye of the murex whelk, used in the manufacture of fine cloths. The words Phoenician and Canaanite may be synonymous in origin, and some modern scholars see the Levantine culture of the 2nd and 1st millennia as Phoenician. But, because the word does not seem to have been used in antiquity before the time of Homer, it seems best to confine it to those elements of Canaanite population and culture that survived the late 13th century.

The development of Phoenicia. Canaanite survivors on the Syrian coast, hemmed in by the neo-Hittite and Aramaean states on the north and east and by the Philistines and Israelites on the south, soon looked westward for opportunity and advancement. The high Lebanese mountains that protected the coastal cities from the Aramaean bedouin also restricted their access to the rich agricultural lands of inland Syria. Yet these same mountains provided timber, and the coast was blessed with many good harbours. The prerequisites of shipbuilding and sailing were thus at hand, and Phoenician skill in these things was a legacy from the Canaanite past. The collapse of Egyptian and Mycenaean sea power left only the Sea Peoples as rivals in the eastern Mediterranean. By the time of the visit to Byblos by an Egyptian envoy, Wen-Amon, in *c.* 1075, Phoenician ports were already busy, and less than a century later David's defeat of the Philistines cleared the way for its commercial and colonial expansion westward.

The legacy of shipbuilding and sailing

At first, the leading cities were Aradus, Byblos, Sidon, Tyre, and Akka. Sidon probably dominated until the 10th century (hence the use of the word Sidonians in the Bible and in Homer to denote all Phoenicians), when Tyre became the chief city, perhaps at times heading a confederacy of the less important city-states. Because relations with the Philistines were bad, there developed amicable relations with the Israelites, although some of the southern cities, including Akka, were victims of David's expansionist policy. For a time, even Tyre may have been an Israelite vassal, and it was perhaps in this capacity that its king, Ahiram, sent David timber and craftsmen to help build the palace at Jerusalem. Solomon returned the southern cities and cooperated with Ahiram in various commercial undertakings on an equal footing. Following a period of civil strife in both Israel and Phoenicia, these good relations were re-established by Omri of Israel in *c.* 870 with the marriage of Omri's son Ahab to Jezebel, daughter of King Ethbaal of Tyre.

By the 9th century the threat of the Assyrian army loomed over the Phoenician coast. Ashurnasirpal II received tribute from Tyre and Sidon and the other Phoenician cities in 876, but the first military confrontation apparently occurred in 853, when small Phoenician contingents joined the allied forces that halted Shalmaneser III at Qarqar. Shalmaneser claimed tribute from the Phoenicians later in his reign, however, as did Adad-nirari III when he reduced Damascus in 804. Tiglath-pileser III's policy of direct annexation affected the coastal cities, and, through the internecine strife and anti-Assyrian rebellion, scholars can trace the destruction of Phoenician cities by the Assyrians, the deportation of the inhabitants, and the imposition of direct Assyrian rule over the coast in the late 8th and 7th centuries. When Assyria fell to the Babylonians (late 7th century), only Aradus, Byblos, and Tyre remained autonomous.

Colonization. The Assyrian invasion must have stimulated the Phoenicians to colonize the Mediterranean basin in an effort to find safer homes and new markets. Their colonizing activities had, in fact, started earlier, though exactly when they had begun is controversial. According to classical historians—*e.g.,* Pliny, Diodorus Siculus—colonizing began late in the 12th century: the foundation of Utica (in modern Tunisia) is attributed to *c.* 1100, of Gades (Cádiz, Spain) to *c.* 1110, and of Lixus (on the Atlantic coast of Morocco) to an earlier date. There is no archaeological evidence to support such high dates, however. Even in the eastern Mediterranean, archaeology suggests that the first colonies, probably in Cyprus, do not antedate the 10th century. This, the time of Ahiram and Solomon, was the golden age of Phoenicia and may also have seen the establishment of the earliest western colonies. These western bases were established by the 9th century; the traditional date for the foundation of Carthage, 814, has been adequately confirmed. The 8th and 7th centuries saw many foundations; some of these, especially after the mother cities began to fall, were colonies of Carthage rather than of Phoenicia proper.

The total number of Phoenician colonies is not known, but evidence indicates at least 50. Phoenician expansion was primarily commercial, with no intent of conquest and no large-scale emigration from the homeland. The importance to Phoenicia of this expansion lay in new markets and ports and new sources of wealth, particularly in such metals as copper from Cyprus and silver, tin, and iron from Spain. Its importance to the rest of the world lay in the channel it provided by which the culture of the Orient was transmitted to the West.

Phoenician civilization. Phoenician civilization was a continuation and development of Canaanite civilization. Despite the more exotic produce that came to Phoenicia from its western colonies or from expeditions to Ophir, its main resource was still timber, supplied as tribute to the Assyrian kings as it had been earlier to the pharaohs. Agriculture was also important, with cereals, vines, olives, and dates carefully cultivated on the hillsides and the narrow plain with the help of irrigation. The availability of iron helped exploit these agricultural and natural resources more efficiently. And the relatively small area available for agriculture not only stimulated overseas expansion but also must largely account for the development of industries in the Phoenician cities that were famed in Greek literature. The Phoenicians apparently excelled in textiles, especially the manufacture of cloth coloured with the murex dye, and in glassmaking. Fine metalwork, ivory carving, quarrying, masonry, and carpentry were also notable. In all these industries the Phoenicians simply—and excellently—carried on the traditions of an earlier age and became famous for their skill because of their greater contact with the Greek world.

Phoenician industries

The same may be said of their art and architecture, which shows the same eclecticism and syncretism that had characterized Late Bronze Age art in the region. Thus, in their ivory carving and jewelry Egyptian influence is strong, while on fine engraved metal bowls Mesopotamian and Anatolian influences predominate. This is not to deny originality and genius to the Phoenician craftsmen, who used the artistic devices of their neighbours in new and exciting ways and produced eminently successful works of art. But the Phoenicians' main achievement lay in preserving the artistic traditions of the Bronze Age through the Dark Age at the end of the 2nd millennium and in passing those traditions to the West.

There was an even greater Phoenician legacy, especially to Greece. Just as the Aramaeans spread the Canaanite alphabet to the East, so did the Phoenicians to the West, where it is the source of all modern scripts. By about the 11th century the Canaanite alphabet had been simplified by the use of linear characters, and in this form it became standardized in Phoenicia. The Greeks adopted it, with minor changes, probably around 800. Shortly afterward —either directly borrowed from the Phoenicians or through the Greeks—a very similar script appeared in Italy for use with the Etruscan language, and from there it passed to the Romans and to posterity.

The Hebrews. The only evidence for the history of the Hebrews' first two centuries in Palestine is from the Bible, which presents a collection of legendary and etiological stories impossible to construe correctly. Three things are certain: by 1200 BC the Hebrews were organized into a number of tribes—traditionally 12—bound together only loosely by their common faith in Yahweh; they were culturally inferior to the existing inhabitants of the land; the task they set themselves, one of conquest, was diffi-

cult. The earliest phase of settlement, in Gilead, east of the Jordan, and in the central Palestinian hill country (though not including Jerusalem), was apparently accomplished without much difficulty; but then the Hebrews began to compete for the land with the powerful Canaanites and Philistines. Thereafter successive crises gave rise to successive leaders—called Judges in the Bible—who united the tribes and delivered them from danger; such deliverers were Barak, who with Deborah defeated a Canaanite confederacy in Esdraelon (Judg. 4–5), and Gideon, who defended the Israelites from an attack by Midianite nomads from across the Jordan (Judg. 6–8). Once a danger had passed, however, the tribes would drift back to their separate camps and villages; only gradually and reluctantly did they realize the need for a more permanent leader and a more unified government. Thus arose the monarchy. Gideon himself was invited to be king, but he refused, doubtless because the concept of an earthly monarch was inconsistent with Israel's covenant with God. But by *c.* 1020 Saul had been made king. Opposition to the idea of kingship was still strong, however, and Saul clearly lacked the support needed to defend Israel against the Philistines. Moreover, a young and popular rival had appeared, David, of the tribe of Judah, who, though Saul's son-in-law, incurred the King's jealousy and was forced into exile, serving for a time in the Philistine army as a mercenary. Upon Saul's death David, with Philistine approval, established himself as leader of a coalition of southern tribes; and for a few years bloody rivalry ensued between him and the leaders of those tribes, mostly in the north, loyal to Saul's son Ishboseth (Ishbaal). In *c.* 1000 David was acclaimed king over all Israel.

The rise of the monarchy

David's reign. Such unification displeased the Philistines, and they launched an attack on the new kingdom near Jerusalem, still a Canaanite city, where David defeated them decisively, forcing their withdrawal to the coastal plain. David then consolidated the union of the two parts of his kingdom by capturing the central town of Jerusalem and making it his new capital. By moving the ark of the Covenant to Jerusalem, David also made it the religious centre of the nation. David then embarked upon a series of campaigns against his neighbours, which rapidly enlarged his territory. By the end of his reign he controlled all the remaining Canaanite cities, the Aramaean city-states of southern Syria, and the whole of the fertile land east of the Jordan, comprising the princedoms of Amon, Moab, and Edom. David ruled from the Gulf of Aqaba to Damascus and beyond; and the Aramaean princes to the Euphrates may have been his vassals, as were the Philistine princes.

David firmly believed in the institution of monarchy and in the centralization of government, and he was wise enough to realize that, although Israel's religion set it apart from the previous occupants of the land, it could not disavow the political and cultural legacy of its predecessors. Thus, he based the governmental machinery and the military organization of his kingdom on those of the Canaanites, though he also preserved the Israelite tribal structure. Nor did he turn his back on the more material aspects of Canaanite culture. The fact that Ahiram of Tyre sent wood and craftsmen to help build his palace in Jerusalem suggests that it conformed to Phoenician and so to Canaanite fashions. In the Israelite literature—for example, in the poetry of this and the immediately preceding periods, which included such masterpieces as the "Song of Miriam," the "Song of Deborah," and some of the Psalms (certain of which David may well have composed himself, as tradition maintains)—are found close analogies in imagery and diction to the Ugaritic epics. Thus, by the time of David's death (*c.* 960), Israel had been transformed from a loose confederation of tribes into a powerful empire and had, along with the Phoenician and Aramaean states, adopted the cultural mantle of the old Late Bronze Age civilization of Syria and Palestine.

Solomon's reign. David's latter years were troubled by intrigue and violence, mainly caused by the problem of the succession. But in a final act of statesmanship David had Solomon anointed king, and for a short time father and son reigned together as co-regents. On David's death Solomon acted swiftly and ruthlessly to secure his hold on the throne. The state was preserved and entered what is commonly regarded as its golden age.

Solomon's reputation for wisdom is clearly founded upon his genuine skill in diplomacy and in his ingenuity in handling difficult situations. Having inherited an overly large state, comprised of many disparate parts, he saw that its survival depended on internal consolidation and economic development. His reign was not marked by costly military campaigns, not even when the Edomites or the Aramaeans of Damascus revolted, though these acts reduced the boundaries of Israel. Solomon built his foreign policy upon amicable relations, promoted by a number of dynastic marriages, which introduced many foreign princesses into his harem. Commercial relations were promoted, especially with Tyre. Trade played a large part in Solomon's conception of government; he exhibited interest in the overland route to southern Arabia, as the story of the visit to Jerusalem by the Queen of Sheba (Saba', in southwest Arabia) proves. He also traded with the north, particularly with Asia Minor, where he bought horses for his army. Though his rule was not militaristic, Solomon realized that his government needed the twin supports of centralized administration and a standing army (which for the first time in Israel was equipped with chariots). As for the administration, Solomon broke completely with the past; he reorganized the entire kingdom into 12 provinces, few of which bore any relation to the old order. This reorganization led to more efficient government, especially in tax collection and the imposition of forced labour, essential though much disliked features of Solomon's government.

Palestine was again a wealthy commercial, cosmopolitan country, as is reflected in its material culture, in which the indebtedness to Canaan and to Phoenicia is clear. Particularly is this true of Solomon's famed architectural monuments. The style of fortifications and administrative buildings he erected owes much to comparable structures of the Late Bronze Age in Syria and Anatolia. Even more striking, the Temple he built at Jerusalem (thereby carrying out a plan his father had been forbidden by prophetic injunction to undertake) was, from its description in the Bible, planned and executed in traditional Canaanite manner. Whether Solomon obtained all his ideas and his craftsmen for his Temple from Phoenicia, as is usually assumed from the account of the help he received from Ahiram, or whether there still survived in Palestine a tradition of such art and architecture, is debatable. What is important is that the Temple, with its adjacent palace and administrative buildings, must have seemed to the poorer Israelite peasantry and tribesmen more the resurrected symbols of the old Canaanite priest-king than the house of God and his anointed.

Israel's culture

Outwardly Solomon's reign was marked by relative peace, prosperity, magnificent achievements in architecture, and probably a flowering of literature (at this time the first attempts were probably made to commit to writing the early tribal traditions, using the Canaanite linear script now common throughout the Levant). But there was much disaffection among the lower orders of society, on whom the burdens of taxation and forced labour bore heavily and who obtained no benefit from the ostentation of the court or from Solomon's prestige among his fellow kings. Particularly in the north, where old antagonisms toward the house of David and hereditary kingship still lived, there was a tendency toward separatism.

Separate kingdoms. Solomon was succeeded (922) in Jerusalem by his son Rehoboam, but the northern tribes, whose requests for administrative reform had been refused, repudiated him and chose Jeroboam, who had earlier led a revolt against Solomon. The united kingdom thus ended and was replaced by two smaller kingdoms—Judah in the south and Israel in the north. The accompanying internal strife enabled the neighbouring vassal princes to revolt, and the Israelite Empire dissolved practically overnight, leaving two second-rank states. Although the actual boundaries fluctuated from time to

time, Israel (until its destruction by the Assyrians in the late 8th century) always included the fertile northern coastal plain and the Plain of Esdraelon and was therefore the richer of the two; its population was diversified, and its boundaries touched those of Phoenicia and the Aramaean states, which often led to warfare but which also exposed Israel more to foreign cultural influences than was the case with Judah, thus making Israel more the true heir of the defunct united kingdom. Judah possessed a more homogeneous population and retained Jerusalem, but it was cut off from the sea by the Philistine cities and comprised only the relatively barren hill country and the Negeb, though for a while it controlled Edom and the Arabian trade route. Judah's relative poverty and isolation helped it to survive, however, until its final destruction in 587 BC.

The dissolution in 922 of Israel's fledgling empire coincided with a revival of Assyria and Egypt, and imperial rivalry for control of the Levant was renewed. The history and culture of Israel and Judah in the 9th–7th centuries follow very closely that of their neighbours; only the Israelites' religion kept them distinct. Thanks to the Bible, scholars know much more about the affairs of the Palestinian states than, for example, about Carchemish, Tyre, or Ashkelon (Ascalon).

External affairs

The external affairs of the two kingdoms were dominated, first, by their own rivalry, particularly by constant friction along their common border, which ran not far north of Jerusalem, Judah's capital. Second, but increasingly important, was the foreign threat, from the north and from Egypt. Shortly after Solomon's death the pharaoh Sheshonk invaded Palestine in an effort to restore Egypt's authority, but he was prevented from following up his success by the chronic weakness of Egypt itself. As it was, both Judah and Israel were devastated but soon revived.

In the north, Israel was concerned mainly with the Aramaean state of Damascus, its rival for the Syrian and Transjordanian trade. This rivalry led the sixth king, Omri (reigned 876–869 or *c.* 884–*c.* 872), to conclude an alliance with Ethbaal of Tyre, marrying his son Ahab to the Phoenician princess Jezebel, thereby restoring close ties with Phoenicia. At the end of the 9th century, Israel profited from the Assyrian defeat of Damascus, and for the next half century it enjoyed a period of comparative peace, coinciding with the long reign of Jeroboam II. Judah benefitted from the relaxation of tension in the north and from the strong rule of Amaziah (reigned 800–783) and Uzziah (783–742). The latter re-established the Red Sea trade, even to the extent of obtaining Phoenician help. The Assyrian danger was never far away, however, and with the Syrian states weakened the next blow fell on Israel (738). Israel was forced to pay tribute; and six years later, having joined Damascus in an abortive revolt against the Assyrians, it was devastated, most of its territory being annexed to the Assyrian Empire. Only the capital, Samaria, was spared; and in 722 Samaria itself fell to Sargon II, its inhabitants deported and replaced by settlers from Syria and Babylonia. The Northern Kingdom ceased to exist.

In the south, Judah had refused to join the rebellion, and it was spared, but only as a vassal state. It remained so for most of the 7th century. Rebellion under Hezekiah led to the destruction of the Judaean cities by Sennacherib in 701. Jerusalem itself survived, and from it the last kings of Judah ruled a much-reduced kingdom, subservient to Assyria until 625, when Josiah shook off the yoke.

Another, more powerful vassal of Assyria also established its independence—Babylon. Nabopolassar of Babylon (reigned 625–605) and his son Nebuchadrezzar II (604–562) overthrew the Assyrian dynasty and took control of its empire. A series of rebellions by the last kings of Judah early in the 6th century led to the Babylonian capture of Jerusalem in 597; the royal house was deported, and a puppet king was installed. He too rebelled, and in 587 the city was captured a second time and utterly destroyed, together with the remaining town of Judah.

The Babylonian exile

The wholesale deportation of the inhabitants of the country to Babylon and elsewhere—the Exile—ends the history of the southern kingdom of Judah and is the most important watershed in the history of the Jewish people, with profound consequences for its future development and its psychology.

Two conflicting tendencies may be noted in the internal development of the Hebrew kingdoms. One was the tendency to make the Hebrew states nations such as others were, and this involved, first, the fostering of the principle of hereditary monarchy. In Judah, where loyalty to the house of David was strong, this was never a serious problem. But in the northern kingdom, with its tribal jealousies, it led to dynastic instability. Only briefly under Omri and Ahab (*c.* 874–*c.* 853) and more lengthily under the Jehu dynasty (842–746) did Israel enjoy stable government; significantly, these times coincided with periods of economic prosperity. Secondly, it involved the desire to become a part of the surrounding world —economically, socially, and culturally. Thus, the rulers of Israel and Judah adopted the administrative machinery of Solomon's state and followed, when circumstances permitted, his example in fostering trade with Syria and with Arabia and in maintaining cultural contact with Phoenicia. The tangible results of this are to be seen most clearly in architecture and art. At about the time of their alliance with Tyre, Omri and Ahab built at Samaria and at Megiddo and Hazor royal quarters and garrison buildings based on Phoenician traditions. At Samaria the biblical reference to Ahab's "House of Ivory" is substantiated by the discovery of hundreds of fragments of beautifully carved and inlaid ivory pieces, direct stylistic descendants of Canaanite ivories of the 2nd millennium, testifying to the wealth and taste of the Israelite court. Less is known of the southern capital, though repairs and improvements to Solomon's Temple and royal buildings probably lived up to the finest precedents. As late as the 7th century fine "Phoenician" masonry was employed for a royal villa at Ramat Raḥel, south of the capital.

The less tangible results of this tendency toward syncretism with the outside world are seen in the Bible. The 8th-century prophets Amos, Hosea, and Isaiah castigate the Israelites for luxury and indolence and for overconfidence, arrogance, and wrongdoing. As prosperity increased, moral, social, and religious standards fell. (Ahab's wife, Jezebel, had brought into Israel the cult of the Phoenician god Melqart [Malkart]; his sister, or possibly daughter, Athaliah, married to Jehoram of Judah, fostered the worship of Baal in Jerusalem; Ahaz, when a vassal of the Assyrians, was compelled to erect an altar to the Assyrian gods in the temple itself.) The real and lasting importance of the tendency toward syncretism, therefore, is that it gave rise, in the prophetic preaching, to the opposite tendency—a rejection of contemporary foreign standards and of worldly values and an insistence upon uniquely high moral and religious ideals. In the prophetic movement Israelite religion reached its apogee, with right conduct and personal morality stressed at the expense of mere formal observance of the cult.

The era of the Assyrian, Neo-Babylonian, and Achaemenian empires. By the end of the 8th century practically all of Syria and Palestine had fallen to Assyria. But constant revolts during the next century show that the people of the region did not readily accept the Assyrian yoke. Hopes that an Egyptian revival would release the Levant from Assyrian rule were not fulfilled, though for a few years (609–605) the pharaoh Necho II controlled the entire region. It was to its vassals that the Assyrian Empire finally fell: Nineveh was destroyed by the Medes and Babylonians in 612, in 609 the Assyrian royal line ended, and in 605 Nebuchadrezzar's defeat of the Egyptians at Carchemish put the western part of the Assyrian domains into Babylonian hands. But Babylonian rule was as oppressive as Assyrian, and revolt following revolt was met with destruction and depopulation. Not until the Neo-Babylonian Empire was itself overthrown by the Achaemenian Persians in 539 was the Levant again to enjoy prosperity.

The main focus during the Neo-Babylonian and Achaemenian periods must again be on Palestine because, thanks to the Bible, scholars know most about that coun-

try; making allowances for geographic and religious differences, it can be assumed that parallel conditions prevailed elsewhere in the Levant. The amount of destruction wrought in Palestine after the fall of Jerusalem in 587 is controversial, but there can be no doubt that a significant proportion of the wealthy and official classes were removed. Archaeology, too, proves that many towns were razed and were not rebuilt for a long time. Perhaps as important was the introduction of new peoples; the Edomites especially moved westward from Transjordan and settled in southern Palestine, in what was later to be called Idumaea, named after them.

The Babylonians. Babylonia seems to have treated the victims of its deportation policy more leniently than had Assyria; some of the exiles became prosperous in their new lives as farmers, tradesmen, and merchants in Babylonia. After Nebuchadrezzar's death (562), the exiled king of Judah, Jehoiachin, was honoured at the Babylonian court. The fact that the Babylonians permitted the exiles to retain their religious and national identity provided conditions in which Israel could take stock of its past and make preparations for a new and formative stage of its history. During the Babylonian Exile many of the historical books of the Old Testament (Deuteronomy, Joshua, Judges, Samuel, Kings) and some of the finest theological writings (particularly the second part of the book of Isaiah) were composed.

Persian rule

The Achaemenids. A new chapter began in 538, when the Achaemenid ruler Cyrus II, having gained control of the Neo-Babylonian Empire, ordered the restoration of the Jewish community and cult in Palestine. Only a few Jews actually took advantage of this, and those who did found too much opposition from the new settlers there and from government officials to rebuild more than the foundations of the Temple. Although the exile was over and Yahweh was once more worshipped in Jerusalem, it was not until Darius I (reigned 522–486) established firm control over the Levant and re-issued Cyrus' edict that the Jewish leader Zerubbabel was able to complete the Temple. Conditions improved gradually, and during the 5th and early 4th centuries the Jewish leaders Nehemiah and Ezra enjoyed excellent relations with the Persian kings and acted as semi-autonomous governors. Nehemiah rebuilt the walls of Jerusalem and began reforms that strengthened and purified Jewish social and religious life. Ezra continued these measures and stressed the reading of the Law, the observance of the sabbath, and the prohibition of mixed marriages. Thus, the separation of the Jews was ensured, and the uniqueness of their faith was maintained.

Outside Judah, the Persian period in Syria and Palestine is one of the most obscure in all their history. It seems clear, however, that it was a time of relative peace and prosperity. The toleration toward the Jews also operated elsewhere. Darius I assigned all Syria and Palestine to one province (or satrapy) called "Beyond the River"—*i.e.*, west of the Euphrates—with a provincial capital at Damascus. This area was subdivided into smaller provinces, including Judah and Samaria. The Achaemenians carefully balanced central control with local autonomy; the governors of the provinces were often natives of the area. Darius' excellent road system, legal reforms, and the introduction of coinage (based on an idea developed in the Greek colonies of Asia Minor in the 7th century) made the Achaemenian Empire one of the most successfully governed in the ancient world. Not only is there no record of rebellion among the Phoenician cities against the Persians during the 6th and 5th centuries, but they actually helped their new masters in a conflict with the Greeks, who were fast becoming Phoenicia's rivals in Mediterranean trade and colonization. Only in the 4th century did the Phoenician cities adopt a pro-Greek policy, resulting in an abortive revolt by Tennes, king of Sidon, in 346.

It is hard to estimate the effect of Persian rule on the civilization of the Levant because of the scarcity of archaeological and literary evidence. A few architectural fragments from Phoenicia are typically Achaemenian, and the standard feature of the Achaemenian palace plan—a square pillared hall—may have influenced later temple buildings of the Roman period, particularly in Syria and Transjordan. The first indigenous coinage to appear in the Levant (in Judah and Phoenicia in the 5th century) was clearly the result of Persian stimulus, though the coins themselves are modelled on Greek counterparts. The main new cultural influences in Syria and Palestine were probably Greek rather than Persian, and perhaps the main contribution of the Achaemenian rulers was the peace and stability that allowed these Western influences to permeate. Settlements of Greeks were already spreading along the coast of Palestine and Syria, and it is in the Persian period that the Hellenization of the Near East can be said to have begun.

SYRIA AND PALESTINE IN THE HELLENISTIC AND ROMAN PERIODS, C. 330 BC–AD 634

Alexander's invasion

The Macedonian conquest. In 333 BC Alexander the Great of Macedon routed the Persian army and invaded Syria. Before pursuing the enemy eastward, he elected to visit Egypt. His march down the Syro-Palestinian coast met with opposition only at Tyre and Gaza (it was at this time that Alexander joined the island of Tyre to the mainland with an artificial mole, so making it the peninsula it is today). Meanwhile, his general Parmenio went to Damascus to secure the satrapic capital, and Macedonian governors were appointed to the provinces. Apparently only Samaria resisted the change of rule; a rebellion in that city was easily put down, and a colony of Macedonian veterans was established there.

Alexander died before he could attend to the organization of his new empire and to the problem of succession, though he appointed Perdiccas regent. No one man, however, could fill the position he had held, and his empire was shared out among his friends and generals—the so-called *diadochoi* ("successors"). Syria and Palestine fell to Laomedon and Egypt to Ptolemy I Soter; but in 320 the latter, anxious to secure his kingdom from attack, seized Syria as well. Meanwhile another general, Antigonus I Monophthalmus, tried to take over the entire empire. The next 20 years saw a struggle of Antigonus and his brilliant son Demetrius I against the other successors, with the Levant changing hands several times. Ptolemy defeated Demetrius at Gaza in 312, enabling another member of the anti-Antigonid coalition, Seleucus I, to establish a kingdom in Babylonia. In 301 Antigonus was killed at the Battle of Ipsus, and a settlement was achieved by which Seleucus occupied Syria and built himself a new capital at Antioch on the Orontes, while Ptolemy retained control over Palestine, Transjordan, and the southern Phoenician coast.

The Ptolemies and Seleucids. Ptolemaic rule of the southern Levant lasted about a century. The Seleucids, who considered this region a part of their inheritance, made several unsuccessful attempts to dislodge the Ptolemies; but not until Seleucid fortunes revived under Antiochus III the Great did the Syrian kings make headway. In 198 the Egyptian army was shattered at Panias, and Ptolemaic rule in Asia ceased; Phoenicia, Transjordan, and Palestine were annexed to the Seleucid Empire.

Roman involvement

Antiochus sought to extend his influence westward, into Greece, but there met with a new power, that of Rome, which had secured its position in the western Mediterranean by defeating Carthage and was beginning to take an interest in eastern affairs. Antiochus' army was defeated by the Romans at Magnesia, and by the Treaty of Apamea (188) he was forced to cede all his territories in Asia Minor beyond the Taurus. Antiochus thus opened the Levant to Roman intervention. In 168, when his successor, Antiochus IV Epiphanes, was about to occupy Egypt to stem a threatened invasion of Palestine, the Roman legate forced him to withdraw to Syria.

After the death of Antiochus IV (164), the entire Levant suffered from internal dissension, family quarrels, and native revolts. Because of Seleucid weakness, the eastern and northern parts of the kingdom fell to Parthian and Armenian invaders; Transjordan and southern Syria (including Damascus) went to the Nabataean Arabs; the Ituraean Arabs founded a principality in Galilee

and Lebanon from which they raided the Phoenician coast, and even individual cities proclaimed their independence and began to issue their own coins. Such turmoil led to the increasing involvement of Rome. First in Asia Minor, where Mithradates VI of Pontus was defeated, then in Syria, where Tigranes II of Armenia was forced to retreat, the Roman legions restored order. In 64 Rome refused to recognize Antiochus XIII as king and annexed Syria. The next year Palestine followed.

Hellenization. The 2nd-century revolts against Seleucid rule were partly a reaction against Hellenism. Greek influence in the Levant had been felt from very early times. But in Hellenistic times the rulers, especially the Seleucids, embarked on a deliberate policy of hellenization in the Near East, founding new cities, Greek in appearance and in their semi-autonomous status. The Ptolemies seem to have been content to leave things much as they were (though they inherited from Alexander the finest Hellenistic city in the world, Alexandria). In their Asian province they founded few, if any, new cities, though they gave new Greek names and new constitutions to some of the older ones. They looked on Palestine as an appendage to their kingdom, to be exploited for the good of Egypt. There was more centralized bureaucratic control of trade, agriculture, industry, and local affairs in Ptolemaic Palestine than in Seleucid Syria, and the level of prosperity, though higher than it had been for centuries, did not reach that attained in Seleucid lands.

But Syria was the heart of the Seleucids' domain, and their missionary zeal in spreading Greek culture therein was unbounded, especially during the first century of their rule. Seleucus I alone founded no less than 16 cities (though not all in Syria); most of them were named Antioch, in honour of his father, Antiochus, or Seleucia, in honour of himself. Little remains of these cities, rebuilt many times since; even those sites that have been excavated—*e.g.*, Jarash, Bet She'an, Antioch, Apamea—exhibit few pre-Roman monuments.

Archaeology thus provides little evidence for the early progress of Hellenization in the Levant. Judging from ancient descriptions and from analogy, however, the first Hellenistic cities in Syria and Palestine were undoubtedly magnificent places, conducive to the sort of Greek life the Hellenistic rulers so ardently hoped to propagate.

The cities testify to the wealth and productivity of the Hellenistic period. Alexander had put into circulation a vast amount of precious metals, which the Persian kings had hoarded, and this provided a remarkable stimulus to trade and industry. Close contacts with the Mediterranean world improved markets, while Ptolemaic enterprise extended commercial relations on the Red Sea and in Arabia. In this prosperity the Nabataean Arab caravan traders laid the foundations of their national state at Petra, while their Syrian cousins did the same at Palmyra.

Many Greek colonists came to the region, bringing the Greek language, philosophy, and religion. As always before, the foreign fashions were partly adopted, partly changed, and partly rejected. This was true of race (many Hellenistic kings and their Macedonian and Greek followers married native women and so helped to mix still further the Levantine ethnic pool), of language (for Greek never replaced Aramaic as the common language of the people), and especially of religion (for in the symbiosis of religions that followed the conquests of Alexander, the oriental cults and deities often proved the more persistent, though frequently under new names and in new temples). Among the educated classes, however, Greek literature and philosophy were generally acknowledged to be superior to anything the Orient had produced. Thus, some of the finest thinkers and writers of the Hellenistic Age were Levantine: Zeno of Sidon, who taught his namesake, Zeno of Citium, founder of the Stoic school of philosophy; Diodorus the Peripatetic of Tyre; Antiochus of Ascalon, one of Cicero's teachers; Poseidonius of Apamea, the historian; Meleager, the poet of Gadara.

The Hasmonean state in Palestine. Many Jews had been attracted by the new Greek customs and had sought to foster them in Judah. A gymnasium was built in Jerusalem, in which Greek games and their attendant religious ceremonies took place. Even young priests wore Greek dress. A rift between these Hellenizers and the more conservative Jews showed itself in rivalry over the office of high priest; the two main contenders during Antiochus IV's reign were Jason and Menelaus. Antiochus not only wanted to support the more Hellenizing candidate, but he also hoped to gain financially by selling the office to the highest bidder. This contest prompted Antiochus to visit Jerusalem on his way back from Egypt in 167. When his soldiers plundered the city and tore down many of its buildings and he himself—in flagrant disregard of Jewish law—entered the Temple, Antiochus could not understand the Jews' reaction to this profanity.

Antiochus was probably a sincere universalist in religious matters, hoping that the syncretism of Greek and Oriental faiths would help unify his realm. To this end he accorded to himself divine status, and he represented himself to his subjects as Zeus on Earth. The monotheistic Jews could not accept this. When, following his devastation of Jerusalem, he turned the sacred precinct over to the cult of Zeus Olympius, proscribed all Jewish practices and observances, and built a fortress overlooking the Temple area, religious Jews had no alternative but to revolt.

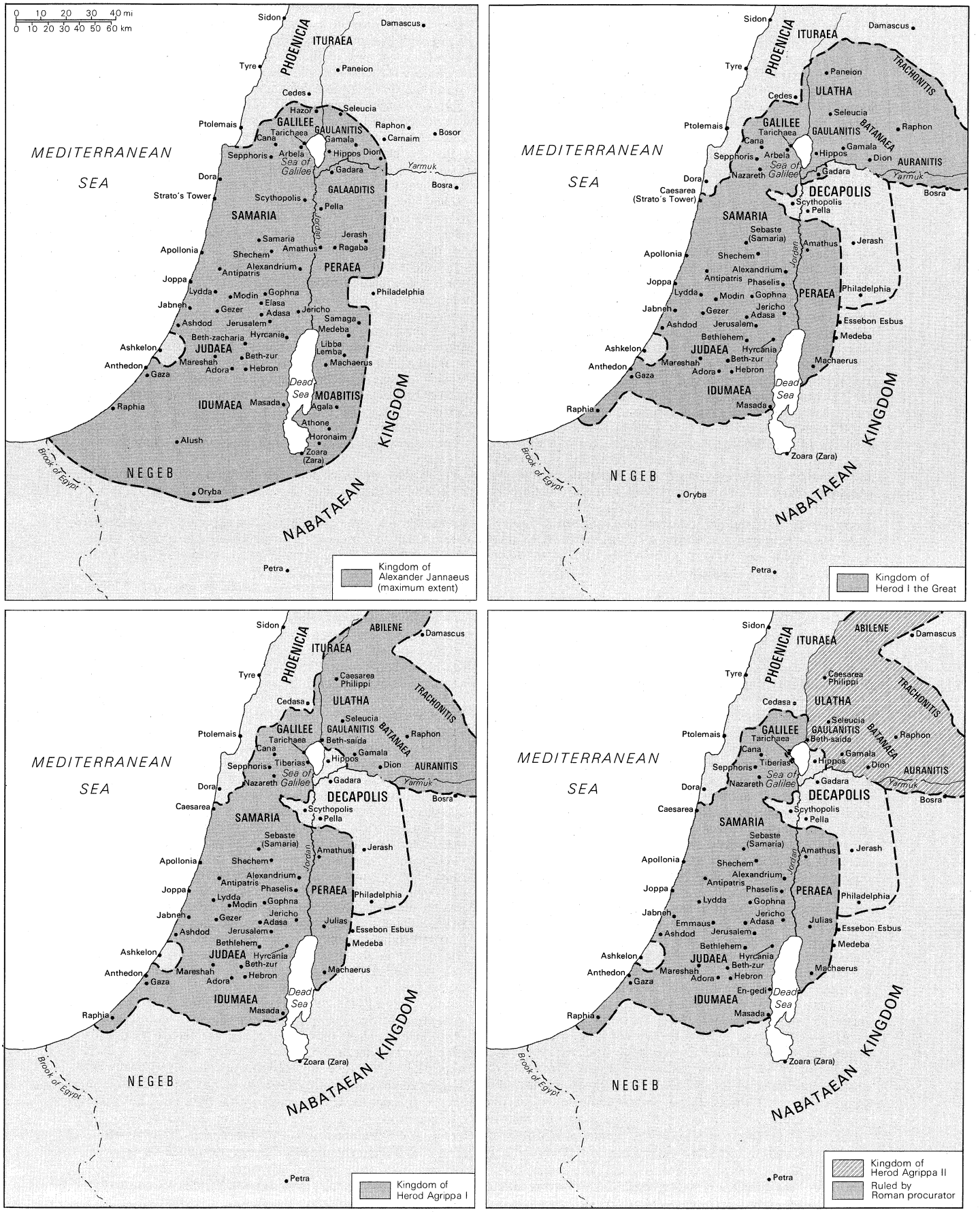
The rebellion of Judas Maccabeus

The rebellion was begun by Mattathias, of the priestly Hasmonean family. At first the rebels confined themselves to tearing down heathen altars and other signs of Greek religion and to harassing the Hellenizing Jews. When Mattathias died (166), he was succeeded by his son Judas, a brilliant military leader called Maccabeus (the Hammer). As the rebel numbers grew, particularly by the adherence of those pious Jews called Hasidim, the conflict became one of total war against the Syrian king. Judas won a number of victories, helped by his knowledge of the terrain and by massive local support. In 164 the rebels took Jerusalem, although the Syrian garrison in its fortress remained. The Temple was cleansed of every Hellenistic feature and was rededicated to Yahweh. The next year Antiochus died and was succeeded by his son Antiochus V, who tried to restore Seleucid authority in Palestine. He reoccupied Jerusalem but made no attempt to renew religious persecution or to profane the Temple. A legitimate high priest was appointed, and a Syrian general was stationed in Jerusalem to uphold the new regime.

With religious freedom won, many pious Jews were prepared to submit once more to alien rule. Not so was Judas; convinced that religious freedom could only be assured by political independence, he continued his war. Judas died in battle in 160 and was replaced by his brother Jonathan. For a time the Syrians were in the ascendancy in Palestine, but at home the dynasty was torn by family rivalry. By skillfully playing off one faction against the other, Jonathan retrieved the Jewish position and was recognized as de facto ruler of Judah. The rival Seleucid princes now vied for his support, and his position became correspondingly powerful: in 152 he assumed the office of high priest (although not of the Zadokite family), and, when Alexander Balas emerged victor of the dynastic conflict in 150, Jonathan was appointed general and joint ruler within the Seleucid kingdom. The struggle did not cease, however; and not until 141 was Jonathan's successor, Simeon, granted independence of Seleucid rule, with the right to mint his own coins. His territory comprised the old province of Judah, a part of Samaria, Joppa on the coast, and a large area east of the Jordan; most important, the Syrian garrison was finally removed from Jerusalem. Thus began the period of independent Hasmonean rule.

Expansion under the Hasmoneans

The period of the next three Hasmonean rulers—John Hyrcanus I (reigned 135–104), Aristobulus I (104–103), and Alexander Jannaeus (103–76)—was a time of territorial expansion. The Jewish rulers conquered practically all of Palestine and Transjordan; only the Greek cities of Ascalon on the coast and Philadelphia in Transjordan and their environs remained outside Jewish control, and the two most important neighbouring states—the Naba-



Palestine (top left) under the Hasmoneans, 168–63 BC; (top right) under Herod the Great, 40–4 BC; (bottom left) under Herod Agrippa I, AD 41–44; and (bottom right) in the time of Herod Agrippa II and the Roman procurator, AD 54–66.

From G. Wright and F. Filson (eds.), *The Westminster Historical Atlas to the Bible* (rev.); The Westminster Press; Copyright 1956 by W.L. Jenkins, used by permission

taean in the east and the Ituraean in the north—were severely weakened. But the Hasmonean state was itself weakened internally by the conflict of the opposing secular and religious tendencies. As the state grew in size and complexity, the old problem reappeared of how to govern successfully without adopting the military and administrative techniques of the outside world. It was impossible, and the Hasmonean kings soon became as manifestly Hellenizers as had been those Jews against whom Mattathias had revolted.

During John Hyrcanus' reign the Pharisees—the Separated Ones—opposed the Hasmoneans both as illegitimate high priests and as Hellenistic rulers. The Hasmoneans looked for support to another sect, the Sadducees. The rivalry of Pharisees and Sadducees and their vying for position within the state contributed to internal dissension that split the Hasmonean house after the death of Jannaeus in 76.

For nine years Jannaeus' wife held the reins of government, but on her death civil war broke out between their sons Hyrcanus and Aristobulus. Hyrcanus was supported by Antipater, an Idumaean, and by the Nabataean king Aretas III. In 65 Aretas and Hyrcanus were beseiging Aristobulus in Jerusalem. The Romans, restoring order in Syria, commanded the Nabataeans to withdraw, thus for the moment saving Aristobulus. The next year the Roman general Pompey arrived in Damascus; taking a dislike to Aristobulus, he reversed the Roman position and gave his support to Hyrcanus. Aristobulus resisted, and the Roman legions marched on Jerusalem; the city was taken with little trouble, but the Temple was fiercely defended and fell only after a three-month siege. The Temple was not plundered, and the traditional observances were resumed on Pompey's command. Hyrcanus was appointed high priest, and Aristobulus was taken captive to Rome. The Jewish kingdom was now under Roman hegemony.

The Romans in Syria and Palestine. Pompey's settlement of the Levant was based as much on the status quo as was consistent with the establishment of peace and order and with maintenance of Roman supremacy. All western Syria and Palestine became the new province of Syria, with its capital at Antioch; the Ituraean and Nabataean states suffered losses of territory and became vassals; Palmyra in eastern Syria was left alone for the moment. The Greek cities were freed from the control of local dynasts, their original semi-autonomous status restored under the general control of the provincial governor. The most important cities east of the Jordan were organized into a league of ten—the Decapolis. Judaea lost most of the territory conquered by the Hasmoneans, and the former kingdom was reduced to its original core in central Palestine, together with Galilee and Peraea. Hyrcanus II remained high priest, but much of the secular power was wielded by his minister, Antipater, under Roman supervision.

The province of Syria

Rome considered the Syrian province so important that it was placed under the direct rule of a proconsul. The first governor, Gabinius, had to suppress a revolt in Judaea of the dispossessed Aristobulus. Under the second governor, Crassus, the Romans were confronted by the Parthian kings of Persia and Mesopotamia; the Roman army was beaten and Crassus slain at the Battle of Carrhae (53). Two years later, the Parthians invaded as far as Antioch but then withdrew. A civil war in Rome between Pompey and Julius Caesar, followed by a period of anarchy after Caesar's assassination in 44, gave rise to unstable conditions in the East; the Parthians invaded again in 40, dislodging the Romans from most of the province. Only with difficulty did Mark Antony restore order. In Palestine, Antipater was made procurator of Judaea and was allowed to make his sons Phasael and Herod governors of Jerusalem and Galilee, respectively. Antipater was murdered in 42, but Mark Antony made Phasael and Herod tetrachs. The Parthian invasion enabled the Hasmonean party, now represented by Aristobulus' son Antigonus, to seize power; and Phasael and Hyrcanus were both captured. Herod escaped to Rome and appealed for help, and the Senate, judging that he would help Rome against Parthia, named him king of the Jews. Herod returned to Palestine with Roman help and began to fight for his kingdom. Jerusalem fell in 37, after an enormous and deliberate slaughter; Antigonus was executed in Antioch at Mark Antony's command, and Herod the Great's reign began.

The kingdom of Herod the Great. Herod's position was not completely secured for several years. Mark Antony was vanquished at Actium (31), leaving Octavian (later Augustus) sole ruler in Rome. Herod met Augustus and convinced him of his loyalty. His kingship under Rome was confirmed, and during the succeeding years he proved his ability as an administrator and in maintaining law and order on Rome's eastern frontier, and he was permitted to make important additions to his kingdom. At its greatest extent, in 20 BC, Herod's kingdom stretched from Idumaea in the south to Galilee in the north and included most of southern Syria and much of Transjordan (Peraea).

Herod ruled as a confederate king, directly responsible to the emperor. He provided auxiliary troops to protect the empire's eastern frontier but paid no tribute, and in internal affairs he was largely independent. He murdered several members of the Hasmonean house, whom he suspected might still threaten his position. His enthusiasm for Hellenistic culture led him to spend more of his country's resources on vast building programs than was economically wise. Jerusalem itself was entirely changed by his massive building programs.

Yet Herod was aware of the peculiar problems of his kingdom, with its mixed Jewish and gentile population, and he probably genuinely wanted to bridge the gap between the two. Within the Jewish parts of his realm he was careful to observe the Jewish law, and thus he managed to prevent a recurrence of a wholesale rebellion against Hellenism and gave Palestine a few decades of relative tranquillity in which the material benefits of Greco-Roman civilization could be felt.

Herod's successors. Herod died in 4 BC, a year or so after the birth of Christ, and his realm was divided between three of his sons: Archelaus received Judaea, Samaria, and the coastal towns; Antipas, Galilee and Peraea; and Philip, southern Syria. But the Romans deposed Archelaus in AD 6 and replaced him with a series of Roman prefects (the most famous of whom is Pontius Pilate). Philip died in 34, and his territory was soon given to one of Herod's grandsons, Agrippa, who had secured the confidence and friendship of the emperor Caligula. In 39 Antipas was also removed, his domain given to Agrippa, and two years later Archelaus' former territory was also given to Agrippa. Thus, until Agrippa's death (44), most of Herod's kingdom was reconstituted. Agrippa ruled much as Herod had, as a rich Hellenistic king in the gentile areas and as a devout Jew in the Jewish. Yet, unlike Herod, he was popular with the Jews (the majority of whom were becoming less averse to Hellenism) and less so with the Romans, who considered him too independent and arrogant. When he died the emperor Claudius refused to accept his son, a minor, as king and turned the realm into the procuratorial province of Judaea.

Jewish resistance

Herod and his successors could not wipe out all resistance to Hellenism among the Jews. There were, of course, varying degrees of resistance. Herod's supporters sought to bridge the gap between the two cultures and thus have the best of both worlds; the Sadducees combined a strict adherence to the Mosaic Law with at least a partial acceptance of Hellenism; the Pharisees were sterner opponents of foreign culture but more flexible in their attitude toward Judaism; the Essenes and other groups hoped for a messiah to deliver them from Roman domination and restore an independent Jewish nation. The Zealots had the most important influence on political events. They condemned anything foreign and insisted that force was the only weapon with which to combat it. At first only a minority, they grew powerful as Roman policy toward the Jews changed. The mad Caligula reversed the formerly tolerant Roman policy by attempting to install his own statue in the Temple, and only his timely murder saved the day. After Claudius' inauguration

of the province in 44, conditions worsened under a series of greedy and incompetent procurators, and in 66 revolt broke out (the First Jewish Revolt—66–77).

The governor of Syria could not quell the uprising, and the emperor Nero entrusted the campaign to the veteran commander Vespasian. With superior forces Vespasian slowly and effectively subdued Galilee and Judaea. When he was recalled to Rome to become emperor himself, he gave his army to his son Titus. In 70 Titus captured and completely destroyed Jerusalem with great slaughter. In 73 the last flames of revolt were put out at Masada, where the last rebels committed suicide rather than fall into Roman hands.

The Jewish nation was not destroyed, however, and the Romans behaved with remarkable moderation. Except in Jerusalem and a few other cities, little material damage was inflicted, and life went on very much as before. Even in the capital, where a Roman legion was stationed, Jews continued to live among the ruins. Although the Romans strengthened their hold on the country and to this end elevated it to the status of a praetorian province, they continued not to interfere in purely Jewish affairs.

In 115–116 rebellions by the Jewish communities in Egypt, Cyrene, and Cyprus were only prevented from spreading to Palestine by the savage measures taken by the governor, Lusius Quietus. In 130 Hadrian, an ardent philhellene, rebuilt Jerusalem as a Roman colony, called Aelia Capitolina, with a shrine to Jupiter on the site of the Temple. But revolt broke out again. Bar Kokhba was proclaimed messiah and commander, and the Romans, taken by surprise, were forced to abandon their new colony. The Temple was partly restored and was rededicated to Yahweh. But Roman might was restored in 135, thus ending the Second Jewish Revolt. Jewish captives were executed or sold into slavery, and the survivors were forbidden to assemble for prayer, practice circumcision, or visit Jerusalem. The colony of Aelia Capitolina was re-established, and Judaea was renamed Syria Palaestina.

Roman Syria and the Palmyrene episode. The rebellions faced by the Romans in Palestine were exceptional for their eastern domains; elsewhere, the population of Syria was glad to accept the benefits of Roman government and the security it brought. Following the wars of the 1st century BC, the security of the eastern frontier was entrusted to a number of client kingdoms: Nabataea and Judaea in the south, Palmyra in the east, Commagene and Cappadocia in the northeast and north. By the early 2nd century AD, these kingdoms had all been absorbed into the empire, and Hadrian strengthened the imperial border with Parthia by a network of roads and fortresses, the remains of which are a conspicuous feature of the landscape of western Palestine and Syria today. In about 193 the emperor Lucius Septimius Severus reorganized the Syrian province, dividing it into two by making the Phoenician coast the separate province of Syria Phoenice.

Client kingdoms

In the 3rd century the Parthians were replaced as rulers of Iran by the more aggressive Sāsānians, who in 260 overran northern Syria, destroyed Antioch, defeated the Roman legions at Edessa, and captured the emperor Valerian himself. From this crisis the empire was saved by a native Syrian prince, Odaenathus of Palmyra. Although within the empire Palmyra had been granted a measure of independence by Septimius Severus, and a line of native rulers had emerged, honoured by the Romans for their loyalty. Odaenathus succeeded in throwing the Sāsānian forces back across the Euphrates and laid siege to the Persian capital at Ctesiphon. In gratitude for his services, Rome granted him the title *dux orientalis* ("commander of the East"). He died soon after, leaving a widow, Zenobia, who exceeded him both in talent and ambition. Shaking off Roman tutelage altogether, she assumed the title Augusta and, in the manner of a completely independent ruler, invaded Anatolia, western Syria, Palestine, and even Egypt, where a Palmyrene garrison was stationed in Alexandria. Her "empire" was short-lived; in 272 Aurelian restored Roman authority throughout the eastern provinces, destroyed Palmyra, and sent Zenobia to Rome as a prisoner.

The effects of Roman rule. Syria, one of the wealthiest of the Roman provinces, was an important supplier of grain to the rest of the empire. Fruits, nuts, and other tree products were also sent to Rome. Palestine was, as always, less favoured agriculturally but shared with its northern neighbour the benefits of a thriving eastern Mediterranean trade. Commerce continued to be the primary activity of the Levantine peoples; the many Syrian merchants who settled in Rome and other western cities brought eastern culture and religion with them.

The process of Hellenization continued under the Romans; Greek continued to be the language of a flourishing literature and philosophy. The emperors, especially Hadrian, Antoninus Pius, and Septimius Severus, founded and embellished cities (Antioch, Baalbek, Palmyra, Jarash, Bozrah, and Petra) in the manner of Alexander and the Seleucids. Although these were given some measure of Roman administrative machinery, they remained both in external appearance and in internal character essentially Greek. As under the Seleucids, Aramaic remained the language of the common people.

Culture under the Romans

Syria and Palestine in the late Roman and Byzantine periods, c. 300–634. *The Christian Golden Age.* The 3rd and 4th centuries were, on the whole, as prosperous as the two preceding ones, and the Byzantine period was to prove a Golden Age for Syria, and especially for Palestine. The emperor Constantine I (306–337) embraced Christianity and made it the favoured if not the sole religion of the empire. As protectors of the new religion, he and his successors evinced a special interest in Palestine, the land of its origin. Constantine's mother, Helena, visited Jerusalem, beginning the cult of pilgrimage to Christian shrines, which contributed greatly to the country's prosperity. The desert fringes of the country in the south and in the east, settled in Nabataean times but then neglected during the 2nd and 3rd centuries, became the scene of intense agricultural and urban development. Churches were built everywhere in the existing towns and in numerous new villages. Syria, especially the north, between Antioch and Aleppo, shared in this prosperity.

Christianity was not without its rivals, both in Syria and Palestine. In the north, pagan cults and Greek philosophy persisted; in Palestine, the Jews, although suffering from the official recognition given to their Christian rivals and (after 428) from exclusion from civil and military posts in the empire, benefitted from the prosperous times and continued to build synagogues in defiance of the law; these, while less assuming and conspicuous than the churches, vied with the latter in the splendour of their interior decoration, particularly in mosaic work, the art form *par excellence* of the period.

Sāsānian and Muslim invasions. Secure from attack in the 4th and 5th centuries, Syria and Palestine suffered from renewed Persian inroads in the 6th century. In 540 the Sāsānian king Khosrow I invaded Syria, sacking Antioch, Apamaea, and other Christian cities. The Byzantine emperor Justinian strengthened Syria's defenses following this incursion, but in the early 7th century the Sāsānians ravaged Syria once more and extended their activities farther south; Damascus and Jerusalem were both taken (614) and Egypt was occupied by the Persian army (616). Byzantine fortunes revived when the emperor Heraclius invaded Persia and forced the Sāsānians to sue for peace. In 629 Heraclius entered Jerusalem in triumph and restored the "True Cross," carried off by the Sāsānians 15 years previously. But while Heraclius was ridding his empire of the Persian foe, the southern borders of that empire were being harassed by groups of Arabs. This was nothing new and hardly gave cause for concern. These Arabs, however, were Muslims, followers of Muḥammad, who had founded a new religion in the Hejaz. Hardy fighters, united in their creed, and boundless in their ambition, they permanently removed Syria and Palestine from Byzantine control in the decade following Heraclius' triumph. (P.J.Pa.)

Heraclius' counter-attack

III. From the Islāmic conquest to 1920

MEDIEVAL PERIOD

The conquest and its sequence. The Arab–Muslim conquest of Syria was accomplished easily. Begun in 634,

it was practically completed by 636. Weakened by the Persian invasion, but, above all, morally and socially alienated by Byzantine domination, by interchurch struggles, and by the ethnic and social gap between a hellenized aristocracy and a subordinate Aramaean population, individual regions never really resisted their new masters and sometimes even welcomed them. As far as is known, there had been Arabs within the borders of Syria for a long time. And accommodation to the new rulers was also made easier by the fact that Islām remained a remote entity—the native population was able to keep its religious freedoms and social customs. The majority of them bent to the new regime and helped to strengthen it by supplying it with administrative staff and by participating in cultural developments. The change in administration resulted in only a partial economic separation from Constantinople, and although coastal ports lost, the towns of the interior gained.

Because it was the passageway to Mecca and to Medina, Syria became the principal political and military centre of the newly conquered area. And its importance in the Muslim empire was further increased when the third caliph (successor to Muḥammad), ʿUthmān, was assassinated by a group that brought ʿAlī (the cousin and son-in-law of Muḥammad) to power; a leading member of ʿUthmān's family, Muʿawiyah, governor of Syria, defeated ʿAlī's forces (657) and became the first of the Umayyad line of caliphs. He kept his seat of government among the Syrian Arabs who had been responsible for his victory; Damascus replaced Medina as the political capital of an empire whose new conquests soon extended from the Aral Sea to the Pyrenees. To the north, only Byzantium resisted, and Syria, as the base for operations launched against it, was of particular strategic importance. The 90 years of the Umayyad dynasty were thus for Syria years of considerable development. Little is known of how Arab–Christian cohabitation was organized in Damascus during the Umayyad period, although it is known that in Syria the Arabs rejected the special town-garrisons (that is, segregation) they had resorted to elsewhere. The evidence suggests that, except under ʿUmar II, Christians were at least well tolerated.

The Umayyad caliphate

The Umayyad period is remembered for cultural, especially architectural, achievements. In Damascus was constructed the Great Mosque, and in Jerusalem the Mosque of ʿUmar and the Dome of the Rock. On the edges of the desert elegant residences permitted the caliphs and the Arab chiefs to link ancestral Bedouin life with that of the new Syrian society. In intellectual life, Arab culture, which was particularly focussed on poetry, was also juxtaposed with the indigenous culture. Meanwhile, both Islām and the Arab language had slowly begun to gain native adepts; they did not, however, expand at the same rate. Even in the second half of the 20th century, in Iran, there are nonarabicized Muslims. And in Syria there are still important Christian communities within a population that is entirely arabicized linguistically.

Under the later Umayyads the strength of the central government declined. There were factions and feuds inside the ruling group: the Arabs of Iraq resented the domination of Syria; the non-Arab converts to Islām (*mawālī*) resented the social gap between them and the Arabs; devout Muslims regarded the Umayyads as too worldly in their lives and policies. After the defeat and death of ʿAlī's son Ḥusayn at the Battle of Karbalā' in 680, sentiment in favour of the family of ʿAlī was still strong. The later Umayyads could not control these discontents. Their rule was finally overthrown and the family virtually destroyed by the new ʿAbbāsid caliphate in 750.

From the 8th to the 12th century. The end of the Umayyad dynasty meant a shift in power from Syria to Iraq. Syria returned to the rank of a simple province, of secondary importance and held in suspicion because of its Umayyad attachments. Even its military role decreased since the war against Byzantium was no longer anything more than a matter of local raids carried out by the frontier people of Cilicia or the Middle Euphrates. Important commerce abandoned it, sucked up by Alexandria, Constantinople, and Baghdad.

The ʿAbbāsid dynasty found itself faced with difficulties that undermined it in its turn. Little by little its most remote provinces began to separate themselves politically, and the army, up to then its source of strength, began to turn into an instrument of disorder.

Starting in the last third of the 9th century and then again, after an interruption, in the second third of the 10th century, Egypt became the seat of certain powerful dynasties: the Tūlūnids, the Ikshīdids, and, most importantly, from 909 to 1171, the Fāṭimids. Syria then became a battlefield for the rival ambitions of these dynasties; first those of the Tūlūnids, who secured control of it from the ʿAbbāsid caliphate; then those of the Ikshīdids and the Fāṭimids, who took control of the centre and the south but had to abandon the north to other Arab dynasties, including the Ḥamdānids and the Mirdasids. The Fāṭimids reintroduced the religious doctrine of Ismāʿīlism, to which their rivals the Qarmaṭians, who were influential in the subdesert zone between Iraq and Syria, also adhered. These two sister but enemy sects came into conflict, notably over the control of Damascus, at the end of the 10th century. The Ḥamdānids were established in the north and had the city of Aleppo as their capital. They engaged in war with Byzantium, but their early successes were followed by the Greek recovery of Antioch and its province in 969. Although practically deserted by the rest of the Muslim world, the leading prince of Aleppo, Sayf ad-Dawlah, was able to mobilize the enthusiasm of the Bedouins, his principal supporters, by organizing the only form of war to which they were adapted—counter-raids in depth against the enemy's rear. He could not put off the end indefinitely, however, and Aleppo became for a century thereafter a kind of protectorate, alternatively under the control of the Byzantines and the Fāṭimids.

The beginning of Egyptian domination

But Aleppo has remained famous in the eyes of posterity, for around the court of the Ḥamdānids lived some of the greatest Arabic writers: the poets al-Mutanabbi and Abū-Firās, the philosopher Al-Fārābī, and the anthologist Abū al-Faraj al-Iṣbāhanī. For a brief period Aleppo became a cultural centre that attracted many people—even the Iraqis.

Certain developments that were to mark Syria for a long time to come occurred during the 10th and 11th centuries. Tribes moving from Arabia or driven from Egypt and Mesopotamia increased the Bedouin element. Among the Christians, a group that had been looked down upon by the other sects retrenched itself on the western slopes of Lebanon. These were the Maronites, who were to become of increasing importance. The Muslim population was affected by the schism found throughout the Muslim world between two Islāmic sects—the orthodox Sunnites and the Shīʿites. Damascus and southern Syria, with Palestine, were primarily Sunnite; Aleppo and the Arabs who dominated the region around it were Shīʿite. At the end of the 10th century, Syria was threatened by the Qarmaṭians, who had established a state in the Persian Gulf. The danger was beaten back, but it returned in a more insidious form as an esoteric doctrine spread by the Ismāʿilis from their centre at Salamīyah (Selemīya) in northern Syria. In southern Lebanon another offshoot of Shīʿism took root: the Druze religion, preached by missionaries sent from Egypt by the Fāṭimid caliph al-Ḥākim.

At the beginning of the 11th century, life in Jerusalem was temporarily affected by Ḥākim's unusual persecution of the Christians, knowledge of which was brought back to the West by Pilgrims. The Church of the Holy Sepulchre was destroyed, but Ḥākim's successors returned to the previous policy of tolerance and, desiring a political reconciliation with Byzantium, even gave the Byzantines a kind of protectorate in the holy places over Christians, particularly those of the Greek rite. In this atmosphere of toleration flourished persons touched by a kind of religious skepticism, such as the great blind poet Abū al-ʿAlā' al-Maʿarrī.

Not of central importance to the great powers, and often

fought over by their dependents or other local potentates, Syria was able to become an area conducive to the development of certain new forms of autonomy. Various types of local militia appeared in many of the Near Eastern towns before the middle of the 11th century. These militias sometimes took the form of secret societies (*futūwwah*) or were sometimes representative of town interests. Of the latter type were the *aḥdāth* (young people) of Damascus, Aleppo, and other towns of Syria and Upper Mesopotamia, who shared effective power with helpless princelings, at least within the town walls. The leader of the *aḥdāth*, the *ra'īs al-balad*, or town leader, was for some time close to being the equivalent of mayor. Tripoli in Syria offers an example of another tendency toward autonomy and independence. Throughout the Muslim world, justice was dispensed locally by the *qāḍī*, who theoretically represented the law but in actuality was closely tied to the local "bourgeoisie." Just as the families of bishops dominated the towns of southern Italy, the family of the *qāḍī* Ibn ʿAmmār founded a prosperous dynasty in Tripoli, one that would probably have lasted much longer if the Crusades had not put an end to it.

The 11th-century coastal towns, such as Tripoli, benefitted from the commercial revival that was taking place in the Mediterranean area, one that involved relations with the Maghrib and, increasingly, with southern Italy. In addition, pilgrims from the West were arriving in greater and greater numbers. Thus Europe and the Near East were in the process of being brought into contact.

Invasion by Turks and crusaders

In the middle of these events Syria fell into the hands of the Seljuq Turks, who had established a sultanate in Western Asia Minor. They occupied Jerusalem and then Damascus, later Aleppo. But after the death of the sultan Malik-Shāh in 1092, the Seljuq Empire fell to pieces; and between 1098 and 1124 the crusaders occupied Antioch, Jerusalem, al-Karak in Transjordan, and the coast. The crusaders organized their conquests into four states, the most important of which was the kingdom of Jerusalem (see CRUSADES). Their situation was precarious, however. The crusaders were always a minority in their states, and they never penetrated far into the interior. They could maintain their position only as long as the surrounding Muslim states were weak and divided. Zangī, the Turkish ruler of Mosul, occupied Aleppo in 1128 and recovered Edessa from the crusaders in 1144. His son Nureddin united inner Syria and annexed Egypt.

Nureddin was also the animator of a program that appealed directly to the pietist sentiments of all Muslims. Under his direction, holy war again became the official and prime task of the regime. Annexation of Muslim territory was presented as necessary to the battle against the infidel, a device that was helpful in discrediting any independently minded princes in the common military action. The triumph of Nureddin's career was the dispatch of an army to Egypt, where the factions around the last Fāṭimids were disintegrating.

Victories of Saladin

Nureddin's campaign was completed by the man who was at the head of the victorious army in Egypt—Saladin. Saladin was one of the numerous Kurds enrolled in the Turkish armies. He put an end to the Fāṭimid rule (1171) and created a strong kingdom of Egypt and Syria. The states established by the crusaders were encircled, and Egypt's resources were put to the service of the holy war. He then defeated the crusaders at the great Battle of Ḥiṭṭīn (1187). He recovered all Palestine and most of the inland strongholds of the crusaders. Soon afterward, however, the Third Crusade recaptured part of the coast, where European rule lingered for another century.

(C.C.)

The Ayyūbids and Mamlūks. After Saladin's death his kingdom was split up among members of his family, the Ayyūbids, who established principalities in Aleppo, Ḥamāh, Homs, Damascus, Baʿlabakk, and Transjordan and ruled them until 1260. The period of Nureddin, Saladin, and their successors was of great importance. Thanks largely to the establishment of Italian trading centres on the coast, and better security, economic life recovered and Syria reached a level of prosperity such as it had not enjoyed for centuries. The Ayyūbid rulers stimulated culture and architecture. Following the Seljuqs, they created a new land system based on the grant of rights over land in return for military service. They were champions of Islāmic orthodoxy against the sects that had gained ground in the previous era. They supported orthodox theology and law and built colleges of a new type, the *madrasah*, as centres of orthodox learning. Their efforts to stamp out heterodoxy were not completely successful. The Ismāʿīlīs kept their strongholds in the mountains, and their secret organization, the Assassins, had some political importance. The Druze religion also spread.

Although strong internally, the state was still in danger from the Bedouin tribes of the desert and from the Mongols, who invaded Syria for the first time in 1260 and sacked Aleppo. They were driven back not by the local rulers but by a new Egyptian military group, the Mamlūks, a self-perpetuating elite of slaves and freedmen, mainly of Turkish and Caucasian origin, who had replaced the Ayyūbids as rulers of Egypt in 1250. In 1260 they defeated the Mongols at the Battle of ʿAyn Jālūt in Palestine; the victorious Mamlūk general, Baybars I, made himself sultan of a reunited kingdom of Syria and Egypt, which he ruled until his death in 1277. This state continued to exist for over two centuries. In 1291 it won back Acre and other coastal towns from the crusaders, who were expelled; and a few years later it took the last crusading stronghold, the island of Ruad (Arwād). The Mamlūks reorganized the Ayyūbid principalities as six provinces, of which Damascus was the largest and most important. Political power was in the hands of the Mamlūk elite, who held land in virtual ownership in return for military service in the cavalry. But there was a local element in the government, the civil servants being drawn mainly from Syrian Arab families with their tradition of religious learning. Local rulers such as the Buhturids of Lebanon were recognized and absorbed into the "feudal" system.

Like the Ayyūbids, the Mamlūks were strictly orthodox in religion. Religious culture flourished and produced a number of great scholars such as the Ḥanbalī jurist Ibn Taymīyah. For religious and political reasons, the Mamlūks dealt severely with the religious minorities living in the coastal mountain ranges: Druze, Maronite Christians, Ismāʿīlīs, and ʿAlawites (or Nusayrīs; adherents of another creed derived from extreme Shīʿism and living in the Jabal an-Nuṣayrīyah). One of the principal reasons for this severity was the fear of the Mamlūks that these minorities might cooperate with the crusaders, should they attempt to return.

In the early Mamlūk period, Syria remained prosperous; the rulers constructed public works, and Venetian merchants carried on their coastal trade. But in 1401 came a blow to economic life: a new Mongol invader, Timur (Tamerlane), sacked Aleppo and Damascus. His empire did not long survive his death in 1405, but the damage had been done. The cities had been burned, a large part of their population killed, and many craftsmen taken away to Central Asia.

Invasion of Timur

OTTOMAN PERIOD

Ottoman government, 16th–17th centuries. Throughout the 15th century, Mamlūk Syria continued to decline, while a new power was growing to the north, that of the Ottoman Turkish sultanate in Asia Minor. Having occupied Constantinople and the Balkans, it began to look south-southward. In 1516 Sultan Selim I defeated the Mamlūks in the Battle of Marj Dābiq and occupied the whole of Syria that year and Egypt the next. While Egypt later recovered some autonomy under the Mamlūks, Syria remained for 400 years an integral part of the Ottoman Empire. It was divided into provinces, each under a governor: Damascus, Aleppo, and later Tripoli and Ṣaydā, or Sidon, of which the administrative centre was later moved to Acre. Damascus, the largest, had special importance as the place from which the pilgrimage to Mecca was organized every year. The governor of Damascus led the pilgrimage when possible, and most of the revenues of the province were earmarked for its expenses.

The tax system continued in principle to be that of

Muslim law—land tax, poll tax on Christians and Jews, and customs duties. But the Ottomans, like their predecessors, gave the right to collect and keep the land tax in return for military service and thus created a virtual landowning class. Later, this system was allowed to decay, and tax collection was turned over to tax farmers (*mültezim*), who themselves became in the course of time a land-owning class. Another official hierarchy, the religious hierarchy of judges, jurisconsults, and preachers, served as an intermediary between government and subjects.

Within this framework of law, order, and taxation, the local communities were left to regulate their own lives. In the desert, the Bedouin tribes were controlled to some extent by gifts, the encouragement of factions, and occasional military expeditions but otherwise were not interfered with. In the mountains, the Maronites and other Christians, Druze, Shīʿite Muslims of Lebanon, the ʿAlawites, and Ismāʿīlīs of Jabal an-Nuṣayrīyah, were watched by the Turkish governors of Tripoli, Sidon, and Damascus, but they were not interfered with so long as they paid their taxes. The Druze princes of Lebanon, first of the Maʿn and then of the Shihabid family, were recognized as governors and chief tax collectors. In the Jabal ad-Durūz south of Damascus there grew up a community of Druze farmers under no control and paying no taxes. The authority of the Christian patriarchs over their communities was recognized. In the corps of *ʿulamāʾ* (learned Muslims holding government appointments) most positions except the highest were held by members of local families having a tradition of religious learning. They continued to be, as under the Mamlūks, spokesmen and leaders of the Muslim citizens.

Economic life

The early Ottoman governors paid much attention to agriculture, and their fiscal system was designed to encourage it. In parts of Syria it flourished during the 16th and 17th centuries, and, apart from cereals for local consumption, cotton and silk were produced for export. Aleppo and Damascus not only were important centres of handicrafts but also served as market towns for the desert and countryside and as stages on the desert routes to the Persian Gulf and Persia. Aleppo and Sidon were also important centres of trade with Europe; French and English had largely replaced Italian merchants, and there grew up also a class of Syrian Christian and Jewish merchants who developed contacts with Egypt, Italy, and France.

Throughout the 17th and 18th centuries the position of the Christians improved. Catholic missions, protected by France, enlarged the Catholic communities of both Latin and Eastern rites, founded schools, and spread knowledge of European languages. The Oriental colleges in Rome produced an educated priesthood, and the Christian communities in Aleppo and Lebanon brought forth scholars such as the famous Joseph Assemani, Vatican librarian, and Arabic writers such as Germanos Farhat, Maronite archbishop of Aleppo. The traditional Muslim Arab culture could still produce a theologian of some originality in ʿAbd al-Ghanī an-Nābulusī and a systematic jurist in Ibrāhīm al-Ḥalabī.

Decline of Ottoman authority. In spite of a widespread revolt in the early 17th century, Ottoman rule was in general stable and effective until the end of that century. After that it declined rapidly, in Syria as elsewhere. Control by the central government weakened; the standard of administration sank; the Janissaries (the elite troops of the sultan) lost their discipline and became a menace to order. The result was a shrinkage of agricultural production, as the villages suffered from the depredations of soldiery and tax collectors and from Bedouin incursions. This was a period of activity in the Syrian desert, and Bedouin tribes, moving northwest from Arabia, extended their control far into the settled land. In the towns there was also a decline. The desert routes were unsafe, and the European merchant colonies were shrinking. But there was still a vigorous commercial life; the standard of craftsmanship was high, and the great tradition of Islāmic architecture was continued.

Ottoman authority did reassert itself to some extent, but in a new form. For most of the 18th century, Damascus was ruled by governors belonging to the ʿAẓm family, loyal to the sultan but with more independence than earlier sultans would have allowed. They controlled the Janissaries, kept back the Bedouin, maintained security, and sometimes extended their authority to other provinces. In the province of Sidon, power was held on similar terms by a ruthless and able Bosnian governor, Aḥmad al-Jazzār (1775-1804), and his group of Mamlūks. Such rulers raised their own armies, but this involved additional taxation and further depressed the condition of the peasants. Agriculture flourished only in the hilly districts, which were virtually beyond Ottoman control, free from Bedouin attacks, and ruled by strong local rulers who protected agriculture and made Acre a prosperous centre of trade.

At the beginning of the 19th century, the population of Syria was perhaps 2,000,000. There were some islands of prosperity: Aleppo and Damascus (each with roughly 100,000 inhabitants), Mt. Lebanon, and certain other secluded districts. In general, however, the country was in decay, the small towns subsisting on local trade, the villagers receding in face of the Bedouin. The Ottoman hold on the country was at its weakest. At Acre, it is true, al-Jazzār maintained his strong rule and was able, with the help of a British fleet and the Turkish army, to hold back Napoleon when he invaded Palestine in 1799. But in Damascus and Aleppo the governors were scarcely able to control the population of city or countryside. The prince of Lebanon, Bashīr II (1788–1840), who had been installed by al-Jazzār and remained quiet while al-Jazzār was alive, gradually extended his control over districts beyond Lebanon. In 1810 the Wahhābīs from central Arabia threatened Damascus.

Egyptian domination. In 1831 the ruler of Egypt, Muḥammad ʿAlī, sent his son Ibrāhīm Pasha at the head of his modern army into Palestine. Helped by Bashīr and other local leaders, Ibrāhīm conquered the country and advanced into Asia Minor. He ruled Syria for almost ten years. The whole country was controlled from Damascus. There and in the provincial centres the governors were Egyptians, but they were assisted by councils representing the population. In political matters Ibrāhīm relied largely on Bashīr, and in financial concerns on a Syrian Christian adviser. New taxes were introduced and strictly collected, agriculture was encouraged, and the Bedouin pushed back. After an abortive attempt to introduce trade monopolies, Ibrāhīm encouraged European traders by maintaining better security and creating special commercial courts. It was in this period that Beirut became important as a centre of trade with Europe. The Christian and Jewish populations were treated with consideration.

European intervention in Syria

After a time, Ibrāhīm's rule became unpopular because his taxes were heavy and because he disarmed and conscripted the population. The European powers (except France) also objected to Egyptian rule in Syria because it was a threat to the Ottoman Empire, the weakness or disintegration of which might cause a European crisis. In 1839 war broke out between Muḥammad ʿAlī and his suzerain, the Sultan. Ibrāhīm defeated the Turkish army, but in 1840 the European powers intervened. After an ultimatum, a British, Turkish, and Austrian force landed on the Syrian coast; there was a local insurrection encouraged from outside, and the Egyptians were forced to withdraw from Syria, which reverted to the Sultan's government.

Ottoman rule restored. The next 20 years were a period of mounting crises. Bashīr II was deposed by the Ottomans because of his support of the Egyptians; his successor was in turn soon deposed. Lebanon became the scene of a struggle for power between Druzes and Maronites, with undertones of social conflict. In the rest of Syria, an attempt was made to apply the new Ottoman administrative system. But the new system of taxation and conscription caused unrest. This situation was worsened by the growth of European influence; the Muslim majority became aware of the weakness of the Islāmic community, and the connection of France with the Catholics and of Russia with the Orthodox both encouraged the minorities

to hope for a more favourable position and focussed on them the hostility of their Muslim compatriots. There was also economic unrest. From the 1820s, European goods flooded the market and replaced the products of local craftsmen. This diminished the prosperity of the artisan class, largely Muslim, but increased that of the import merchants, mainly Christians and Jews.

The tension thus generated burst out in 1860 when a civil war of Druzes and Maronites in Lebanon touched off a massacre of Christians by a Muslim mob in Damascus. The Ottoman government sent a special commissioner to punish the guilty and suppress disorder. France sent an expeditionary force, and a European commission discussed the future of the country, coming to the conclusion that Lebanon (the mountain itself but not the coastal towns) should be an autonomous district (*mutaṣarrifīyah*) but that no change should be made in the rest of Syria.

From then until the collapse of the Ottoman Empire, Syria continued to be governed as a group of Ottoman provinces. From 1888 there were three: Damascus, Aleppo, and Beirut. The new administrative and legal system was more carefully applied, and a new type of educated official gradually raised its standards. The introduction of railways and telegraphs made possible a stricter control. A French-built railway linked Beirut and Damascus, with a later extension running north to Aleppo, and in 1908 the Hejaz Railway was opened to take pilgrims from Damascus to Medina. Railways and better security encouraged agriculture. Transjordan and al-Jazīrah, the district between the Upper Euphrates and Tigris, ceased to be wholly nomadic. In other parts, however, the growth of population outran the increase of production, and there began a large-scale emigration to North and South America. The trade of Beirut increased steadily; its modern harbour was built by a French company, and by 1914 its population had grown to about 150,000. Aleppo (population about 200,000) and Damascus (250,000) both had a flourishing trade, but the crafts declined, and the desert routes suffered from the opening of the Suez Canal.

In Beirut and other cities there was a considerable change in social life. The upper and middle classes adopted the clothes and social customs of western Europe, and Western-style schools flourished. In 1866 the American Protestant Mission opened in Beirut the Syrian Protestant College (later the American University of Beirut), and in 1881 the French Jesuits opened the Université Saint-Joseph in the same town. The government opened schools, and young men of the great Arab families of the towns began to attend the higher schools in Constantinople and to go on to civil or military service.

Under Sultan Abdülhamid II (1876–1909) the Muslim Arabs of Syria were reasonably content. His emphasis on Islāmic solidarity frightened the Christians, but the Muslims on the whole still regarded the Ottoman Empire as the political embodiment of Islām. Syrian Arabs played a leading part at the Sultan's court.

Emergence of Arab nationalism

After the Young Turk revolution of 1908, relations between Arabs and Turks grew worse. Power fell into the hands of a Turkish military group whose policy stimulated the growth of opposition. Arab nationalist feeling became more conscious, and political parties, both open and secret, were organized by Syrians in Cairo, Constantinople, and Paris as well as in Syria itself. Their explicit demand was for improved status within the empire and some local autonomy, but there was a growing feeling in favour of independence and the establishment of a Syrian or Arab state. This was the aim of the secret Fatāt Party founded in 1911. Many Christians, particularly Maronites and other Catholics, wanted an independent Lebanon with extended frontiers and under French protection. Among the nationalists were some who looked to England and France for allies in the struggle for independence, but others were conscious of the danger of domination by those countries. France had paramount influence on the Syrian coast and in Lebanon, while in Palestine all the powers competed for influence. And in the 1880s there had begun Jewish immigration from Russia and eastern Europe, aiming to create a national home for the Jews under international guarantee.

World War I. When the Ottoman Empire entered World War I in 1914, Syria became a military base. In 1915 an Ottoman army under German command attacked the British position on the Suez Canal, and from 1916 a British and imperial force based in Egypt, with a French contingent, undertook the invasion of Palestine. By the end of 1917 Gen. Sir Edmund (later Field Marshal Viscount) Allenby had occupied Jerusalem and by November 1918 the rest of Syria. Most Christians and Jews welcomed the occupation; among the Muslims a large proportion had remained loyal to the empire, as being all that was left of the political independence of Islām, but the nationalist societies had made common cause with the ruler of the Hejaz, Sharīf Ḥusayn, forming an alliance with Britain against their Turkish suzerain. An Arab army under the command of Ḥusayn's son Fayṣal was formed in the Hejaz, with Syrian and other Arab officers and British help led by T.E. Lawrence. It took part, under Allenby's general command, in the Syrian campaign, operating on the east flank in Transjordan and helping to capture Damascus.

Allied vacillation over Syria's future

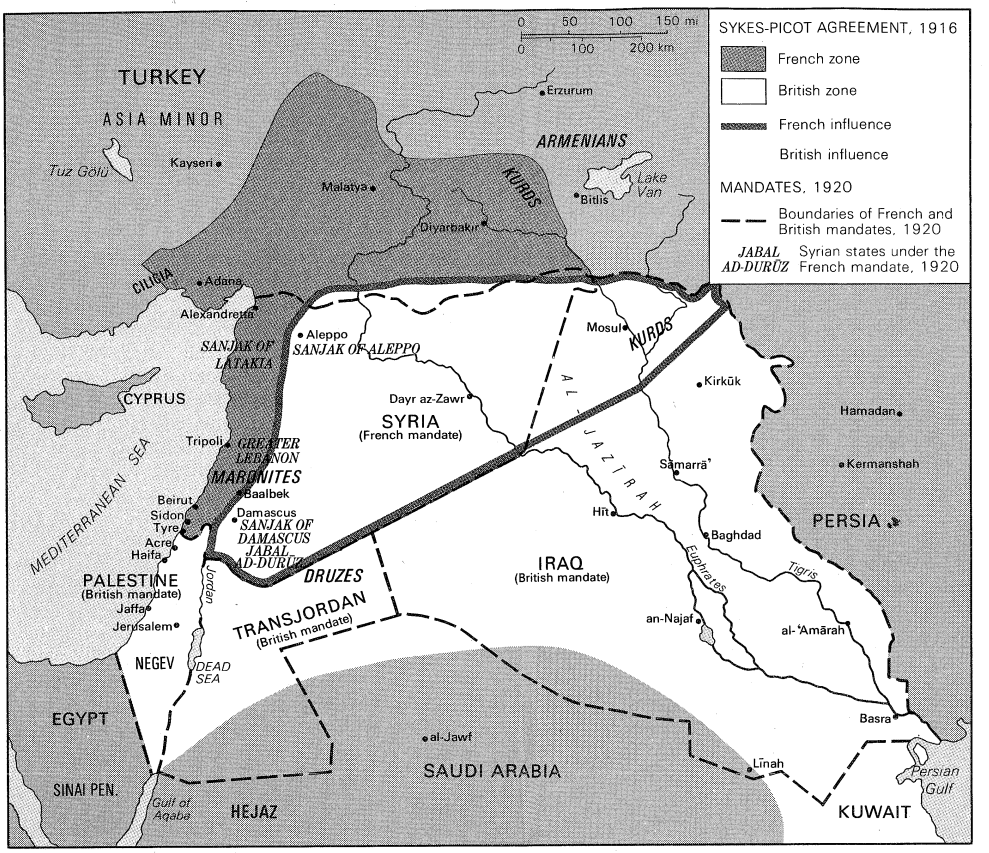
When the war ended, Allenby installed an Arab military administration, under Fayṣal, in Damascus and the interior. The French took over the coast, with Beirut as their centre, and the British took over Palestine. There followed four unsettled years while the fate of Syria was being decided. During the war the British government had made promises, to Ḥusayn and other Arab leaders, that the Arabs would be independent in those countries that they helped to liberate, subject to certain reservations, the precise extent of which has never been clear. Then, in November 1918, Britain and France declared their intention of establishing in Syria and Iraq "national governments drawing their authority from the initiative and free choice of the native populations."

By the Sykes-Picot Agreement of 1916, France was to be free to establish its administration in Lebanon and on the coast and to provide advice and assistance to whatever regime existed in the interior; Britain was to have similar rights in Transjordan and Iraq. Palestine was to be under international control, but France later agreed that it should be in the British sphere, while Britain had already, by the Balfour Declaration of November 1917, pledged conditional support for the establishment of a Jewish national home. It proved impossible to reconcile the various claims and pledges. (A.H.Ho./V.E.I.)

IV. Syria and Palestine since 1920

Events of the war and postwar negotiations had made the *amīr* Fayṣal, the most widely known and respected leader. The government he established in Damascus, although rudimentary, unfamiliar in form, and weak, appeared almost the only rallying point for nationalists. There, in July 1919, a national congress met to demand independence for Syria. Strongly opposed to the French, the nationalists were willing to settle for an interim mandate by the Americans or the British. Just 10 weeks later, Britain turned over effective control of what subsequently became Syria and Lebanon to the French, and Gen. Henri Gouraud was appointed French high commissioner there. By December fighting had broken out between the French and Fayṣal's forces on the frontier between the west and east zones.

On March 8, 1920, the Arab nationalists held a second congress in Damascus, again declaring for complete independence. Three days later, on March 11, 1920, Fayṣal was proclaimed king of Syria as Fayṣal I, but neither France nor Britain would recognize him. In April the Allied supreme council, meeting at the Conference of San Remo, ignored Syrian sentiment and offered the mandate for "Syria and Lebanon" to France. General Gouraud, faced with the problem of imposing the mandate upon the interior, took advantage of a frontier clash to crush the Arab army and occupy Damascus. Fayṣal was dethroned and forced into exile. In Syria, while he had not yet won the esteem of the whole community, the manner of his overthrow left a legacy of bitterness.



Projected and actual divisions of Syria and Palestine by the World War I Allies.
From Z. Zeine, *The Struggle for Arab Independence*

SYRIA AND LEBANON UNDER FRENCH MANDATE, 1920–41

The French organized Syria into four states: Aleppo; Damascus; the ʿAlawite country (the Latakia region); and Jabal ad-Durūz (the predominantly Druze mountain region to the south of Damascus). The state of Greater Lebanon was created from parts of the former Ottoman vilayet of Syria and of the former Ottoman vilayet of Beirut, together with Mt. Lebanon, which had enjoyed autonomy as an Ottoman sanjak since 1861. A high commissioner for Lebanon and Syria resided in Beirut. In 1922 a decision of the high commissioner established a loose federation between the four Syrian states; Lebanon was treated thereafter as a separate entity.

The new administrative organization, although obviously designed to divide and weaken opposition to French rule, had some historical justification. The ʿAlawites (or ʿAlawīs; also called Nuṣayrīs) of the Latakia region and the Druzes of Jabal ad-Durūz, both offshoots from Islām, had for centuries led an existence separate from the majority Sunnites. The provinces of Damascus and Aleppo were distinct under the Ottoman Empire and had a strong sense of local pride. Mt. Lebanon, nucleus of the French-created Greater Lebanon, was geographically and politically separate. And, finally, the province of Alexandretta, with a sizable Turkish population, was in 1923 made a sixth division of the mandate.

Whatever the rationale, however, these divisions were immediately opposed by the Syrian nationalists. In partial concession, the French reunited Aleppo and Damascus in 1925. This only encouraged the nationalists, who then formed the People's Party to demand complete national unity.

Formation of the People's Party

The Druze revolt. Minor outbreaks occurred in various parts of Syria throughout the first five years of the mandate. The first major rising against the French began in 1925 in Jabal ad-Durūz. A large French force sent against the Druzes was defeated, and the revolt spread to other parts of Syria and to the Druze portions of Lebanon. When the insurgents secured a foothold in Damascus, the French bombarded the city. A major military campaign was required to subdue the revolt in 1927.

Establishment of the Lebanese Republic. Developments were quieter in Lebanon, where the Christians, who had formed the overwhelming majority in the old sanjak of Mt. Lebanon, hailed the establishment of the state of Greater Lebanon as an achievement of Lebanese national aspirations. The territories annexed to the old mountain sanjak were preponderantly Sunnite, Shīʿite, or Druze in population; of these three non-Christian groups, only the Sunnites raised serious objections to being included in the Lebanese state.

With the support of a substantial majority of the population, the French were able to introduce numerous administrative reforms. The French governor was assisted by an elected council drawn from the country's major religious sects. The Lebanese leaders who sat in the council demanded a larger measure of self-rule, but in other respects they cooperated with the authorities and accepted the many important reforms the French introduced.

Lebanese administration

In response to the Lebanese demands, the French authorities convened the council toward the end of 1925 to draft a constitution. Upon its adoption in May 1926, the state of Greater Lebanon became the Lebanese Republic. Three days later the legislature elected Charles Dabbas, a prominent Greek Orthodox lawyer and former journalist, as first president of Lebanon.

Constitutional developments in Syria. Achievement of a constitution took longer in Syria. In 1928, a year after the suppression of the Druze revolt, a constituent assembly was summoned. This assembly, in which the People's Party and other nationalists had a large majority, drafted a constitution giving no recognition to the French mandate. The draft was therefore rejected and the assembly prorogued. Further negotiations with the nationalists ended in deadlock, and the French finally introduced a constitution that organized the state of Syria (Damascus and Aleppo) as a republic similar to that of Lebanon. The ʿAlawite state was established as a separate republic, and the Druze state maintained administrative autonomy. The first parliamentary elections in 1932 yielded, under French pressure, a majority of moderates,

but even this hand-picked assembly refused French terms for a treaty until 1936, when a treaty acceptable to both sides was finally concluded. Subsidiary agreements and letters permitted the inclusion in the Syrian republic of the 'Alawite state, the Druze state, and the sanjak of Alexandretta, with special status. Rioting among the Turkish population of the sanjak of Alexandretta over proposed inclusion in Syria caused the French to permit the Turks to send troops to help supervise elections. Their presence secured a bare victory for the Turks in the elections. Two years later, on September 2, 1938, the sanjak of Alexandretta was proclaimed the autonomous Republic of Hatay. It was subsequently incorporated into Turkey, and its loss became a new Syrian grievance against France. The Franco-Syrian Treaty of 1936 was never ratified by France. Upon the outbreak of World War II in 1939, the French took measures to freeze the situation both in Syria and in Lebanon.

Problems and crises in the Lebanese Republic. While the politics of Syria, from the suppression of the Druze revolt until World War II, were dominated by the conflict between the nationalists and the French, the politics of Lebanon following the proclamation of the constitution were more concerned with personal and factional rivalry focussing on Émile Eddé and Bishara al-Khuri, two Maronite leaders. The internal struggle led the French to dissolve the Chamber of Deputies and to suspend the constitution in May 1932.

This provided the French with the opportunity to complete the work of administrative, fiscal, and judicial organization begun in the early years of the mandate. It also provoked a strong Lebanese opposition under the leadership of Bishara al-Khuri. When the French announced that a new president would be elected in 1936, the two men were again the rival candidates. Eddé won with a majority of one vote. Later in the year, following the conclusion of the Franco-Syrian treaty, Lebanon concluded a similar treaty with the French, envisaging complete independence for Lebanon before the end of 1939. In correspondence subsidiary to the treaty, Eddé agreed to observe certain principles, foremost among which was the equitable representation of the various religious sects in government and administration. Accordingly, Eddé asked Khayr ad-Dīn al-Ahdab, a Sunnite Muslim notable and journalist, to form a government in January 1937, thereby establishing the precedent for Sunnite premierships under Maronite presidents.

PALESTINE AND TRANSJORDAN UNDER BRITISH MANDATE, 1920–48

When the Conference of San Remo established the French mandate for Syria and Lebanon in 1920, it gave the British a mandate in Palestine, including Transjordan. When the French overthrew Faysal and his Arab regime in Damascus, his brother Abdullah recruited a private army in the Hejaz and announced his intention of marching on Syria to expel the French. On his way north, in January 1921, he entered Transjordan and established political dominance over its tribes and villages, with his seat in the city of Amman. The British recognized him as *amīr* (prince or governor), and in return he affirmed the British mandate, subsequently split from the rest of the Palestine mandate in 1923, and renounced his declared intention of invading Syria.

The state of Transjordan

The Balfour Declaration. In "Palestine," the definition of which was a part of the controversy, the British were committed to the establishment of a Jewish "national home." The strangeness of the term provides a hint of the complexity of the background. Zionism, like other movements of national self-expression, was a product of the political stirrings of the 19th century; like many others, it arose from bitter experiences in eastern Europe. A Zionist movement of considerable influence was created in England, America, Germany, and France, and before 1914 settlers had been sent to establish communities in Ottoman territory and attempts had been made to win approval for special status for a Jewish community there. Attempts were also made to enlist the support of the German government, which then enjoyed close ties with the Sublime Porte, to promote the scheme. When the first Russian Revolution occurred in 1917, the British believed that much of its leadership was Jewish and sympathetic to Zionism; to keep Russia on the British side and in the war seemed vital. In America (upon which war-impoverished Britain was increasingly dependent for financial support), the British government believed, supporters of Zionism were powerful and influential. For these reasons, the British Cabinet devoted much time and talent during 1917 to planning how to win Jewish sympathy. The result was a declaration, published in Europe in November 1917 as a statement from the British foreign secretary, Arthur James (later Lord) Balfour, to Lord Rothschild, head of the English branch of the banking family. This declaration, subsequently incorporated in the mandate document, was one of the most remarkable exercises in English of the 20th century; just 67 words long, it formed the starting point for political controversy for the next 30 years. It contained two juxtaposed, balanced, but—as history was to show—mutually exclusive parts: "His Majesty's Government views with favour the establishment in Palestine of a national home for the Jewish people, . . . it being clearly understood that nothing shall be done which may prejudice the civil and religious rights of existing non-Jewish communities."

Balfour Declaration

While the declaration was actively publicized in America, in Russia, and on the battlefronts, it was withheld by censorship in the Middle East. When, after the Allied victory, it was published there, and the scope of Zionist ambition was made clear, storms of protest followed. Public opposition forced Faysal, who had flirted with the Zionists in his more urgent fear of the French, to denounce the Zionist movement. Even when Faysal was overthrown and the mandate system was established, the Palestine Arab community felt so assured of general Arab support and was so encouraged by the private sympathies of British officials as to see no need to make concessions to Zionism beyond accepting the Jews already established in Palestine as a minority with guaranteed rights.

In September 1921 the British promulgated a constitution that provided for the establishment of a Palestinian state based upon the expectation of Arab-Zionist cooperation; since the constitution made concessions to Zionism that the Arabs were unwilling to accept, however, it was not applied.

Since no constitutional Palestinian state could be established, the British continued direct rule. They also continued, however, to seek ways to create auxiliaries or partners in government, and in 1923 they offered each of the two major communities an "agency." The Jewish community accepted and formed the Jewish Agency, but the Arabs refused. As so often in the following years, the Arabs' political action amounted to abstention, while the Jewish community profited by creating a shadow government that enabled them to concentrate their efforts on achievement of their major objectives.

Zionists and Arabs

Until 1925, Jewish immigration steadily increased, but from 1926 it fell, and in 1927 there was a net emigration. The Palestinian Arabs relaxed in the belief that the Zionist movement was melting away to nothing. The economic shock of 1929 suddenly reversed the trend and shattered the calm. A demonstration in Jerusalem at the Wailing (Western) Wall—which was also the outer wall of the Muslim Ḥaram (al-Ḥaram ash-Sharīf) area—was followed by the first large-scale attacks upon Jews by Arabs. In the course of the troubles, the *muftī* of Jerusalem, Amīn al-Ḥusaynī, emerged as the leader and champion of the Palestinian Arab cause.

The troubles of 1929 led the British to establish a special commission under Sir Walter Shaw, which reported in March 1930 that the conflict was mainly the result of the disappointment of Palestinian Arab hopes for independence and of the fact that Jewish expansion was creating a "landless and discontented" Arab class in the country. The Shaw Commission urged that restrictions be imposed on the management of the Jewish national home. In May 1930 the British sent a commission under

a senior civil servant, Sir John Hope Simpson, to study the whole scope of the mandate and recommend a solution to the conflict. The commission, noting Arab fears of losing the country, urged an immediate halt to immigration. The British government accepted the advice and issued the so-called Passfield White Paper to implement it. Zionist reaction was immediate, angry, and effective. Ignoring the mandate government, leaders of the movement concentrated their efforts on London and forced cancellation of the White Paper.

The Palestine revolt and the Peel Commission and later commissions. In February 1931 the British prime minister had stated, in a letter to *The Times* of London, that continued Jewish immigration "can be fulfilled without prejudices to the rights and positions of other sections of the population of Palestine." It was not this statement, however, but the election of Adolf Hitler in Germany that was to play the vital role in the years ahead. From 1933 the growing Nazi persecution increased the flow of Jewish refugees into Palestine. The Arabs there, not involved and mainly ignorant of German politics and feeling justified by the Shaw, Hope Simpson, and Passfield episodes, lapsed into a moody and sullen abstention, noncooperation, and boycott of British goods. Meanwhile, the Jewish community, chafing under existing restrictions and increasingly fearful of the rise of Nazism, about which they were better informed, protested and rioted. The British found themselves caught in an impasse.

Nazi persecution

In April 1936 an Arab High Committee was formed to unite the Palestinian Arabs in opposition to the Jews; its formation was followed by a renewal of Arab attacks on Jews, which soon developed into a civil war that was to last three years. In November 1936 a new commission, under Lord Peel, arrived to study the situation. The Arab leaders boycotted the commission until just before its departure. The report, published in July 1937, concluded that cooperation between Arabs and Jews in a Palestinian state was impossible; to the dismay of the Arabs, it recommended the partition of Palestine. Partition, it made clear, was no easy solution, since the establishment of a Jewish state would involve radical movements of population to secure the necessary Jewish majority, even in the parts of Palestine where the Jewish population was largest.

In September 1937, nonofficial representatives from the various Arab countries met at Blūdān in Syria and announced the complete rejection of the Peel proposals. In Palestine, the publication of the Peel report was followed by intensified Arab terrorism and violence. The British thereupon outlawed the Arab High Committee and arrested and exiled its leading members. The *muftī* and a few others escaped arrest and fled to Syria, which became the headquarters of a continuing Palestinian Arab insurrection.

In despair, the British government appointed a new commission under Sir John Woodhead to reconsider the partition plan suggested by the Peel Commission. The Woodhead report, published in November 1938, proposed a reduction of the Jewish share of Palestinian territory to about 400 square miles (1,000 square kilometres) around Tel Aviv—the only area where the Jews constituted a majority. The Woodhead scheme was completely rejected by both Zionists and Arabs. To seek an exit from the impasse, the British convened a conference in London, where the British conferred separately with Arab and Jewish delegations in February 1939.

When no common ground was found, the British government decided to impose its own terms, which were set forth in a White Paper published in May 1939. It stated that there would be no partition and that it was not British policy that the country should become either a Jewish or an Arab state. It envisaged the establishment within 10 years of an independent "Palestine State." In the intervening period, Jews and Arabs would be invited to take an increasing share in the administration. Jewish immigration into Palestine would be limited to a total of 75,000 persons during the next five years, after which no further immigration would be allowed without Arab consent. Land purchases by Jews from Arabs would be prohibited in some areas and restricted in others, in accordance with regulations to be published by the high commissioner.

The British White Paper of 1939

As a proposal for the final settlement of the Palestine question, the White Paper was opposed by both the Zionists and the Arabs. As a means for freezing the situation for the duration of the war, however, it succeeded. Between 1939 and 1945, Palestine was relatively quiet; both the Arab and Jewish communities, for the most part, cooperated with the British, though extremists in both communities carried on their anti-British wars.

THE ACHIEVEMENT OF SYRIAN AND LEBANESE INDEPENDENCE

In Syria and Lebanon, stability was precluded by the fate of France in the war. After that country's fall to the Germans in June 1940, a regime collaborating with Germany was established by Marshal Philippe Pétain at Vichy, while in London Gen. Charles de Gaulle formed a government in exile to cooperate with the Allies. Vichy control of Syria and Lebanon proved dangerous to the British positions in Iraq, from which vital oil was flowing, and in June 1941 British and Free French forces, proclaiming the two mandates' sovereignty and independence, invaded them. By mid-July the Allied occupation of Syria and Lebanon was complete. Gen. Georges Catroux, the representative of the Free French government in Cairo, took over control in Syria and Lebanon as delegate general rather than as high commissioner, to signify the proclaimed new order, and in February 1942 Britain appointed Maj. Gen. Sir Edward Spears as minister to Syria and Lebanon.

World War II and independence

Syria, 1941–43. In September 1941 Catroux formally proclaimed the independence of Syria as a sovereign and politically and territorially indivisible unit including the ʿAlawite and Druze territories; its independence was something less than complete, however, since it was committed to a policy strictly in alignment with that of France. In February 1942 the ʿAlawite and Druze territories were transferred to Syrian authority. Not until March 1943 did Catroux yield to British and Syrian nationalist pressure and decree the full restoration of the Syrian constitution. Elections for a new Chamber of Deputies, held in July, returned the National Bloc to power with a substantial majority; the chamber elected Shukri al-Kuwatli president. The new government demanded complete independence at once, but the Free French authorities declined to transfer powers and services to the Syrian government unless it agreed to enter into special treaty relations with France.

Lebanon, 1941–43. In September 1939, following the outbreak of war in Europe, the French had dissolved the Lebanese Chamber of Deputies, dismissed the Cabinet, suspended the constitution for a second time, and confirmed Émile Eddé as president and head of state by appointment. In April 1941 Eddé, who had quarrelled with the Vichy French authorities, was forced to resign; a Maronite judge, Alfred Naccache, replaced him as chief of state and was still in power when General Catroux formally proclaimed the independence of Lebanon. Subsequently Naccache was appointed president. In March 1943 the Lebanese constitution was restored along with that of Syria; Naccache was removed from his position and replaced by a Protestant notable, Ayyub Thabit, pending general elections.

Meanwhile, political developments had been taking place in Lebanon at a more fundamental level. After the proclamation of Lebanese independence in November 1941, and particularly after the arrival of Sir Edward Spears in Beirut as British minister to Lebanon and Syria, the nationalist leader Bishara al-Khuri and his Constitutional Bloc, backed by the British, began to press for full independence. The Christians of the Constitutional Bloc were joined by prominent Sunnite Muslim leaders of Arab nationalist persuasion, and together they reached a gentleman's agreement on the principle of Christian–Muslim cooperation in the country after independence—an agreement later known as the National Pact. By the unwritten terms of this pact, the Muslim Lebanese were

Christian–Muslim cooperation in Lebanon

to accept the independence, sovereignty, and territorial integrity of Lebanon and cease to demand union with Syria or inclusion in a pan-Arab state; the Christian Lebanese were to admit that Lebanon has an Arab character and agree to cooperation with the Arab states to the greatest possible extent.

By the summer of 1943, as the time for elections approached, the question of how to distribute the parliamentary seats among the various sects became pressing. When President Thabit, whose sympathies lay with the National Bloc, announced that the number of seats would be fixed at 54, allowing 32 for the Christians and 22 for the Muslims and Druzes, strong Muslim protests forced his removal from office. He was replaced as chief of state by a Greek Orthodox notable, Petro Trad, who raised the number of parliamentary seats to 55, giving 30 seats to the Christians and 25 to the Muslims and Druzes. This 6:5 ratio, it was understood, would be maintained in subsequent Lebanese parliaments.

In the ensuing elections the Constitutional Bloc and its allies won an overwhelming victory. The new chamber elected Bishara al-Khuri president, and Riad as-Sulh became premier. On November 8 the Sulh Cabinet presented the chamber with a special bill proposing an amendment of the constitution to remove all references to the French mandate. The bill was unanimously passed, in the absence of the pro-French National Bloc leader Eddé. On the morning of November 11, however, French marines and Senegalese troops arrested President Khuri and his ministers, and the French delegate general, Jean Helleu, announced the suspension of the constitution, the dissolution of the chamber, and the appointment of Émile Eddé as chief of state.

A general strike, angry demonstrations and riots, and the establishment of a temporary Lebanese government in a mountain hideout near Beirut, all with open encouragement from the British, forced the French to reverse themselves, and President Khuri and his party returned to Beirut in triumph. The Free French government then terminated the mandate in both Lebanon and Syria.

While the war continued and French troops remained stationed in the two countries, the French continued to press to secure French interests by special treaty relations. In May 1945, clashes between French troops and rioters in Syria, culminating in a third bombardment of Damascus by the French, led to further British intervention. In the course of the following year, the French forces were compelled to evacuate Syria and Lebanon, which finally became completely independent.

THE PARTITION OF PALESTINE AND THE EMERGENCE OF ISRAEL

Between 1922 and 1939 the Jewish population in Palestine had risen from 83,790 to 445,457 (30 percent of the total inhabitants). Tel Aviv had become a Jewish city of 150,000. Until 1939 the plans to establish a Jewish state in Palestine had been compromised by the fact that there were not enough Jews there to secure a majority in a portion of Palestinian territory of the required size. As eastern Europe fell under German domination, however, Nazi persecution drove many more Jews to seek refuge in Palestine by illegal immigration; and when the systematic slaughter of the Jews of Europe began in 1942, the flow of immigrants became a flood.

Jewish militance

As the Jewish population swelled with this influx of fugitives, the nature of Zionist activity in the country began to change and to tend more toward violence. Following the first Arab attacks on the Jews in 1921, a secret Jewish army called Haganah (Defense) had been formed. Until 1936 Haganah restricted itself to purely defensive action, but during the years of the Arab revolt it became more aggressive; it also received some legal recognition when the British administration formed a Jewish Settlement Police drawn exclusively from Haganah and placed nominally under British command. In 1937 a more clandestine Jewish militia representing the extreme revisionist party within the World Zionist Organization had been formed—the Irgun Zvai Leumi (National Military Organization). During the early years of the war, Irgun followed the lead of the Jewish Agency and cooperated with Haganah; it soon resumed its old extremist course, however, as Jewish fugitives freshly arrived from Poland joined its ranks and took over its control. The Irgun leaders were convinced that Britain had betrayed the Zionist cause—an opinion that was shared by another fanatical organization, the Stern Group, or Gang, whose leader, Abraham Stern, was killed in a police raid in 1942. In the last years of the war, the Irgun and the Stern Group began resorting to terror to remind the British of their commitment to Zionism.

Meanwhile, the Jewish Agency, under its veteran leader Chaim Weizmann and younger leaders such as David Ben-Gurion, tried to maintain British goodwill by offers of help to the British war effort. They proposed formation of a Jewish Legion to defend Palestine, and when Britain showed a disinclination to accept this proposal Palestinian Jews joined the British forces as volunteers.

Pressures for a Jewish state. A new Palestine policy was decided upon at a Zionist conference held in May 1942 at the Biltmore Hotel in New York City; it called for unrestricted Jewish immigration into Palestine and for the ultimate establishment of the country as a Jewish commonwealth. By the end of the war, Zionist political activity in the United States had succeeded in winning the U.S. government's support for Zionism, and Britain, unable to resolve the predicament on its own, was pleased to admit American involvement.

Biltmore Resolution

At the end of the war in Europe, the Jewish Agency addressed a memorandum to Britain demanding the full and immediate implementation of the Biltmore Resolution. Another memorandum followed in June 1945 demanding that immigration visas be issued for 100,000 European Jewish refugees awaiting admission into Palestine. By the time Japan surrendered in September, Haganah had gone into alliance with the Irgun and the Stern Group to present a united front in Palestine. The Jewish Agency stood ready to assume the provisional government of the Jewish state. In the absence of a unified Arab leadership in Palestine, Arab leaders from the neighbouring countries of Egypt, Syria, Lebanon, Transjordan, Iraq, Saudi Arabia, and Yemen, which in March 1945 had formed the League of Arab States, proclaimed their intent to take up the defense of the Arab cause in Palestine.

In the United States, Pres. Harry S. Truman took up the Zionist cause and urged that the European Jewish refugees be immediately admitted into Palestine. Beset by U.S. pressure, fighting a costly and unpopular war against Zionist guerrillas in Palestine, and seeing no practical solution, Britain participated in yet another commission, the Anglo-American Committee of Inquiry, which published its conclusions in April 1946. In essence, it recommended the immediate admission to Palestine of 100,000 Jewish refugees from Europe, the withdrawal of all restrictions on Jewish purchase of land, and the eventual incorporation of both communities in a binational state under United Nations trusteeship. The British government refused the central and immediate demand, admission of 100,000 refugees, and suddenly found itself involved in a war. Forced to maintain a large and costly military establishment when its electorate demanded demobilization and easement of the tax burden, with no end of the bloodshed in sight and with all its practical political and diplomatic ploys used and options closed, the British referred the question to the United Nations.

The General Assembly voted on May 15, 1947, to create a Special Committee on Palestine (UNSCOP) to submit "such proposals as it may consider appropriate for the solution of the problem of Palestine." When it arrived in Jerusalem, UNSCOP was boycotted by the Arabs but actively aided by the Zionists. Few issues had been more studied than Palestine, and UNSCOP found nothing new but urgency. The only solution, it suggested, was partition, but it urged that the consequences of partition be mitigated by the maintenance of economic union. On November 29 the UN General Assembly approved, with slight frontier modifications, the UNSCOP recommendations.

UNSCOP

The Palestine war. The UN decision was a major Zionist victory. Not only did it affirm the Zionist right—the fundamental point at issue and bitterly opposed by the Arabs—to establish a Jewish state in Palestine, it gave the state a territory that, although smaller than that proposed by the Jewish Agency, was far out of proportion to the relative numbers of Jews to Arabs in the country. It comprised more than half the territory of Palestine, including the greater part of the valuable coastal area. Shocked and angry, the Arab leaders refused to recognize the validity of the UN decision and declared their determination to oppose it by force.

By January 1948, volunteers were arriving from the Arab countries to help the Palestinian Arabs; but these irregulars, unorganized and poorly led and equipped, were soon overwhelmed by the superior Zionist forces. By May 13 the latter had secured full control of the Jewish share of Palestine and captured important positions in the areas allotted to the Arabs. The Irgun stormed and captured the village of Deir Yāsīn and massacred much of the population, terrorizing the Arab villagers, who began a mass exodus from Palestine.

Deir Yāsīn

On May 14 the State of Israel was proclaimed and immediately recognized by the Soviet Union and the United States. On the following day, as the British announced the end of their mandate in Palestine, troops of the new, poorly trained, shoddily equipped, and uncoordinated armies of Egypt, Transjordan, Syria, Lebanon, and Iraq entered the country.

They were no match for Israel, and the separate armistice agreements that were finally signed in 1949 by Egypt, Lebanon, Transjordan, and Syria left Israel in possession of all the areas it had won by conquest: the whole of Galilee, the whole of the Palestinian coast minus a reduced Gaza Strip (now occupied by Egypt), all of the Negev, and a strip of territory connecting the coastal region to Jerusalem, including the northwestern section of Jerusalem. The remaining parts of Jerusalem (including the Old City), along with what remained of the Arab share of Palestine, were taken over by Transjordan, which then became the Hashemite Kingdom of Jordan. No entity remained that was officially called Palestine. The departure of nearly 1,000,000 Palestinian Arabs had meanwhile left Israel with the substantial Jewish majority it urgently required.

SYRIA AFTER INDEPENDENCE

The humiliating failure of the Arab intervention in Palestine brought serious discredit to the governments of the Arab countries involved, but nowhere more than in Syria. There it provided the excuse for the first of a series of military coups.

Fundamental to the Syrian problem was the ethnically, religiously, and socially heterogeneous nature of the emerging republic. The new state united the ʿAlawite and Druze territories, which had formerly enjoyed separate status, with the predominantly Sunnite regions of Damascus, Homs, Ḥamāh, and Aleppo. While the ʿAlawites and Druzes formed compact communities in their respective regions, they were not the only religious minorities in the country. In northern Syria there were communities of Ismāʿīlī Muslims. Throughout the country, and particularly in the cities, there were large communities of Christians. Kurds and other less important ethnic minorities, notably the Armenians and the Syriac-speaking Christians, were to be found in various parts.

Syria's heterogeneous population

In addition to this religious and ethnic heterogeneity, there was an equally important social heterogeneity; the population of Syria was composed of townsmen, peasants, and nomads, three groups with little in common. Economic differences added further complexity; in the cities the ostentatious wealth of the notables contrasted sharply with the poverty of the masses. Those same notables were also the owners of large agricultural estates on which the peasants were practically serfs. It was the Sunnite landowning notables of Damascus, Homs, Ḥamāh, and Aleppo who had organized themselves as the People's Party, and later as the National Bloc, and led the resistance to the French. When Syria achieved independence, they took over power and endeavoured to forge a unitary state. They proved unequal to the task.

By 1949 the small but rising middle class, among which new social ideas were developing, and minorities, mainly the ʿAlawites and Druzes, who resented the growing threat to their particularism, were increasingly opposed to the government. The rulers, having tasted power after so long a struggle for independence, self-righteously refused those concessions that might have saved them. Moreover, they appeared to be more devoted to achieving pan-Arab goals than to solving the problems closer to home. In the years immediately following World War II, Iraq and Saudi Arabia were making rival bids for pan-Arab leadership. The ruling National Bloc in Syria readily divided into two new parties: a National Party headed by Shukri al-Kuwatli, which represented the business interests of the Damascus notables and supported Saudi Arabia; and a resuscitated People's Party, which represented the interests of the Aleppo notables and supported Iraq. The Socialist and secular Arab nationalist Baʿth party was recruiting followers among students and army officers, winning support particularly among the ʿAlawite, Druze, and other minorities that were strongly represented in the army.

The colonels. The end of the short-lived civilian order in Syria came in March 1949 when Col. Husni az-Zaim overthrew the Kuwatli government in a bloodless coup. Zaim was himself overthrown in August by Col. Sami al-Hinnawi. A third coup, led by Col. Fawzi Salu in conjunction with Col. Adib ash-Shishakli, followed in December; by November 1951 Shishakli had removed his associate by a fourth coup.

Military dictatorships in Syria

The first four military dictators of Syria were officers of no particular ideological commitment, and the regimes they led may be described as conservative. All four ruled in association with veteran politicians of the former civilian regime. Among the politically minded army officers at the time, many were Baʿth Socialists, while others were followers of the Arab Socialist Union of Akram al-Hourani. Colonel Shishakli was a first cousin to Hourani and came to power with his assistance. Opposing the Baʿth and pro-Hourani Socialist officers were officers of a radically different political persuasion, who followed the Syrian Social Nationalist Party (the Parti Populaire Syrien, or PPS), an authoritarian party of the Fascist type devoted to the establishment of a pan-Syrian national state that would include Syria, Lebanon, Palestine, Transjordan, Iraq, and Cyprus. Because its concept of the Syrian nation included Iraq, it received strong backing from the Iraqi government. After coming to power, Shishakli parted company with Hourani and turned to the PPS for support.

Shishakli was overthrown in February 1954 by a military coup led by Col. Faisal al-Atasi. The PPS forthwith lost its influence in Syrian politics, and an attempt to regain it by violence in the following year led to the suppression of the party and the purging of PPS elements in the army. From that time the Baʿthists in the army had no serious rival, although the party was not yet in a position to take power alone. In the four years that followed, it had to associate itself in rule with civilian leaders of the old school and also with the Communists.

The union with Egypt, 1958–61. The years that followed the overthrow of Shishakli in Syria saw the rise of Pres. Gamal Abdel Nasser of Egypt to uncontested leadership of the pan-Arab unity movement. As a party committed to pan-Arabism, the Baʿth could not dissociate itself from this movement; and the veteran pan-Arab Syrian politicians found themselves in the same position. The Arab world was meanwhile being drawn into the Cold War. The continued interference of pro-Western Iraq in Syrian affairs forced the coalition regime in Syria to turn more and more to Egypt for support and also to establish the first friendly contacts with the Communist countries. By 1958 the fear of Iraq and its Western involvements, on the one hand, and a growing reluctance to become committed to the Communist countries, on the other, convinced the Syrian coalition that the only way out was union with Egypt. In February 1958 Syria gave up its

sovereignty to become, for the next three and one-half years, the "Northern Province" of the United Arab Republic (U.A.R.) of which Nasser was president.

The union of Syria with Egypt, though widely acclaimed, proved a bitter disappointment. Far from accepting the Syrians as full partners, the Egyptians tended to treat them as subordinates. In September 1961 a coup led by rightist Syrian army officers, in league with a coalition of civilian leaders, re-established Syria as a sovereign and independent state.

The "secessionist" regime, 1961–63. The coup of 1961 paved the way for a return of the old class of notables to power. Parliamentary elections were held, and Nazim al-Qudsi of the People's Party was chosen president. The "secessionist" regime, though civilian at the surface, was still under army control, and in the army the Ba'th was still powerful. The regime had made hardly any concessions to the Socialism of the Ba'th and the pro-Nasser pan-Arabists. While the latter approved of the Socialist measures introduced under the union with Egypt (such as the land reforms and the nationalization of banks and large business enterprises), the secessionist regime set out quickly to undo or at least amend them, thus playing into the hands of the Ba'th. In March 1963 the Ba'th and their supporters in the army seized power.

Ba'thist Syria after 1963. The new government was a coalition that included pro-Nasser elements who demanded reunion with Egypt. A month before the Ba'th coup in Syria, however, the Iraqi branch of the party had seized power in Baghdad. A Ba'thist union between Syria and Iraq seemed imminent, but it was opposed by Egypt and by the pro-Nasser Arab unionists in Damascus and Baghdad, and the Ba'thists were unable to establish a union that would exclude Egypt. The Ba'th leaders of Iraq and Syria flew to Cairo for unity talks with President Nasser, but Nasser would agree to a union only on his own terms, and the talks failed. The Ba'th regime in Iraq did not outlive the year. In Syria the pro-Nasser Arab unionists were thrown out of power by an eighth coup in October 1964, and a Ba'th military dictatorship was established.

The Ba'thists in Syria were soon faced with a serious problem. Although their party was led by Syrians, it was pan-Arab in ideology and organization and had branches in Iraq, Lebanon, and Jordan. Non-Syrian Ba'thists were members of the central committee of the party, which had its headquarters in Damascus; the continued subordination of the Syrian branch of the party to the central committee gave non-Syrian Ba'thists a considerable say in Syrian affairs. Thus it was necessary for the Syrian branch of the party to become independent of the pan-Arab central committee. In May 1965 the party split. A ninth coup in February 1966 set up a new military regime in the name of the Syrian Ba'th. The central organization of the Ba'th was reconstituted in Lebanon.

Ba'th split

For some years the Syrian branch of the Ba'th party, with its 'Alawite leadership, successfully maintained itself in power. Nureddin al-Atassi, a Sunnite notable from Homs, served as chief of state until November 1970, while Col. Salah al-Jadid wielded ultimate power. Salah al-Jadid was then deposed in favour of Hafez al-Assad, who on March 14, 1971, was sworn in as president. (For the Syrian role in the Lebanese civil war, see below.)

LEBANON AFTER INDEPENDENCE

For many years Lebanon successfully maintained its parliamentary democracy, despite serious trials. The main problem for Lebanon was to implement the National Pact between the Christians and Muslims. In the early years of independence, so long as no urgent call for pan-Arab unity came from outside, the National Pact faced no serious strains.

The Khuri regime, 1943–52. Under Pres. Bishara al-Khuri, there was close cooperation between the Maronite president and the Sunnite leader Riad as-Sulh, who was premier most of the time. So successful was the partnership that an attempt was made to perpetuate it by seeking a temporary amendment of the constitution to permit the president, in 1949, a second six-year term. The parliamentary elections of 1947 were blatantly rigged to produce a Parliament favourable to the amendment. This, together with the open favouritism of the president toward his family and friends, and the gross corruption he allegedly condoned, made Khuri increasingly unpopular after his re-election in 1949.

The military coups that overthrew the Kuwatli regime in Syria in March 1949 and the monarchical regime in Egypt in July 1952 encouraged the opponents of Khuri in Lebanon. In July 1949 the Syrian Social Nationalist Party (PPS) tried to overthrow the regime by force. The coup failed and its leaders were seized and shot. The PPS took its revenge by securing the assassination of Khuri's premier in 1951. The mounting opposition to the Khuri regime culminated in September 1952 in a general strike that forced his resignation. Camille Chamoun was elected by the Parliament to succeed him.

The Chamoun regime and the 1958 crisis. The presidency of Camille Chamoun coincided with the rise of Nasser in Egypt. During the Suez war (October–December 1956), Chamoun earned Nasser's enmity by refusing to break off diplomatic relations with Britain and France, which had joined Israel in attacking Egypt. Chamoun, known for his pro-Western leanings, was accused of seeking to align Lebanon with Iraq, Turkey, Iran, and Pakistan in the Western-sponsored Baghdad Pact.

Lebanon's difficult position

Matters came to a head following the parliamentary elections of 1957, which allegedly were manipulated to produce a Parliament favourable to the re-election of Chamoun. In the summer sporadic troubles occurred in the Druze area of ash-Shūf. When Syria entered the union with Egypt in February 1958, the Muslim opposition to Chamoun in Lebanon hailed the union as a triumph for pan-Arabism, and there were widespread demands that Lebanon be associated in the union. In May a general strike was proclaimed, and the Muslims of Tripoli rose in armed insurrection. The insurrection spread, and the army was asked to take action against the insurgents to prevent the overthrow of the Chamoun regime by force. The commanding general, Fuad Chehab, however, refused to attack the insurgents for fear that the army, which was composed of Christians and Muslims, would split apart. United Arab Republic units from Syria took over peripheral parts of the country to help the insurgents. The Chamoun government took the issue to the United Nations, accusing the U.A.R. of intervention, and UN observers were sent to Lebanon. When in July the pro-Western regime in Iraq was toppled in a coup by apparently pro-U.A.R. forces, President Chamoun immediately requested U.S. military intervention, and on the following day U.S. marines began to land outside Beirut. The temporary presence of U.S. troops had little immediate effect on the internal situation, but the insurrection slowly faded out. To succeed Chamoun as his term ended, Parliament turned to the commander of the army, General Chehab; Rashid Karami became the new premier.

The Chehab and Hélou regimes, 1958–70. The crisis had been resolved by judicious compromise, and the Chehab regime was successful in maintaining the compromise and promoting what was renamed the National Unity (rather than National Pact) of the Lebanese People. By his refusal as army commander to take offensive action against the insurgents in 1958, Chehab had earned the confidence of the Muslims as no Christian leader had before. Once in power, he proceeded to allay the long-standing Muslim grievances by associating the Muslims more closely in the government and administration and by attending to neglected peripheral areas of Lebanon where Muslims predominated. Internal stability was further promoted by the re-establishment and maintenance of good relations with the U.A.R. (which, even after the Syrian secession in 1961, remained highly popular with the Muslim Lebanese). The economic boom that had begun under the Chamoun regime as the result of the flight of capital from the unstable Arab world into Lebanon continued under the Chehab regime.

National Unity of the Lebanese People

Charles Hélou, a former journalist and member of Khuri's Constitutional Bloc, was elected to succeed Che-

hab in 1964. His government was essentially a weaker version of the Chehab regime, but it managed to keep the internal peace and to thwart attempts to draw Lebanon into military alignment with Syria and Egypt.

ISRAEL AND THE ARAB STATES

Jordan and the Palestine question after 1949. When King Abdullah concluded his armistice with Israel in 1949, he incorporated into his kingdom the area of the old Palestine mandate held by his troops. In the eyes of many Palestinians, Abdullah appeared as the ultimate enemy, and on July 20, 1951, he was assassinated in Jerusalem by a Palestinian terrorist. The brief reign of his mentally unbalanced son Talal was followed by that of his grandson Hussein, from his coming of age in 1953.

The reign of King Hussein

Hardly had King Hussein taken the oath of loyalty to the constitution when the Ba'thists and their associates began to foment trouble among his discontented Palestinian refugee subjects. They also encouraged military plots against the King. After the establishment of the U.A.R. in 1958, fear of the consequences for his country and regime led the King to conclude a federal union with Iraq. The overthrow of the Iraqi monarchy in the same year brought a quick end to this federation; subsequently, British paratroops had to be flown to Jordan to prevent a Nasserist overthrow of the monarchy there.

Syria's secession from the U.A.R. in 1961 brought temporary relief to the regime in Jordan. During the years that followed, United States aid made it possible for Hussein to promote considerable economic development in his country, which, apart from the Palestinian West Bank, was composed mainly of desert. There was also a boom in tourism, centred on the Old City of Jerusalem. Hussein's Palestinian subjects, though a source of political trouble, included some of the best educated and talented elements in the Arab world.

Formation of PLO and al-Fatah

Until 1964 the Palestinians were the object of actions by the Arab states; largely leaderless and unorganized, they were recipients of United Nations aid and a source of cheap labour in the "host" countries. Their cause was argued for them by others, and such political activity as they evinced was in relationship with one or other of the Arab states or leaders. In 1964, however, the Arab summit meeting that proposed the United Arab Command also led to the formation of a Palestine Liberation Organization (PLO). About the same time, Palestinians living in various Arab countries formed a secret organization called the Palestine National Liberation Movement, known from a reversal of its Arabic initials as al-Fatah ("the Conquest"). Both the PLO and al-Fatah undertook the training of guerrilla units for raids on Israel. Until 1967 these raids emanated mainly from Jordan. The distribution of the refugee camps focussed attention on Jordan and Lebanon and created tensions in both countries that were to explode in later years.

Israel after 1948. The emergence of Israel as a Jewish state on the former territory of Palestine was the central political issue of the Middle East from the end of World War II. Viewed from any perspective, it was a remarkable achievement. The energy, enthusiasm, and skill of the Zionists were unmatched among the Arabs. Essentially, Israel represented the coming of a modern, European state into an underdeveloped area, and in that fact lies the crux of its achievement and its problems.

Deriving proportionally huge external financial and military support from Western governments (its relations with the Soviet Union, despite immediate recognition in 1948, have been strained), the Israelis also benefitted from a highly trained and motivated citizenry to create a unique nation-state. Ironically, they were assisted in this process by the hostility of the surrounding Arab countries, because the state of siege in which they were forced to live helped to unite them against the common danger and gave them a sense of mission. And, finally, hanging as a tragic backdrop to the state was the memory of the Nazi holocaust, both a justification and a cause of the final push to success of the Zionist movement.

In 1948 Israel emerged as a parliamentary republic, with David Ben-Gurion, the head of the Jewish Agency, as premier of the provisional government. Parliamentary elections held in January 1949 yielded a Knesset (Parliament) in which the Mapai (Labour Party) of Ben-Gurion had the ascendancy. The Knesset elected Chaim Weizmann, president of the World Zionist Organization, as the first president of Israel (he was succeeded, upon his death in 1952, by Itzhak Ben-Zvi, who was succeeded in turn, upon his death in 1963, by Schneor Zalman Shazar). In keeping with Zionist principle, the Knesset proclaimed Jerusalem the capital of Israel (only a section of the new city was actually held by Israel) and passed the Law of Return, which gave every Jew the right to immigrate. Ben-Gurion formed a succession of coalition governments that kept him in power continuously (except in 1953–55, when his Mapai associate Moshe Sharett was premier) until his resignation in 1963. Levi Eshkol (also of the Mapai) succeeded him at the head of another succession of coalition governments. Upon Eshkol's death in 1969, the Israel Labour Party, which had been formed in the previous year to unite the Mapai with the more leftist Ahdut Avodah and the Rffi (a splinter party of the Mapai founded in 1965 by Ben-Gurion in association with Maj. Gen. Moshe Dayan), chose Golda Meir, former foreign minister and secretary general of the party, to succeed Eshkol as premier. Labour remained in power till 1977.

Israel Labour Party

To large numbers of Jews throughout the world the state of Israel symbolized the hopes and aspirations of the Jews as a people. To its neighbours, Israel remained an alien presence forcibly established on formerly Arab soil and consequently unacceptable. At a minimum, they demanded a return to frontiers fixed by the 1947 UN partition and repatriation of the Palestinian Arab refugees who had fled the country after 1948; to neither condition was Israel prepared to yield. Because of Arab hostility toward Israel, the government had to maintain restrictions on the civic and political liberties of its Arab minority—a blemish on Israel's democracy, which was otherwise the most successful in the Middle East.

The frontiers of Israel were periodically exposed to guerrilla raids from one or another of the neighbouring countries. Even more important, Israel's access to the Red Sea, essential for trade with East Africa and the Far East and for receiving needed oil supplies from Iran, was barred by Egypt's control of the Suez Canal and by the Egyptian and Saudi Arabian control of the Strait of Tiran at the mouth of the Gulf of Aqaba. After the Suez war of 1956, Israel secured a guarantee for passage through the Strait of Tiran in return for its retreat from Sinai and the Gaza Strip, occupied in the course of that war. The port of Elat was enlarged to handle the Red Sea trade.

The Arab–Israeli War of 1967 and after. While the direct cause of the Six-Day War of June 1967 was Egypt's closure of the Strait of Tiran, the war was, in a more profound sense, another battle of the wars of 1948–49 and 1956, focussed on the issue of Israel's legitimacy in the Middle East. Inexorably, also, it was coloured and magnified by the global confrontation of the United States and the Soviet Union.

From time to time, Israeli forces took reprisals for attacks on Jewish settlements by Arab guerrilla bands—attacks that became particularly frequent after 1964 and that came mostly from across the Syrian border. Tension had been building up on the border from the early 1960s as Israel asserted its control over the demilitarized zones. In November 1966 Nasser pledged his support for Syria if Israel should invade; subsequently, Syrian bombardment of Israeli villages from across the border was intensified, as were the Israeli reprisals and warnings. In May 1967 Syria announced that Israel was massing troops on its border; Nasser thereupon called for the immediate withdrawal of the UN Emergency Force from the Israel–Egypt border and closed the Strait of Tiran to Israeli shipping and to all other vessels carrying strategic material to the Israeli port of Elat; he also secured a promise of solidarity from Jordan and received promises of support from Saudi Arabia, Algeria, and Iraq. In Israel, Premier Levi Eshkol, under popular pressure, handed over the Ministry of Defense to Maj. Gen. Moshe Dayan.

Withdrawal of the unarmed UN troops after May 16 signified to the Israelis that ultimately they were alone. On May 30, King Hussein signed a mutual defense pact with Nasser, convincing the Israelis that they were surrounded. Two hours before dawn on June 5 the Israeli Air Force caught on the ground and largely destroyed the Arab world's only effective military force, the Egyptian Air Force. In Sinai in the following days, the Israeli Army smashed the Egyptian troops. Caught by its five-day-old pact with Egypt, Jordan also was sucked into the war. Israel offered Jordan a sort of informal separate peace, but Hussein realized that he could not accept and survive. The price he paid for survival was loss of the bulk of his kingdom, the whole of the West Bank of the Jordan. Syria in its turn was the object of a furious frontal assault that cleared the Golan Heights and threatened Damascus.

Jordan–Egypt pact

By the time the UN Security Council managed to effect a cease-fire (June 11), the Arab states had lost huge territories, much of their productive capacity, perhaps 5 percent of their best labour force, millions of dollars worth of industrial equipment, and untold amounts of potential tourist and other revenues; much more importantly, they lost psychologically and politically. That fact was to set the tone for the events of the following years. Central to the history of the Middle East, the defeat of the Arab governments gave a powerful impetus to the Palestine guerrilla movement.

Militancy of the Palestinian refugees. Following the cease-fire, the more militant members of the Palestine Liberation Organization (PLO) took over its leadership and began making guerrilla raids into Israel and Israeli-occupied territory. In riposte, Israeli forces attacked the "host" countries, and Lebanon and Jordan found themselves paying the bill for the policies of the Palestinians. The Palestinians, moreover, increasingly insisted upon their right to act as an independent state, under arms and immune from the laws of Jordan. Clashes and confrontations punctuated the months leading up to the high point of Palestinian power in 1970. In June a Palestinian group apparently tried to assassinate King Hussein and fired on his palace with rockets. Next, they demanded the arrest and dismissal of the King's uncle, who was commander of the armed forces. Surprisingly, King Hussein announced his decision to bow to most of the guerrilla demands, that concession, he said, to be his last. But the guerrilla groups, never really effectively coordinated, much less centralized, continued to push their program, and by August the country was an armed camp or, more accurately, a collection of armed camps, in which the government appeared to be but one among many rival powers. The situation obviously could not continue, and the spark that set off the explosion, known among the Palestinians as "Black September," was the most flamboyant of the Palestinians' acts, the hijacking on September 6–9 of four commercial airliners, three of which were flown to northern Jordan and there held hostage with their passengers and crews. Humiliated, the King resolved to destroy the guerrillas as a contender for power in Jordan. The Jordanian Army, given its head in mid-September, swept through the refugee camps, wiping out all resistance, disarming the guerrillas, and deporting many of the leaders. Driven from Jordan, the PLO focussed its activities upon Lebanon.

Black September in Jordan

Israel, confident that its policy had worked in Jordan, naturally turned similar pressure upon Lebanon. Already, in December 1968, an Israeli Army airborne group had seized Beirut International Airport and blown up most of the Lebanese commercial air fleet; this spectacular raid was followed by others, particularly along the southern frontier. Perhaps the most dramatic of these took place in April 1973, when Israeli commandos spent almost a complete day in Beirut hunting down and killing some of the best known leaders of the PLO. Outraged, the Palestinians accused the Lebanese of complicity.

The war of 1973. When Gamal Abdel Nasser died in September 1970, he was succeeded by Anwar as-Sadat. Few expected from Sadat a continuation of Nasser's strident pan-Arab policies, and the Egyptian offensive of October 6, 1973, thus came as a shock to many who had overlooked the clear signs that war was about to break out. During 1973 the Soviet Union increased arms deliveries to both Egypt and Syria; Syria alone received in the first six months of 1973 almost $200,000,000 worth of arms and aircraft. By the end of June, Sadat of Egypt and Assad of Syria had formed a rough plan of operations. In September King Hussein apparently was informed in general terms and approved the plan. A major air battle off the Syrian coast, in which Israeli planes shot down at least eight Syrian aircraft, appears to have resolved all Syrian doubts on Sadat's conditions for cooperation. Ironically, Sadat himself publicly warned that war was imminent in a speech on September 28.

Fighting started almost simultaneously on the Syrian and Egyptian fronts at about 2 PM local time on October 6—for the Jews, Yom Kippur, the holiest day of the year, and for the Muslims, the anniversary of the crucial battle for the conquest of Mecca by the Prophet Muḥammad, the 10th of Ramaḍān. The war ended on the Egyptian front effectively on October 24 but dragged on on the Syrian front until a cease-fire was achieved in May 1974. Except for action by a Jordanian unit in Syria, Jordan was not directly involved in the fighting, and Lebanon managed to stand aloof. The Palestinians, as before, were ignored in the formal military engagement.

The Syrians were by no means so successful as the Egyptians, either in the fighting or in the cease-fire negotiations. They ended the war with a net loss of territory and with appalling losses of men and material. Bitter because the Egyptians had signed a separate cease-fire, they felt that the Egyptians were being wooed away from the Arab cause by the United States. The Soviet Union, feeling outmanoeuvred and in danger of losing its diplomatic position in the Middle East, encouraged the Syrian "hawks." The cease-fire on the Syrian front was thus more difficult to achieve than that in Sinai. Negotiations were hard, and the U.S. secretary of state, Henry A. Kissinger, spent 28 days travelling back and forth between Israel and Damascus to achieve acceptable results. In the cease-fire, Syria regained 300 square miles of territory lost in the war, together with small but symbolically important pieces of ground lost in the fighting of 1967; Israel retained the Golan Heights and an observation post on Mt. Hermon; and a complex buffer zone under UN supervision was established around the Syrian town of al-Qunayṭirah.

CIVIL WAR IN LEBANON, 1974–76

Events in Lebanon meanwhile were moving toward the outbreak of one of the most destructive civil wars in modern history. The essential issues were separate but strongly and immediately affected one another. First came the fact that the political structure of Lebanon was based on a sort of floating consensus—an agreement among leaders of the various religious and ethnic minorities over their respective roles in the state. Under the strong regime of Pres. Fuad Chehab and his somewhat weaker successor, Charles Hélou, the personal and group ambitions of the various sectors of society were held in check, and the organs of the state, especially the army and the security police, were employed to prevent recourse to violence. The immediate result was a long period of domestic calm, stretching from the civil war of 1958 to 1969.

Background to the war. During that period, however, major changes were taking place in Lebanese society, most significantly urbanization, which brought 40 percent of the Lebanese population to the city of Beirut. But the city, like the country, failed to make its people homogeneous; Beirut became a reflection of Lebanon as a whole. Each quarter took on a religious affiliation, with virtually every village having an outpost established in one suburb or more. And the newcomers suffered from deep and growing social and economic contrasts with their more affluent close neighbours. Those who remained in the rural areas lost even more ground, relatively speaking. By the early 1970s, agriculture, in which nearly half the population was employed, accounted for less than 11 percent of the country's gross domestic product, while the

Lebanese social change

share of urban commerce was rising. An already existing schism was deepened between the groups, which increasingly took on an urban Christian *versus* rural Muslim coloration.

A second factor, the role of Lebanon in the Arab world, was an issue almost as complex as that of Israel. Many Syrians still felt that the French decision to separate Lebanon from Syria in 1920 was invalid and an affront to the very idea of pan-Arabism; many in Lebanon, often for different reasons or only temporarily, agreed or professed to agree. Lebanon's noninvolvement in the wars of 1967 and 1973, its strong and often heavy-handed security policies, and rumours of its secret understandings with Israel all focussed attention on this issue.

Third, the Palestinians thought of Lebanon, after the ruinous Jordanian campaign against them in September 1970, as their last refuge, and by 1973 roughly one person in each 10 in Lebanon was a Palestinian. Yet the Palestinians were constantly made aware of their separate and inferior status; landless and mostly poor, they were exploited as a source of cheap labour. Increasingly their politics became radicalized, and they found common cause with those Lebanese who were poor, rural, and mainly Muslim. As they acquired structure, motivation, and arms during the period after the fall of their base in Jordan, they were gradually sought out as allies by radical groups in Lebanon. Faced, finally, in September 1975, by an Egyptian–Israeli interim agreement on Sinai, the Palestinians concluded that they were being deserted by the Arab states and would shortly be suppressed in Lebanon by a second Black September.

Palestinian radicalization

Beginning of the war. Toward the end of the presidency of Charles Hélou, the Palestinians began to clash with Lebanese security forces. The situation was increasingly complicated by the inability of the PLO to control the more radical of the factions into which the Palestinians were divided. Under an agreement announced in Cairo on November 3, 1969, the Lebanese government gave the Palestinians virtually a free hand in the refugee camps and at forward posts in the south along the Israeli frontier. In return, the PLO promised not to intervene in Lebanese politics; this, for reasons noted above, was impossible for the Palestinians and not desired by the Lebanese left. It was particularly significant that at this period the minister of the interior, charged with police functions, was the leftist leader Kamal Jumblatt. When the Lebanese failed to restrain the Palestinians, the Israelis began to raid the south with increasing severity, and this fact encouraged the Lebanese Christian right, particularly the militant Phalangist Party (with discreet army backing), to attack the Palestinians with its well-organized and well-armed militia.

Franjieh election

It was in this atmosphere that Suleiman Franjieh was elected president, by only one vote in the Parliament, on August 17, 1970. Franjieh was known as a man of personally violent tendencies, and the Lebanese right looked to him to suppress the left and their Palestinian allies. In his first acts, he seemed to justify that hope. Economic events, however, soon got out of control. Lebanon was caught in a severe inflation that enriched those on one side of the social chasm while it exacerbated the distress and bitterness of those on the other side. The government appeared incapable of addressing this problem; in its attempt to make an appearance of reform, it discarded or destroyed the strong instruments of state control that it had inherited from the aftermath of the civil war of 1958. The more radical of the guerrilla groups, angered by failure of the Lebanese to prevent Israeli incursions and encouraged by the vacillation of the regime, determined upon a series of dramatic terrorist attacks. When their plots were foiled, they retaliated by abducting several Lebanese soldiers. Heavy fighting then broke out between the army and the armed refugees in May 1973, and thereafter, despite a series of truces, events moved rapidly toward civil war. By early 1974 the Palestinians were providing the force for the Lebanese National Movement led by Kamal Jumblatt. Into this arena then stepped the relatively deprived Shī'ite Muslims, especially those of the rural area of al-Biqā' (Beqaa) valley, under the leadership of a Muslim *imām*, Sayyid Musa as-Sadr; his Movement of the Deprived was based on a call for Shī'ite political rights, economic justice, and defense against Israeli raids in the mainly Shī'ite south.

Hardly a day passed during the rest of 1974 and 1975 without a battle somewhere in Lebanon. The country was torn apart, and the central government virtually ceased to exist. The army, long the mainstay of the government, was immobilized by the nature of the conflict. And the combatants, amply supplied by various foreign groups, turned upon one another with a ferocity—and firepower—almost unequalled in such a small area of the world.

Final phases of the war. During the early phases of the civil war, the Christians had received encouragement from President Sadat of Egypt, while President Assad of Syria had encouraged and aided the left and the Palestinians. The Lebanese and resident Palestinians fought furiously, and gradually the left and the Palestinians began to win the war. By the early months of 1976, it seemed clear that the Christians were losing and that either they would be defeated (so that Lebanon would be reconstituted as a left-dominated, Palestine-oriented state) or Lebanon would be partitioned. Either case appeared to the Syrians likely to bring Israeli intervention. This realization forced a reversal of Syrian policy, first toward restraint on the left and then toward support for the Christians. Ironically, both the Syrians and the Israelis, so opposed to one another on other issues, took up the cause of the Lebanese Christians with essentially the same tools: Syria supplied arms and prevented the Palestinians from taking strategic points, while Israel blockaded Sidon and Tyre, through which arms to the left and the Palestinians had arrived, trained a large military contingent in Israel and supplied it with tanks, and shipped tanks and other equipment into the Christian sector. During the summer of 1976, the Israeli defense minister, Shimon Peres, paid at least four clandestine visits to the de facto Christian capital at Jūniyah, and the Israeli Army occupied sizable parts of southern Lebanon; from the east, beginning in early June, Syrian military units entered the country with about 450 tanks and 20,000 soldiers.

Syrian reversal of position

This realignment led to two new developments. First, the non-Syrian left, represented by Iraq and Libya, resumed close contacts with the PLO, which earlier it had denounced as conservative and ineffectual, and this, in turn, caused President Sadat of Egypt to seek closer relations with the PLO leader Yasir Arafat in order to win Palestinian support for his claim to Arab leadership and to prevent attacks on the Egyptians' separate disengagement agreement (September 1975) with the Israelis. Second, the Christians, with strong support from Syrian tanks and soldiers, began to win the civil war. In late June, Christian forces launched attacks on the two huge Palestinian refugee camps of Jisr al-Bāshā, which fell after one week, and Tall az-Zaatar, which, becoming a symbol of Palestinian resistance, withstood a tragic and bloody siege until mid-August.

In April Pres. Suleiman Franjieh had signed a constitutional amendment to permit the Parliament to elect a new president. Elias Sarkis was elected on May 8 but was not inaugurated until September 23, and then in a Syrian-held town. Meanwhile, Lebanon was effectively partitioned along the "Green Line," which passed through the centre of Beirut (east–west) along the main road to Damascus: to the north was a Christian government, to the south a leftist (Druze–Muslim–Palestinian) government led by Kamal Jumblatt. Repeated attempts were made to bring the fighting to an end, until finally a formal "summit" meeting, held in Cairo on October 25–26 established an Arab League peace-keeping force of 30,000 troops (of whom about 70 percent were Syrian), under the overall command of President Sarkis and headed by a Lebanese Sunnite officer. By the end of November, despite continued minor clashes, the civil war came to a close.

Consequences of the war. The civil war was a catastrophe for the Lebanese, whose country now lay in ruins. For the Palestinians, it not only cost perhaps 20,000 killed and twice that many wounded but also shattered

their leadership and placed in jeopardy alliances and friendships—even in distant Kuwait—upon which they had relied for the 30 years of their dispersion. The Syrians appeared stronger than before, but their victory could turn out to be almost worse than defeat; having got into Lebanon, they faced the problem of extricating themselves without leaving behind strong and enduring enmities. Egypt, whose leadership of the Arab world was certainly endangered by the war, emerged with the divisive separate truce (September 1975) with Israel all but forgotten and with general respect for its role in making peace. Jordan, likewise, found the tragic memory of its Black September overshadowed by not one but many incidents, each blacker than Black September. Only Israel among the states of the Middle East appeared to have "won." The unity of its enemies appeared to be shattered. The Palestinians lost their major bid; Syria acted, apparently, out of fear of their intervention and against its fellow Arabs, virtually in concert with Israel; the Lebanese Christians, likely to regain much of their former power, were in Israel's debt; and the modus vivendi with Egypt was more secure than before. More important, the revulsion and horror of the war had caused Arabs everywhere to question, as never before, the very dream of pan-Arabism. (K.S.S./W.R.P.)

BIBLIOGRAPHY

Syria and Palestine to c. 1550 BC: The following chapters in *The Cambridge Ancient History*, 3rd ed. (1970–), summarize the evidence for Syria and Palestine available up to *c.* 1965–69. All have a full bibliography that can supplement the titles given here. In volume 1, see SIR MAX MALLOWAN, "The Development of Cities from Al-'Ubaid to the End of Uruk 5"; ROLAND DE VAUX, "Palestine During the Neolithic and Chalcolithic Periods," and "Palestine in the Early Bronze Age"; MARGARET S. DROWER and JEAN BOTTERO, "Syria Before 2200 B.C."; and GEORGES POSENER, JEAN BOTTERO, and KATHLEEN M. KENYON, "Syria and Palestine, *c.* 2160–1780 B.C." In volume 2, see JEAN R. KUPPER, "Northern Mesopotamia and Syria"; and KATHLEEN M. KENYON, "Palestine in the Middle Bronze Age." ROBERT J. BRAIDWOOD, "Near Eastern Prehistory," *Science*, 127:1419–30 (1958), provides a description of the Pre-Pottery Neolithic site of Jarmo in northeastern Iraq. J.B. HENNESSY, *The Foreign Relations of Palestine During the Early Bronze Age* (1967), analyzes objects traded among Palestine, Egypt, and western Asia, and thereby establishes a chronology for the Palestinian Early Bronze Age. KATHLEEN M. KENYON, *Amorites and Canaanites* (1966), deals with the breakup of the Early Bronze Age civilization of Syria and Palestine and the development of the Middle Bronze Age culture. The same author's *Archaeology in the Holy Land*, 3rd ed. (1970), a summary of the archaeological evidence for Palestine from the beginning of settlement down to the postexilic period, includes a bibliography of the main excavated sites. JAMES MELLAART, *Earliest Civilizations of the Near East* (1965), is a popular account of the earliest settlements in Anatolia, Iraq, Iran, Syria, and Palestine—includes accounts of the author's finds at Hacilar and Çatalhüyük in Anatolia.

From c. 1550 BC to AD 630: (*Syria and Palestine in the era of the Egyptian and Hittite empires*): PAUL GARELLI, *Le Proche-Orient asiatique, des origines aux invasions des peuples de la mer* (1969), the best one-volume history, covers the entire Near East but should be used selectively for the Levant (very full bibliography). J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament*, 2nd ed. (1955) and *The Ancient Near East in Pictures Relating to the Old Testament* (1954), are excellent source books of translated texts and annotated illustrations, covering a wider period and area than this article does. JOHN BRIGHT, *A History of Israel* (1959); and MARTIN NOTH, *Geschichte Israels*, 3rd ed. (1956; Eng. trans., *The History of Israel*, 2nd ed., 1958), are the two best single-volume histories of ancient Israel in the English language, approaching the subject from rather different viewpoints. HAROLD H. ROWLEY, *From Joseph to Joshua* (1950), although rather out of date, contains a valuable discussion of all the various theories concerning the Patriarchal and Exodus narratives. The conclusions are not always reliable. ARVID S. KAPELRUD, *Ras Sjamra-funnene og det Gamle Testament* (1953; Eng. trans., *The Ras Shamra Discoveries and the Old Testament*, 1963), is a brief introduction to the subject. (*Syria and Palestine from c. 1200 to c. 330 BC*): For this period, see BRIGHT and NOTH (*op. cit.*); and ANDRE DUPONT-SOMMER, *Les Araméens* (1949). DONALD B. HARDEN, *The Phoenicians*, 2nd ed. (1963); and SABATINO MOSCATI, *Il mondo dei Fenici* (1966; Eng. trans., *The World of the Phoenicians*, 1968), are both good, well-illustrated accounts of Phoenician history and civilization. LUDWIG KOHLER, *Der Hebräische Mensch* (1953; Eng. trans., *Hebrew Man*, 1956), is a brief, but stimulating and scholarly, survey of Hebrew life and thought. ROLAND DE VAUX, *Les Institutions de l'Ancien Testament*, 2 vol. (1958–60; Eng. trans., *Ancient Israel: Its Life and Institutions*, 1961), is a thorough and detailed study of the civil, military, and religious institutions of ancient Israel (very full bibliography). EKREM AKURGAL, *Die Kunst Anatoliens von Homer bis Alexander* (1961; Eng. trans., *The Art of Greece: Its Origins in the Mediterranean and Near East*, 1968), is a profusely illustrated study of Assyrian, Aramaean, and Phoenician art and their influences. (*Syria and Palestine in the Hellenistic and Roman periods*): For these periods, see ROBERT M. GRANT, *A Historical Introduction to the New Testament* (1963); AVIGDOR TCHERIKOVER, *Hellenistic Civilization and the Jews* (1959; orig. pub. in Hebrew, 1930), a fairly detailed study by one of the foremost authorities on the subject; A.H.M. JONES, *The Herods of Judaea* (1938), still one of the best and most reliable sources, despite the vast number of more recent books on the Herodian family; and the same author's *Cities of the Eastern Roman Provinces*, 2nd ed. (1971), the standard work on the subject, with full bibliography.

From the Islāmic Conquest to c. 1920: MARIUS CANARD, *Histoire de la dynastie des H'amdanides de Jazîra et de Syrie*, vol. 1 (1951), a detailed study of the Ḥamdānid dynasty; MARSHALL W. BALDWIN and KENNETH SETTON (eds.), *A History of The Crusades*, 2 vol. (1955–62; 2nd ed., 1969–), contains chapters by leading scholars on all aspects of the Crusades, particularly good on the Muslim viewpoint—covers the Seljuqs, Ayyūbids, and Fāṭimids in detail; STEVEN RUNCIMAN, *A History of the Crusades*, 3 vol. (1951–54), a sound general history of the Crusades; A.S. ATIYA, *The Crusades in the Later Middle Ages* (1938), a very useful and readable account; CLAUDE CAHEN, *La Syrie du nord à l'époque des Croisades et la principauté franque d'Antioche* (1940), a detailed account of Syria at the time of the Crusades, intended for the serious student; MAURICE GAUDEFROY-DEMOMBYNES, *La Syrie à l'époque des Mamelouks d'après les auteurs Arabes* (1923), another detailed work; NICOLA ZIADEH, *Urban Life in Syria Under the Early Mamlūks* (1953), a useful supplement to the political histories, more general in scope than its title implies; PETER M. HOLT, *Egypt and the Fertile Crescent, 1516–1922* (1966), the best work on Syria for the Ottoman and early modern periods; A.N. POLIAK, *Feudalism in Egypt, Syria, Palestine, and the Lebanon, 1250–1900* (1939), useful for the study of institutions; HENRI LAMMENS, *La Syrie, précis historique*, 2 vol. (1921), a useful outline of events that tends to be rather controversial; PHILIP K. HITTI, *History of Syria*, 2nd ed. (1957), a general history that includes Lebanon and Palestine; and *The Origins of the Druze People and Religion* (1928), a good general account of this important minority; IRA M. LAPIDUS, *Muslim Cities in the Later Middle Ages* (1967), a penetrating introduction to the Mamlūk period; ABDUL-KARIM RAFEQ, *The Province of Damascus*, 1723–1783 (1966), covers some of the same ground for the Ottoman period; MOSHE MA'OZ, *Ottoman Reform in Syria and Palestine, 1840–1861* (1968), for 19th-century political and administrative change.

Syria and Palestine since 1920: STEPHEN H. LONGRIGG, *Syria and Lebanon Under French Mandate* (1958), tells in detail the story of the two countries from 1918 until 1947 and is indispensable as a reference on the subject. KAMAL S. SALIBI, *The Modern History of Lebanon* (1965), traces the history of the country from the 18th century to date and deals with the period since 1918 in pt. 2, entitled "Greater Lebanon." There is no survey of the history of Syria since 1920, but useful information on the subject is found in PATRICK SEALE, *The Struggle for Syria* (1965), which gives an intelligent and vivid account of the peculiar developments in the Syrian Republic between 1943 and 1958. ALBERT HOURANI, *Syria and Lebanon: A Political Essay* (1946), retains its importance as a sensitive and well-thought-out study of the two countries as they emerged into independence. The material available on the Palestine question, Israel, and Arab–Israeli relations is vast, but hardly any of the accounts available are objective. CHRISTOPHER SYKES, *Crossroads to Israel* (1965), tells a convincing story of the developments that led to the establishment of Israel on the former territory of Palestine. NORMAN BENTWICH, *Israel* (1952), considers the developments in Israel during the early years after its establishment. See also KAMAL S. SALIBI, *Crossroads to Civil War: Lebanon 1958–1976* (1976), dealing with the events leading up to the civil war and the war; and WILLIAM R. POLK, *The United States and the Arab World*, 3rd ed. (1975).

(K.M.K./P.J.Pa./C.C./A.H.Ho./V.E.I.)

Syrian and Palestinian Religions

Syrian and Palestinian religions, originating in the Stone Age and continuing into Hellenistic times, absorbed religious beliefs and practices of ancient Near Eastern religions—as well as their own indigenous cultic phenomena—and transmitted them to biblical, Greek, and Roman religions. The Syrian and Palestinian religions influenced Judaism, Christianity, and Islām insofar as these religions reacted to Syrian and Palestinian religious beliefs and practices.

NATURE AND SIGNIFICANCE

At least as early as the 3rd millennium BC, the area comprising what was later known as Syria and Palestine was the home of a linguistically and ethnically mixed population and the meeting place of cultural influences from all sides: Egypt, Mesopotamia, and Asia Minor. This complexity increased in later centuries and had as its counterpart a very complex religious picture. The typical political unit in the area was a city-state of limited geographic extent. Each of these states had its own pantheon and cult, and within the larger states there might be different ethnic and social groups carrying on their distinct religious beliefs and practices.

Common features

Certain common features can be discerned, however. The religion of each of these peoples was characterized by the worship of numerous gods. These gods were mostly conceived in anthropomorphic terms and images and were identified with, or guardians of, important elements of the universe and vital aspects of nature and of human society. The competing powers in the divine world were thought to be organized, rather loosely, into a council of gods, in which the storm god played an especially dominant role. Images of the gods stood in their houses, the temples, where organized priesthoods saw to it that they were supplied with food and clothing. Important conceptions as to the nature of the world, and of human society, were embodied in myths, which often took the form of narrative poems of considerable length and complexity. The dominant concerns of official religion were for the economic bases of society—that is, for abundant crops and fertility of herds and flocks—and for the state, especially for the well-being of the ruling king and the continuance of his line. Scholars have reasonably full information only about the religion of the Canaanites—one of the ancient peoples of Palestine–Syria—as attested in the Ugaritic texts (see below *Sources*) from the Late Bronze Age (15th–13th centuries BC), supplemented by other materials; hence, for practical reasons, Canaanite religion must occupy a central place in a description of the area's religions. Fortunately, there is good reason to suppose that Canaanite religion was typical and representative and maintained its dominant characteristics through the centuries, so that knowledge of that religion probably indicates the general nature and features of most of the other religions of the area.

The religions of ancient Syria–Palestine are of great intrinsic interest because of the variety of phenomena they encompass and because of the great antiquity of the conceptions they preserve, some of which doubtless derive from the Stone Age (*c.* 4000 BC). But beyond this interest, these religions are of extraordinary importance to the student of ancient Israelite religion. On the one hand, Canaanite religion constituted the antithesis to Israelite religion; and, on the other hand, it supplied a rich heritage of religious language, ideas, and ritual practices on which the Old Testament writers drew to an extent that is only beginning to be realized. Furthermore, Syrian religion enjoyed a vogue not only in the Hellenistic world (3rd century BC–3rd century AD) but centuries earlier, in the Minoan–Mycenaean age (*c.* 3000–*c.* 1100 BC), it exercised an influence on early Greek religion.

SOURCES

Evidence for reconstruction of the ancient Syrian and Palestinian religions has increased astonishingly within the last 50 years, due to materials recovered by excavation; even so, however, the record is full of gaps. Moreover, ancient religious texts tend to be more difficult to interpret than historical records, business documents, or other such recovered materials.

Old Testament, Greek, and Roman sources

The Old Testament was, until comparatively recent times, the most ancient of sources available to scholars for study of the religions of Syria–Palestine. It supplies the names of some major deities, such as Baal (male fertility god) and Asherah (female fertility goddess); some cultic terms; and evidence for the major concerns of the native religion. The story of the Israelite prophet Elijah's contest with the prophets of Baal on Mt. Carmel (I Kings 18) is especially rich in information. But biblical evidence is scattered, fragmentary, and limited in value by its polemical character. Greek and Latin sources have long been drawn on to fill in the gap in knowledge of this subject, especially for Punic religion—referring to the Syrian-derived religion of Carthage in North Africa—and for Syrian religion in the Hellenistic world. Along with inscriptions and coins, scattered passages in classical writers shed light on Syrian religion. Thus, the Roman historian Dio Cassius (*c.* AD 150–235) describes the devotion of the Syrian-born Roman emperor Elagabalus (reigned AD 218–222) to the sun god of his native Emesa, and Apuleius (flourished 2nd century AD) paints a vivid picture of the obscene rites that attended the worship of the Syrian Goddess (*Metamorphoses* 8–9). *De Dea Syria* (On the Syrian Goddess), attributed to the satirist Lucian (*c.* AD 120/125–*c.* 190), describes the temple and cult of the goddess at Bambyce (Hierapolis) in Syria. The 4th-century Christian historian Eusebius in his *Praeparatio evangelica* (Preparation for the Gospel) gives extensive extracts from the "Phoenician History" of Philo of Byblos (2nd century AD), who in turn claims to have translated the work of one Sanchuniathon, a Phoenician priest. The work is a treatise on Phoenician religion, especially cosmogony (origin of the world) and theogony (origin of the gods). As long as no independent confirmation of its authenticity was available, it was possible to dismiss Philo's work as a farrago of elements current in Hellenistic religious syncretism, until the Ugaritic evidence referred to above confirmed its authenticity and antiquity. One can scarcely accept Philo's dating of Sanchuniathon to before the Trojan War (*c.* 12th century BC), but a leading biblical archaeologist, W.F. Albright (1891–1971), has dated him to about the 6th century BC.

Phoenician, Punic, and Ugaritic sources

Phoenician and Punic inscriptions were first effectively deciphered in 1837 by the German linguist H.F.W. Gesenius (1786–1842). Punic inscriptions are very numerous but are mostly votive and provide little beyond the bare names of deities and a few epithets, plus theophoric names (those including the name of a god) of the offerers and some sacrificial terms. An inscription called the Sacrificial Tariff of Marseilles (2nd or 3rd century BC) lists various sorts of offerings and the payment due to the officiating priest in each case; a similar tariff from Carthage is also extant. Phoenician and Aramaic funerary, votive, and building inscriptions, the earliest of them dating to the 10th century BC, have been discovered in Syria–Palestine itself. Egyptian texts of the New Kingdom (18th–20th dynasties, *c.* 1567–*c.* 1085 BC) make fairly frequent mention of Syrian deities who had come to be venerated inside Egypt. A text in the Hittite language contains an originally Canaanite myth concerning Ilkunirsha (perhaps "El, Creator of the Earth") and his consort. Evidence for a very early stage of Semitic religion (early 2nd millennium BC) can be gleaned from Amorite theophoric personal names that have been preserved in cuneiform texts from Mari on the mid-Euphrates and from Alalakh in North Syria.

By far the most extensive and most important sources, however, are the Ugaritic texts. Since 1929, excavations at Ras Shamra on the Syrian coast opposite Cyprus have been yielding texts written in an alphabetic cuneiform script on clay tablets; the language—a Semitic tongue closely related to Phoenician and Hebrew—and the script are called Ugaritic from Ugarit, the ancient name of the city. Many of the tablets contain religious texts, myths, incantations, rituals, offering lists, and lists of gods. The tablets date from the Late Bronze Age, but the mythological texts may have been composed centuries earlier.

Written evidence is supplemented by a great quantity of strictly archaeological material from numerous excavations in Palestine and Syria: temples, cultic utensils, amulets, and images and symbols of the gods.

HISTORICAL AND SOCIAL BACKGROUND

Prehistory through the 3rd millennium BC. Behind the history of religion in Syria–Palestine stretches a long prehistory that is known only in fragments from the evidence of archaeology. By the 7th millennium BC, at Jericho, some characteristic elements of later religion can be discerned. British excavations of the prepottery Neolithic phase of the city discovered animal figurines that were probably votive offerings, and a figurine of a mother goddess proffering her breasts. A small shrine held a sacred standing stone, a type of cult object used very widely in later ages. Veneration of ancestors is attested for Stone Age Jericho in the groups of carefully preserved skulls, with plaster molded over the bones to reproduce the features of the living person.

Evidence of ancient religion at Jericho

By the 3rd millennium BC, when written evidence begins to be available in meagre quantity, Syria–Palestine was already inhabited by speakers of a Semitic language of a Canaanite type. Moreover, a large share of the ancient place-names are Semitic. Some, such as Beth Yerakh (House of the Moon god), shed a bit of light on the religion of this early time. Other place names suggest occupation of some areas by non-Semitic groups. Moreover, there is evidence that the Egyptians in the Old Kingdom (*c.* 2686–*c.* 2160 BC) already controlled Byblos on the Phoenician coast and exercised cultural and political influence over much of Syria–Palestine. Characteristic features of the religious history were thus already present—that is, a mixed population and an openness to foreign influence.

2nd millennium BC. In the early centuries of the 2nd millennium, the Amorites, a Semitic people, pushed into the centres of civilization all around the Fertile Crescent. At Mari on the Euphrates and also at many other sites in Syria and Mesopotamia, Amorites had gained the upper hand. The Egyptian Execration Texts (documents by means of which the Egyptians cursed their enemies) show that groups of Amorites also dominated throughout Palestine. Since many of the Amorite personal names—the only source for knowledge of the Amorite language—contain the name of a god and an assertion about the god, a reconstruction of some aspects of the religion is possible, though the limitations of the evidence are obvious.

Following the wave of Amorite settlement came a period of incursion by Hurrians and Indo-Europeans, beginning about the 18th century BC. By the 15th century BC, northern Syria was dominated by these elements, and many cities even in southern Palestine had kings with non-Semitic names. In the 15th century BC, kings with Indo-European names ruled a strong Hurrian state known as Mitanni, which controlled all of north Syria. The rise of the Hittite Empire spelled the end of Mitanni, and through the remainder of the Late Bronze Age, Syria–Palestine was the scene of a struggle for domination between Egypt and the Hittites. At the same time, the Akkadian language was in widespread use for communication between states throughout the Near East, and scribes in Syria–Palestine were trained in the literary, religious, and scholarly works of Mesopotamia. Trade by sea with Cyprus, Crete, and the Aegean flourished. As might be expected, then, the Late Bronze Age presents a complex religious picture: a great many local cults were scattered throughout Syria–Palestine, and the possibilities for transmission and borrowing from one area to another were practically unlimited. A treaty (*c.* 1380 BC) of the Hittites with Mitanni illustrates the complexity in its list of the gods of Mitanni who are regarded as witnesses to the treaty, for the deities named include Hurrian figures such as Teshub, the storm god, and Shimegi, the sun god; the Indian gods Mitra, Varuṇa, Indra, and the Nasatyas; and originally Mesopotamian deities, such as Anu and Enlil.

Syncretistic developments

The political picture changed decisively at the transition from the Late Bronze to the Iron Age (*c.* 1200 BC). At about that time, certain peoples with ties to older nomadic groups established themselves as independent kingdoms in various parts of the region: the Aramaeans in various cities in Syria, down as far south as Damascus; the Ammonites, on the highlands east of the Jordan; Moab, east of the Dead Sea; and Edom, south of the Dead Sea. The Hittite Empire disintegrated, leaving behind a number of states in Syria that maintained Hittite cultural traditions. The "Sea Peoples," among them the Philistines, established themselves along the Mediterranean coast from the Egyptian border north to Dor on the coast of Palestine. Israel became master of most of Palestine west of the Jordan and held some lands east of the river. The direct descendants of the ancient Canaanites, the Phoenicians, were left in control of a small strip along the Syrian coast.

Only meagre evidence exists for the religion of the Aramaeans, Ammonites, Moabites, Edomites, and Philistines. Such evidence as survives suggests that they continued the traditions of Late Bronze Age polytheism in various ways, with no fundamental change in type. Even the non-Semitic Philistines worshipped gods under names familiar in the region for centuries before their arrival, such as the grain god Dagon and the fertility goddess Astarte. Israel introduced a new element, an exclusive monotheism that brought it into a life-and-death struggle with local religions.

1st millennium BC. The Phoenicians found an outlet for their energies in trade and colonization of lands bordering the Mediterranean, especially in Spain, Sardinia, and North Africa. Carthage, founded in the late 9th century BC according to tradition—a date that agrees well with modern research—became the principal Phoenician foundation. Carthage fell to Rome in 146 BC, but veneration of the old gods survived for centuries thereafter. In Syria–Palestine, although the old independent city-states were incorporated in successive empires—Assyrian, Neo-Babylonian, Persian, Hellenistic, and Roman—the local cults survived. At famous centres such as Baalbek, Emesa, and Hierapolis, a descendant of the ancient Canaanite religion flourished, and the Syrian goddess Atargatis was widely worshipped in the Eastern Roman Empire. After the triumph of Christianity, and later of Islām, vestiges of Syrian religion survived only at the level of folk religion.

MYTHOLOGY AND RELIGIOUS PHENOMENA

The gods. Ancient Ugaritic scholars prepared ordered lists of their gods, and one such order is followed here, omitting numerous minor figures. The first in the series of divine names is Il-a-bi, which admits various interpretations: "the god of the (or my) father(s)" or "the father god" or still others. The lack of certainty on this point leaves open the question as to whether this figure is in any way related to the "god of the fathers" or "the god of my father" frequently mentioned in the Old Testament stories of the patriarchs Abraham, Isaac, and Jacob.

El. The next deity is El, whose name is the same as the common noun "god," a word that appears in varying forms in most Semitic languages. El is the titular head of the pantheon and as such is called king. He is the father of the gods (except for Baal) and is also the creator of man. Phoenician and Aramaic texts give him the title Creator of the Earth, a title expanded in the Bible to "maker of heaven and earth" (Gen. 14:19). He is identified with the Hurrian Kumarbi and with Greek Cronos. El is thought of as an old man with a long gray beard and as such has the title "father of years" (other translations of the epithet have been preferred). As another epithet, "Bull," suggests, he was renowned for his sexual powers, both in myth and in artistic representations, where he is depicted with the attributes of the ithyphallic (having an erect penis) Egyptian god Min. As the "Kindly One, Merciful El," he is the bestower of good gifts on men in various mythic episodes. In one strange burlesque, he is almost a figure of fun, roaring drunk at his own party. In Ugaritic religion, El is relatively inactive compared to the storm god, Baal; but in Sanchuniathon, El is more active and rather sinister. Like the Greek Cronos (Saturn), he emasculates his father, Heaven, and he slays some of his children.

The Creator God

Dagon. Dagon (Dagan) comes next in the list, presumably because he, not El, was thought of as the father of Baal and because of his great importance in the city's cult, even though he plays no role in extant myth. The original meaning of the name is uncertain, but it became a common noun in Ugaritic and Hebrew for "grain," which sufficiently indicates that he was associated with fertility of crops. Worship of Dagon is attested at Mari almost a millennium earlier than at Ugarit, and Dagon later became the chief god of the Philistines.

Baal. Baal was the storm god and as such occupied a prominent place in both cult and myth. The ancient proper name of the deity is Hadad, but this title is much less common at Ugarit than Baal, which is also a common noun meaning "Lord, Master." As storm god, Baal is called "cloud rider," a title adopted in Israel for Yahweh (Ps. 68:4). He is also "the Prince" and "the Conquering One." His residence is on a mountain north of Ugarit on the Syrian coast whose Semitic name was Zaphon. The Hittite name was Khazzi, which becomes Kasios in later

Importance of the storm god

Giraudon



Storm god riding on a bull, brandishing a mace and thunderbolt. Relief from Arslan Tash; 8th century BC. In the Louvre, Paris.

Greek references to the god and the mountain, the Canaanite Olympus. Under the title Baal Zaphon, the god was worshipped as a patron of sailors, a cult transferred to a site in the Egyptian delta near Pelusium. Though there is now no doubt that Baal/Hadad was a great god, comparable to Zeus in nature and importance, the situation is complicated by a number of factors. Since *baal* means "lord," it is often combined with place-names to form divine names of the pattern "Lord of such and such a place" (*e.g.*, Baal-peor, "Lord of Peor"). In some cases these names may designate a separate minor deity rather than the great storm god; on the other hand, the "baals" of various localities may in other cases be local manifestations of the one major figure. In a given case, it is often impossible to decide. In the Old Testament, a plural form, *ha-bealim* ("the Baals"), is of frequent occurrence. It may refer to a variety of local deities, but in many cases the plural is a "plural of majesty" referring to a single god; an honorific plural, *baaluma*, occurs at Ugarit.

Yarikh and Kothar. Yarikh, "Moon," was a male deity widely and very anciently worshipped, but little mythological detail survives. In general, the sun—a goddess at Ugarit—and the moon and stars play relatively minor roles at Ugarit. Kothar was the Canaanite craftsman god, similar to the Greek Hephaestus. The name means "skillful," and his other title is Khasis, "clever."

Fertility goddesses

Asherah, Anath, and Astarte. Three goddesses are especially prominent: Asherah, Anath, and Astarte. Asherah, the consort of El at Ugarit, has the fuller name "She who walks in the Sea" and is also called "Holiness." The Old Testament uses the name Asherah frequently, but usually for a wooden cult object associated with her worship rather than for the goddess herself. Astarte (Ashtoreth in the Old Testament) is a goddess of love and war, the approximate equivalent of Mesopotamian Ishtar. Overshadowed by Anath in Ugaritic mythology, she is much more prominent in cultic texts. In the Old Testament, she is the most frequently mentioned of the pagan goddesses; and her name is also a common noun, "sheep breeding."

Anath is the chief Ugaritic goddess of love and war. The consort of Baal, she plays an important role in the myth of Baal's death and resurrection—that is, in the cycle of fertility and infertility in nature. Her outstanding trait is her ferocity in battle. In one scene, she wades in blood and wears heads and hands as ornaments. There is evidence that suggests a connection between Anath and the ferocious Indian goddess Kālī. In the Hellenistic period, Anath and Astarte are blended into a single figure, Atargatis. She is the "Syrian Goddess," whose cult was famous for its *galli*, devotees who voluntarily emasculated themselves for the goddess and thereafter dressed as women.

Mot, Melqart, Adonis, and other gods. Mot is what his name, "Death," implies, god of the dead and of all powers that oppose life and fertility. Resheph is a related figure, a god of the underworld and of pestilence. Some abstractions were deified and occur in the god lists, notably Mishor and Suduk, approximately "Justice" and "Righteousness." Many other deities are attested at Ugarit, and other sites swell the list of names.

Carthaginian deities

The chief god of Tyre was Melqart, who was also head of the pantheon at Carthage. The name means "King of the city"—that is, "of the underworld city" in the most plausible interpretation. Greeks and Romans identified him with Hercules. Baal Hammon, perhaps "Lord of the Incense Altar," is mentioned only very rarely in inscriptions from Syria but becomes very important at Carthage, where the name occurs in hundreds of dedicatory inscriptions. Little is known of his nature. The principal Carthaginian goddess is Tanit (or Tinnit). No explanation of the name is generally accepted; usually it occurs in the form "Tanit face of Baal." Little of a positive nature can be asserted as to her nature, though, not surprisingly, she seems to have to do with the heavens and with fertility.

Adonis, whose name is a Semitic word, "Lord," is known only from relatively late non-Semitic sources. A youthful god of vegetation beloved of the goddess of love, he is according to one myth killed by a boar, his blood staining the river Adonis. Women took part in rites of mourning for him and set out ephemeral "Adonis gardens" in his honour. The cult reached Athens by the 5th century BC from Cyprus and was widespread in the eastern Mediterranean.

The chief god of Moab was Chemosh; his name occurs also in a list of gods at Ugarit, but little can be said about his nature; still less is known of Milkom, god of the Ammonites.

Myths and legends. *Theogony and cosmogony.* Philo of Byblos gives Sanchuniathon's account of the origins of the cosmos, the gods, and the arts of civilization. Many of the figures in the latter portion of the account can now be identified with ancient Canaanite deities, and in the sequence of divine generations, the account is strikingly like that of the Greek poet Hesiod's (flourished *c.* 800

BC) *Theogony* and the Hittite–Hurrian creation myth of Kumarbi, but sufficiently distinct to provide assurance that it is an ancient and independent variant of a common tradition. Thus, Sanchuniathon tells of four generations of gods: Eliun or Hypsistos ("the Most High"), Uranos, El (=Kronos), who castrates his father, and Demarus (=Zeus), all of this very like the *Theogony* and Kumarbi, but with numerous differences in details. This very archaic theogony is preceded by an atheistic cosmogony that is probably the product of later Phoenician speculative thought; it resembles the cosmogonies of the Pre-Socratic philosophers.

The Baal epic. From Ugarit comes the longest of known Canaanite myths, the Baal epic, probably composed in the early 2nd millennium BC. (The order of the separate tablets is not completely certain.) Baal is challenged to battle by the god Yamm (Sea). Equipped by Kothar with special clubs, Baal defeats Yamm. Baal, though supreme in power, has no house like those of the other gods; finally Kothar builds and equips a house for him, magnificent with gold and silver. After a gap in the text comes an episode in which Baal descends into the jaw of death, the god Mot. When the news reaches El, he mourns extravagantly, as does the goddess Anath, who then sees to the burial of Baal's body. She seeks out Mot and cuts him to pieces. Then Baal is resurrected and "the heavens rain oil, the watercourses flow with honey."

Other myths. Briefer myths recount the wedding of Nikkal, a moon goddess, to the moon god Yarikh, and El's seduction of two females, who bear the children Dawn and Dusk. Creatures of voracious appetite, they are driven into the wilderness; at this point the text breaks off. In one mythological fragment, Anath espies the beauty of her brother, then "she eats his flesh without a knife; she drinks his blood without a cup"—a striking parallel to the omophagy (eating of live victims, occasionally human) practiced by raving devotees of the Thracian and Phrygian god Bacchus. Elsewhere there is an allusion to Baal's defeat of Lotan, "the primeval serpent . . . the twisting serpent." Lotan is the same as the Hebrew Leviathan, and the same motif occurs in the Old Testament (*e.g.*, Isaiah 27), with Yahweh as the dragon slayer.

Legends. Other Ugaritic epics are less strictly mythological in that they involve human characters. The Keret epic concerns a king Keret whose sons have all died. In a vision Kindly El advises him to undertake a military campaign to get a new wife. After careful preparation, Keret besieges the city Udum and compels its king, Pabel, to grant him his daughter Hurriya in marriage. Hurriya bears Keret many children. In the last episode of the extant story, Keret falls ill and is saved only by El's healing intervention. The so-called *Aqhat Epic* concerns the sage, Danel (the ancient worthy Daniel mentioned in Ezekiel 14), who has no son. In response to Danel's prayer and offerings, El grants him a son, Aqhat. Later Danel gives Aqhat a marvellous composite bow made by Kothar. The goddess Anath covets it, but Aqhat spurns her and she arranges for him to be killed. Danel and his daughter Paghat discover the crime and take steps to avenge it; here the text breaks off.

PRACTICES AND INSTITUTIONS

The importance of sacrifice

Sacred customs and ceremonies. Animal sacrifice occupied a central place in the religion. Few details about the rites observed are known, but since much of the Ugaritic terminology is the same as later Israelite terminology (*e.g.*, "sacrifice," "burnt offering," "peace offerings," "wave offerings"), scholars contend that they are justified in conceiving Canaanite sacrificial practices to be similar in many respects to those of the Israelites. The animals commonly presented were sheep, goats, cattle, and birds; the Marseilles Tariff (see above) mentions also young deer and other game. The extent to which human sacrifice was carried on is uncertain, but there is no doubt that it was practiced to some extent, judging from biblical allusions and evidence for human sacrifice at Carthage. Sacred prostitutes, both male and female, were among the personnel of some shrines, but details as to the extent and significance of this institution are lacking.

Kings, priests, and prophets

Religious authorities. The king and queen were important religious figures, both because offerings and prayers were made for them and because they took a leading part in certain rituals, to judge from brief references in Ugaritic texts. The relatively elaborate official cult at Ugarit was led by a "chief priest" and by several lesser orders of priests. There were special temple singers and "drawers of water for the sanctuary" (compare the role of the biblical Gibeonites). The temples were economic as well as cultic centres. Estates were assigned to orders of priests, and a certain chief priest is said to have been also "chief of the shepherds"—that is, overseer of the temple flocks.

Prophets—that is, persons inspired by a god to deliver a message in his name—were a very ancient feature of the religion. In the Mari texts (dated *c.* 18th century BC), there are repeated references to ecstatics, some of them women, who appear unsolicited in the temples to deliver messages from the gods to the king. A similar phenomenon is attested for Byblos in the Wen-Amon report, an Egyptian text of about 1100 BC: a youth falls into an ecstatic state and delivers an oracle on behalf of the Egyptian envoy, which is respected by the ruler of Byblos. The prophets of the Tyrian Baal used dancing and self-laceration to achieve a state of frenzy (I Kings 18). A long charm against snakebite from Ugarit suggests that not only this variety of incantatory magic but others as well were cultivated in Syria–Palestine.

Shrines and festivals

Sacred places and sacred times. The temple, or house of god, was the place where the god, symbolized by his image, dwelt and thus the point at which heaven and earth met. Since, according to biblical tradition, Solomon's Temple was built by a Phoenician architect, the biblical descriptions apparently give us the fullest and most coherent idea of Syrian temple architecture. The basic tripartite plan, with porch, holy place, and most holy place, is illustrated in the small temple excavated at Tall Tainat in north Syria (9th–8th century BC). In addition to the elaborate temples in the cities, there were the more rustic shrines known as "high places" (Hebrew *bamot*), built on natural ridges or artificially built up and often associated with the cult of dead heroes. The principal cult objects were statues of the gods, but typical also is veneration of various deities in the form of unshaped stones, well exemplified in the black stone of Emesa that Elagabalus brought to Rome.

Very little is known about the sacred calendar of the region. At Ugarit, important sacrifices were offered on the day of the new moon and on the 14th of the month. Since the three great feasts of ancient Israel—that is, Unleavened Bread, Weeks (Pentecost), and Booths—were evidently native agricultural feasts originally, it is probable that from time immemorial the farmers of Syria–Palestine kept a spring festival at the beginning of barley harvest, a feast at the end of the grain harvest, and a fall vintage festival.

RELIGIOUS ART AND ICONOGRAPHY

Though excavations have yielded statues and reliefs depicting gods or symbols of the gods, it is frequently difficult to be certain about identification of a given representation with a particular deity known from written evidence. Baal and similar storm gods are frequently shown as riding on a bull and brandishing a mace or thunderbolts. The symbol of Tanit, very commonly found on Punic steles, is basically a triangle having a disk at its apex, with a horizontal bar in between; this basic form is often embellished. The source of the symbol and its significance are not known. In Syria–Palestine the most commonly found religious artifacts are small terra-cotta figurines depicting females, often naked and with exaggerated sexual characteristics. Some of these are probably intended to represent fertility goddesses, but none can be certainly identified with any one goddess. Winged human-headed lions (cherubim) and similar composite creatures are very frequent in Canaanite religious art, as in the cultic stand from Taanach (Palestine; 10th century BC).

BIBLIOGRAPHY. For translations of Ugaritic texts and major Phoenician inscriptions, see J.B. PRITCHARD (ed.), *An-*

cient Near Eastern Texts Relating to the Old Testament, 3rd ed. (1969). For the material remains of Canaanite religion, see Pritchard's *The Ancient Near East in Pictures,* 2nd ed. (1969). A one volume condensation of the earlier editions of both these works is available in paperback, *The Ancient Near East: An Anthology of Texts* (1958). G.R. DRIVER, *Canaanite Myths and Legends* (1956); and T.H. GASTER, *Thespis: Ritual, Myth and Drama in the Ancient Near East* (1950), contain some Ugaritic materials not in the Pritchard volume. E.H. GIFFORD, *Eusebius, Praeparatio Evangelica,* 4 vol. (1903), may be consulted for Philo of Byblos' Phoenician History in English. For the historical background, see W.F. ALBRIGHT, "The Role of the Canaanites in the History of Civilization," in G.E. WRIGHT (ed.), *The Bible and the Ancient Near East* (1961), and *Archaeology and the Religion of Israel* (1953), for an excellent summary sketch of Canaanite religion, rich in detail. M.H. POPE, *El in the Ugaritic Texts* (1955), is an important study of a major god. All the gods and myths are treated authoritatively by M. POPE and W. ROLLIG in H.W. HAUSSIG (ed.), *Wörterbuch der Mythologie,* vol. I (1965).

(D.R.H.)

Systems Engineering

Systems engineering, a relatively new field, is not so much a branch of engineering as it is a technique for applying knowledge from the older branches in effective combination to solve a multifacetted engineering problem. It is related to operations research (*q.v.*) but differs from it in that it is more a planning and design function, frequently involving technical innovation, whereas operations research concentrates on the analysis of how existing systems work. Probably the most important aspect of systems engineering is its application to the development of new technological possibilities with the specific objective of putting them to use as rapidly as economic and technical considerations permit. In this sense it may be seen as the midwife of technological development.

The word systems is frequently used also in other combinations, especially when elements of technological advance are not so important. Systems analysis is an example. Systems theory or sometimes systems science is frequently applied to the analysis of physical dynamic systems, usually linear or nearly so. An example would be a complex electrical network, perhaps with one or more feedback loops, characterized by the partial reversion of the effects of a process to their source.

In the development of the various engineering disciplines in the 19th and 20th centuries, considerable overlap was inevitable among the different fields; for example, chemical engineering and mechanical engineering were both concerned with heat transfer and fluid flow. Further proliferation of specializations, as in the many branches of electrical and electronic engineering, such as communications theory, cybernetics, and computer theory, led to further overlapping. Systems engineering may be seen as a logical last step in the process. Systems engineers frequently have an electronics or communications background and make extensive use of computers and communications technology. Yet systems engineering is not to be confused with these other fields. Fundamentally a point of view or a method of attack, it cannot be identified with any particular substantive area. In its nature and in the nature of the problems it attacks, it is interdisciplinary, a procedure for putting separate techniques and bodies of knowledge together to achieve a prescribed goal in an effective manner.

Systems engineering as a point of view

In general, a systems engineering approach is likely to differ from a conventional design approach by exhibiting increased generality in its basic logical framework and increased concern with the fundamental objectives to be achieved. Thus, at each stage the systems engineer is likely to ask both why and how, rather than merely how.

In addition to systems engineering, it is important to define systems themselves. The systems with which a systems engineer is concerned are first of all man-made; second, they are large and complex; their component parts interact so extensively that a change in one part is likely to affect many others. Unless there is such interaction, there is little for the systems engineer to do, at least at the systems level; he can turn immediately to the components themselves. Another important characteristic of systems is that their inputs are normally stochastic, that is, the inputs are essentially random functions of time, although they may exhibit statistical regularities. Thus, one cannot expect to foresee exactly what the system will be exposed to in actual operation, and its performance must be evaluated as some sort of statistical average of the responses to a range of possible inputs. A calculation based on a single precisely defined input function will not do.

Systems may also vary depending on the amount of human judgment that enters into their operation. There are, of course, systems such as electrical circuits or automated production equipment that may operate in a completely determinate fashion. At the other extreme, there are management and control systems, for both business and military purposes, in which machines in a sense do most of the work but with human supervision and decision making at critical points. Clearly these mixed human-machine systems offer the greatest variety both of possibilities and problems for the systems engineer.

THE DEVELOPMENT OF SYSTEMS ENGINEERING

The systems approach stems from a number of sources. In a broad sense it can be regarded as simple extension of standard scientific methodology. It is a common procedure in science (and elsewhere) to list all of the factors that might affect a given situation and select from the complete list those that appear critical. Mathematical modelling, perhaps the most basic tool in systems engineering, is a technique encountered in any branch of science that has become sufficiently quantitative. Thus, in this broad sense, the systems approach is simply the inheritor of a tradition that is generations, if not centuries, old.

In looking for more recent and more specific sources for the systems approach, on the other hand, there are two in particular that stand out. First is the general field of communications, particularly commercial telephony, where systems engineering first appeared as an explicit discipline in its own right. Traces of the systems approach are to be found in telephone engineering at least as far back as the beginning years of the century, and systems ideas were fairly common in telephony by the 1920s and 1930s. When Bell Telephone Laboratories, the research arm of the American Telephone & Telegraph Company, was officially incorporated in 1925, its two principal engineering divisions were called respectively Apparatus Development and Systems Development. A complete formal doctrine of the role of systems engineering, however, first emerged in the years after World War II, as part of an effort to redefine the policy and structure of the research and development. This doctrine set the engineering effort on a level of logical parity with the research and development efforts and made it of almost comparable actual size, at least with research. The systems engineer had a multitude of functions, with special emphasis on effective utilization of scientific and technical advances in planning new communications systems. This particular set of ideas, of course, reflected the special needs of telephony. Nevertheless, as an example and a point of departure, it had a wide effect. It seems to be one of the reasons why so esoteric a subject as systems engineering advanced as rapidly as it did.

Relation to operations research

A second major source for systems engineering is operations research, which originated in a recognizable form in Britain during World War II and initially was concerned with the best employment of military equipment. Typical examples included determining the best employment of a given number of bombers, the best way of arranging convoys against submarine attack, and the best way of using interceptors against a bombing attack. Operations research was effective in such cases and has flourished ever since in both civilian and military contexts.

There exists a clear distinction between operations research and systems engineering. Because operations research is concerned with the best employment of existing equipment, technological uncertainties do not arise. Systems engineering, on the other hand, is normally concerned with the planning of new equipment and such uncertainties may be important. In practice, nevertheless,

systems engineering and operations research have a good deal in common. In particular, they share many of the same analytic techniques. This results in large part from the fact that a systems engineer is likely to evaluate the effectiveness of a tentative design by the same methods an operations research man would use with actual hardware.

Another reason for overlap is the fact that the distinction between new and existing equipment is not quite clear-cut. Newness in equipment is a relative matter. If the new equipment is sufficiently well based on existing design techniques and seems to involve few enough technical uncertainties, the issue becomes unimportant. The question is one of degree and, to an extent, of judgment.

Most of the present character of systems engineering derives historically from the early 1950s. There were some noteworthy events in the years just after the war, including, for example, the discovery of linear programming in 1947 and the founding of various organizations for continued development of the field in the late 1940s. On the whole, however, this was a period of consolidation of earlier advances. Thus, in the communications field, the principal systems were some long-distance transmission systems that had been initiated before the war and had been interrupted by war activities.

In the 1950s the pace of growth accelerated appreciably. The first general textbook on systems engineering appeared in 1957 and was followed by a number of other works that treated both industrial and military applications. These publications proved sufficient to establish systems engineering as an accepted academic discipline, and courses in it are now taught in many universities throughout the developed countries of the world. Professional societies and journals exist in France, Great Britain, India, Japan, West Germany, and the United States.

The development of systems engineering after 1950 stemmed, in large part, from the impact of great advances in adjacent fields, notably communications and electronics. An automatic control system is a good example. A control system has the prime characteristic that the components interact extensively and that the system as a whole has certain properties—*e.g.*, stability—that cannot be said to adhere to any individual component. Thus control systems furnished convenient textbook examples for systems engineering.

The development of information theory as a basic starting point for communications engineering, in the years just after World War II, was also influential in shaping the evolution of systems engineering. The various subsystems in many complete systems were found to be held together by what were, in effect, communication channels. Thus ideas of information transfer from one part of the system to another proved useful in understanding the operation of the structure as a whole.

Use of the computer

Systems engineering also profited from the advent of computers (*q.v.*), which affected the field in two principal ways. In the first place, they were a powerful tool for analyzing complex systems by means of extensive calculations or direct simulation. In the second place, they could be used to digest large amounts of data or as actual constituents of complex systems, especially those concerned largely with information transmission. This opened up the possibility of processing information as well as simply transmitting it in such systems.

The impact of military weapons problems on systems engineering began soon after World War II. A landmark date was 1945, when development of Nike Ajax, a United States air-defense missile system, was initiated.

In 1945, available rocket propulsion seemed barely sufficient to give the missile a satisfactory tactical range. It was discovered that achievable range depended on several parameters, such as the weight and size of the warhead, fineness of the missile's aerodynamic design, degree of manoeuvrability provided by the control system, the shape of the trajectory, and the average speed along it. Thus an effective systems-engineering effort could be mounted, in which a variety of combinations of the missile's properties could be explored, with the objective of achieving the best balance between range and other tactical characteristics.

Control and feedback questions were also important aspects of the overall systems problem. The whole system was in fact a gigantic feedback loop because the missile was controlled by orders sent it from a ground computer, and the computer input included information on what the tracking radar observed the missile to be doing. Thus there was a closed feedback loop from missile to computer and back to the missile again. There were in addition such subsidiary feedback loops as that through the autopilot controlling the attitude of the missile, and the dynamical response of the system was further affected by the need to process the radar signals to remove radar "jitter." The analysis of such elaborate dynamical systems involving interlaced feedback paths has become an important special part of the general systems area.

In the 1950s and 1960s systems engineering also grew in other directions, largely as a result of weapons systems projects associated with the Cold War. Thus the Ajax study was concerned with the dynamics of a single isolated missile. On the other hand, the defense systems that grew up in the 1950s involved the coordinated operation of a large number of missiles, guns, interceptors, and radars scattered over a considerable area. These were all held together by a large digital computer, which thus became the central element of the system. The SAGE system in the United States (for semiautomatic ground environment) is a good example.

During the same years the systems approach also became increasingly identified with management functions. Thus the phrase "systems engineering and technical direction" came into use to describe the role of a systems engineer responsible both for the initial planning of a project and its subsequent management. So-called planning, programming, and budgeting techniques (PPB) were developed to provide similar combinations of systems engineering and financial management.

In nonmilitary fields, systems engineering has developed along similar though more modest lines. Early applications were likely to stress feedback control systems in large-scale automated production facilities, such as steel rolling mills and petroleum refineries. Later applications stressed computer-based management information and control systems somewhat like those that had earlier been developed for air defense. In recent years the systems approach has occasionally been applied to much larger civilian enterprises, such as the planning of new cities.

SYSTEMS ENGINEERING TECHNIQUES, TOOLS, AND PROCEDURES

Techniques. If a system is both large and complex in the sense in which we have defined these terms, it may be difficult to find out how it works. In a substantive sense, a large part of the content of systems engineering consists of techniques for the investigation of such relatively complex situations.

Modelling and optimization. Perhaps the most fundamental technique is the flow box diagram, a graphical display composed of boxes representing individual components or subsystems of the complete system, plus arrows from box to box to show how the subsystems interact. Though such a representation is very useful in an initial study, it is, of course, essentially qualitative. A more effective approach in the long run is construction of a so-called mathematical model, which consists of a set of equations, or sometimes simply of tables and curves, describing the interactions in quantitative terms. It is not necessary for the mathematical model to be exact, as long as it serves its purpose. It frequently consists of piecewise linear approximations to basically nonlinear situations (*i.e.*, a series of short straight lines that roughly approximate a curve). After the model has been constructed and checked, a number of mathematical techniques can be employed (including straightforward computing) to find out what it says about the actual operation of the system. Often these calculations will have a probabilistic or statistical flavour.

When the components or subsystems interact signifi-

cantly, it may be possible to achieve essentially the same final level of performance in many different ways. Limited performance by one subsystem may be offset by superior performance somewhere else. These optimization studies, called trade-offs by people in the field, are important in suggesting how to achieve a given result in the most economical manner. They are, however, at least equally valuable in suggesting whether or not the proposed result is in fact a reasonable goal to aim for. It may be found, for example, that a modest reduction in performance will permit radical savings in overall cost or, conversely, that the postulated equipment is capable of much better performance than is asked of it, at only nominally greater expense. (It may also turn out that the equipment can supply useful functions not originally contemplated. Computing systems, for example, can usually perform extra chores of record keeping at little increased cost.) For all of these reasons, studies of such variables are an important part of systems engineering, both in the early exploratory phases of the project and in the final design.

Optimization studies

Identifying objectives. The formulation of suitable objectives for the final system is so important a part of the systems-engineering process that it deserves special attention. It is, of course, always possible to state the general objectives of a system in vague or perfectionist terms. A sufficiently clear, precise, and comprehensive statement to serve as a basis for engineering studies, however, is another matter. Unless the situation has been well explored in the past, the real choices are not likely to be obvious when the work begins. Thus, the first task of the systems engineer is to develop as clear a formulation of objectives as possible. This will usually involve computations and consultation with others interested in the system. Because the final statement must reflect value judgments as well as purely technical considerations, the systems engineer does not try to do this thinking entirely by himself but attempts to serve as a working focus and catalyst. Although issues of this sort naturally present themselves with particular force near the beginning of a systems study, they may recur in subsequent steps. The question of objectives is never really out of the systems engineer's mind.

The principal reason why a satisfactory statement of objectives may present such a problem is simply that most systems have multiple objectives, often in conflict with one another. In airplane design, for example, there are a multitude of desirable characteristics, such as range, speed, payload, and safety to be maximized, as well as undesirable characteristics, such as noise generation and air pollution, to be minimized. Because the same design cannot do the best job in all of these directions, a compromise achieving the most desirable overall performance is required. The most attractive compromise, which may require both study and ingenuity, is not likely to be found at all until some hard thinking has been done about what characteristics are really needed.

Especially difficult problems in defining objectives may arise when an existing technology is transplanted to some new disciplinary area. An example might be the application of electronics, perhaps computer techniques, to medicine and education. It seldom happens in such cases that the best system is based on a simple one-for-one substitution, such as direct replacement of the teacher by electronics. It is much more likely that the most effective plan will turn out to be a rather complicated mixture of the old and the new. This conclusion, however, is likely to raise basic issues about the actual objectives of the new system, issues made no simpler by the interdisciplinary nature of the situation.

To make adequate comparisons between competing objectives, a logical frame of reference, broad enough to include both, is needed. Thus, the systems engineer may study many situations in the framework of more than one system or a whole hierarchy of systems of steadily increasing generality. In the airplane example, the airplane itself is a possible system, as are the group of planes owned by one airline, the total number of airplanes in a particular country, and all that nation's transportation facilities. Though the simplest system—the airplane itself—is a satisfactory reference for specific design problems, a more general framework may be needed to approach broader problems. Thus the individual airplane designer may seek to ameliorate air-traffic congestion by improving airplane takeoff and landing characteristics, permitting better utilization of existing airports. The airlines in turn may suggest construction of more and better airports. From the point of view of the transportation system as a whole, the best step might be to invest more money in high-speed rail facilities, to carry part of the air-traffic load. In systems engineering, the error of studying the problem within too narrow a framework is called the error of suboptimization.

Work at different levels of generality

The stress on systems objectives has one further consequence worth mentioning; *i.e.*, that systems engineering is likely to be strongly user-oriented. This results naturally enough from the fact that systems objectives usually relate to overall performance, which is what the final user is interested in. The identity of technical interest between the systems engineer and the final user is usually marked; systems engineering is likely to give special consideration to such qualities as reliability, ease of maintenance, and convenience in operation. Moreover, the final step of a systems-engineering project is typically an evaluation phase that attempts to find out how well the system works in the hands of the user.

Design considerations and examples. Chemical-processing plants usually consist of many different items of equipment. A typical plant, for example, might include distillation columns, heat exchangers, absorption columns, and chemical reactors. These feed each other; in addition, partially processed material may be recycled. Thus, the different pieces of equipment interact extensively, and many opportunities exist for improving efficiency by rational compromise. An inefficient reactor, for example, can be balanced against high-efficiency absorption and distillation columns. An adequate systems-engineering study takes account of still other considerations, which go beyond the individual physical plant. Capital demands must be balanced against operating costs. Moreover, the plant under study may be one of a complex of plants at the same site, dependent on common feedstocks and common utilities, and this in turn may be only one of several complexes at different sites, manufacturing the same or related products. Thus, the best design for an isolated plant may by no means be the best when all of these other facilities are taken into consideration. The idea of a hierarchy of systems applies here because it is only by going to a sufficiently general system, perhaps the operations of the company as a whole, that one can expect to deal satisfactorily with all of these considerations.

Another example is furnished by the design of the commercial transport plane mentioned above. In such a design, the aerodynamic lift, the drag of fuselage and wings, the control apparatus, the propulsion system, and such auxiliary hardware as the landing gear all interact substantially. One cannot be disturbed without affecting the others. Thus, if designers make the fuselage fatter and the wings smaller in an effort to carry more payload at the same or higher speeds, a new control system might be needed because of the changes produced in the overall mechanical and aerodynamic characteristics of the vehicle. Stronger and heavier landing gear might be needed to withstand higher landing speeds. Almost surely, the new design would call for larger engines and fuel tanks to compensate for greater aerodynamic drag. Thus the designers would have lost ground in some respects and gained in others. The new plane might be more useful for short flights when not much fuel must be carried but less useful for long ones. Obviously, the system objective—the kind of airplane actually wanted—must control the direction of any such study.

The study becomes more interesting if a possible advance in basic technology is considered, such as an improvement in propulsion or aerodynamics, and it is desired to determine how it might best be applied in a new airplane design. The central systems-engineering question

then would probably encompass the relation between the available new plane characteristics and the needs of the existing air-transportation system. Clearly, such an investigation can be made only by going to one of the upper levels in the systems hierarchy.

Finally, to operate the new airplane successfully, a whole series of supporting functions may be required, including routine checkout, maintenance, and spare parts supply, in addition to functions directly involved in the plane's flight. Though, under normal circumstances, these might readily be handled by the existing operating staff, it is part of the user orientation of the systems approach that the systems engineer is expected to anticipate any new requirements here and make sure they are properly planned for.

There exist entire areas of technology that, from a present-day point of view, have an obvious systems flavour. Examples are transportation, electric-power supply, and communications. Communications was the area in which systems engineering first developed. Telephone engineering exhibits most of the characteristics that have been associated with systems problems. A national or international telephone network is inherently large and complex. Moreover, there are many ways in which its function might be accomplished. In addition, telephone traffic—the number of connections that must be set up simultaneously and their geographical dispersion—is essentially random in character, so that a statistical or probabilistic approach is mandatory. Furthermore, man–machine relations are particularly conspicuous in telephony because the customer participates in every call. Finally, the field is abnormally responsive to advances in electronics, so that an orderly approach toward the utilization of new technology by means of systems engineering is particularly rewarding.

Systems engineering in telephony

Tools. The most obvious aspect of systems-engineering tools is their great diversity. A leading text in the field states that "there is virtually no scientific discipline which may not be used in the design of some large-scale system" and singles out probability, mathematical statistics, computing, system logic, queuing theory, game theory, linear programming, cybernetics, group dynamics, simulation, information theory, servomechanism theory, and human engineering. To this list might also be added decision theory, nonlinear programming, some elements of econometrics, and communications theory as related to random processes.

In spite of this diversity, many of the tools of systems theory can be grouped under a few major headings. The analytic problems associated with optimization, for example, are a recurrent theme. Probability and statistics are also major areas that carry with them numerous more specialized topics, such as queuing theory and much of communications theory. Finally, computing is a major field for the systems engineer. If all else fails, he can still mount an attack by direct calculation or simulation.

These fields are all essentially mathematical in nature. The systems engineer may need knowledge and skills of other sorts as well. The mathematical model and associated objective functions that conventionally begin a systems analysis, for example, are only satisfactory to the extent that they adequately represent the real physical situation. The adequacy of a mathematical model in this sense, however, is a matter of physical or engineering rather than mathematical judgment. In some circumstances, the systems engineer may also need to know something about experimental procedures in general and, in particular, about ways of maximizing the amount of information from a given testing program. This is particularly likely to happen in urgent high-risk projects, like weapons-system programs, in which intermediate testing failures are bound to occur, and the systems engineer, as part of his overall responsibility, must decide what to do next. Even in simpler cases, in which the project should close in a final test and evaluation phase, the systems engineer is responsible for ensuring that this work is adequately carried out. A closely related question is the monitoring of testing procedures for routine quality-control purposes. This also is logically part of the systems engineer's duties, a reflection of his basic user orientation. When reliability is very important, as in space programs, it may be a major responsibility.

The systems engineer in testing programs

When the systems engineer's job is defined as including significant management responsibility, some acquaintance with modern management techniques is an obvious requirement. The techniques of particular importance are those that bear most directly on cost figuring and scheduling technological developments.

Finally, systems engineering is used in new situations that may involve the application of new discoveries in science or technology to existing technical areas or the application of known science or technology in new contexts. In either case the systems engineer obviously needs considerable substantive knowledge of the fields involved if he is to make reliable plans.

It is apparent that no single person is likely to meet all of these qualifications. Thus, systems engineering on any significant scale almost invariably involves a team approach.

Procedures. Not all systems-engineering projects follow exactly the same course. Too much depends on the nature of the problem and of the specific organizational setting in which the work is done. Nevertheless, if such variations are ignored, a sort of general pattern of procedures for systems-engineering studies can be perceived. Most authorities recognize a sequence of about five major steps, including systems studies, exploratory planning, development planning, studies during development, and current engineering (individual nomenclatures differ somewhat).

In a broad sense the activities reflected in the early phases (systems planning; exploratory planning) are obvious. The starting point is a recognized need or requirement and the end point a preliminary design of a system to satisfy that need. Mathematical modelling and so-called trade-off studies are used to find efficient proportions for such a system.

When the situation is sufficiently simple, the systems engineer's task may be almost as straightforward as this suggests. In other circumstances, however, the job may be much more difficult. Five points requiring special attention follow.

Exploration of objectives. It is particularly important that the objectives of the system be thoroughly explored and carefully stated. An offhand statement of a requirement seldom is complete or quantitative enough to serve as a basis for design. The desirability of going to a higher level in the systems hierarchy may need to be evaluated. Questions of economic constraints, even if they cannot be fully explored in the preliminary design, are introduced at this time. Finally, the preliminary study must take adequate account of the conditions to be met by the system in actual use. Many systems, such as telephone, highway, and air-defense systems, must be evaluated ultimately by the amount of traffic they can carry. It would not be sufficient, then, to stop the preliminary design with a demonstration in principle, by showing that the proper sequence of operations can be realized for a single telephone call or a single vehicle or enemy intruder. The traffic problem must be met head-on. Similarly, the preliminary study should deal explicitly with such questions as the extent to which the system should be expected to keep operating under conditions of damage or other departures from nominal operating conditions.

Systems involving traffic

Consideration of other possible systems. In many situations there may be more than one possible systems answer to a given need. The systems engineer tries to avoid settling prematurely on one design without giving other possibilities adequate attention, particularly when dealing with systems that depend on radical breaks with past technology. If radical new systems work at all, they are likely to lead to much greater advances than slight improvements of an old system.

Data collection. The systems-engineering team is likely to find that it needs data it does not have. Thus data collection begins as soon as these needs can be foreseen. It is usually more economical in the long run to make the effort ambitious enough to provide a data base through-

out the life of the project and not just in the preliminary phases. As a closely related issue, physical information may be lacking on components or phenomena that may become important in a proposed system. The systems engineer attempts to foresee such needs in order to initiate a suitable experimental program as quickly as possible.

Organization of the team. The internal organization of the systems-engineering effort may also raise unexpected problems. Because the group must cover so much ground, it is natural to make it fairly large, to include all of the necessary kinds of expertise. On the other hand, if the group becomes too large (perhaps seven to ten members), it begins to lose internal coherence. The members tend to pursue their individual interests and may become overly active partisans of particular systems concepts. The close-knit integration of the parts with the whole, on which the whole systems idea depends, begins to fail. Later on, as the system develops, it will naturally begin to subdivide into somewhat independent pieces, and larger numbers can be accommodated.

These phases of the project culminate in the generation by the systems-engineering group of a preliminary design for the proposed system. The preliminary design is supposed to be a feasibility demonstration, to show at least one way to realize the proposed system physically. It is also supposed to show that the system will be feasible economically in the sense that it will not be impossibly expensive.

Review and development. With the preliminary design in hand, the appropriate authority decides whether or not to proceed with the project. If the decision is made to go ahead, responsibility begins to shift from the systems engineers to the development engineers, until the development engineers soon carry the principal work load.

The first step in this shift is a detailed review of the ground already covered in the preliminary design, with the objectives of establishing a detailed blueprint for the development and of providing enough information for satisfactory forecasts of schedules and expenses. In the course of this review, the development engineers may find improved or cheaper ways of accomplishing some of the functions in the preliminary design. If the systems engineer has done his planning well, it will generally be possible to adjust the system to take significant advantage of these possibilities.

After this detailed review, actual development may proceed directly to the final goal, or, depending on circumstances, may be interrupted by the construction of development test models and prototypes. In either case, the development engineer rather than the systems engineer is the central figure in the effort. The systems engineer, however, is responsible for seeing that the original aim of the system is maintained and that intermediate test results are consistent with the overall plan. These stages in the effort correspond approximately to the phases labelled "development planning," "studies during development."

Production and operations

After the development phase comes production and operations. This corresponds approximately to the item, "current engineering." Production and operating people dominate, although the systems engineer also has important responsibilities. He devises tests to find out to what extent the final product does indeed realize the initial systems objectives. Occasionally, readjustments in the design may be necessary. It usually turns out that the apparatus in the field is not used in exactly the way the designer intended. The systems engineer must be alert for such situations and their causes. This becomes particularly important information if there is likely to be a successor design sometime in the future.

APPLICATIONS OF SYSTEMS ENGINEERING

The procedures just described do not distinguish between relatively new technical situations and somewhat older ones. Many useful systems are, in effect, modifications of previous designs. The proportions of the subsystems may be changed, but no substantial function has been added or left out. Chemical-processing plants and information systems, for example, are likely to be of this sort. The basic task of the systems engineer in such a situation is relatively straightforward; it is essentially a matter of reoptimizing the existing design to meet the new conditions.

In other circumstances, however, the basic systems concept represents a more radical break with the past. The new concept may involve the introduction of new functions or the realization of old functions in new ways. On the other hand, it may merely involve a radical change in system parameters (constants). Many complex systems are sensitive in one way or another to changes in scale. A systems concept that seems awkward and uneconomic for certain ranges of basic parameters may be quite inviting for others. Sometimes it is enough just to understand clearly what physical factors are holding the system back. Formal optimization procedures are normally better suited to small, continuous variations in linearized or nearly linearized systems (systems the operation of which can be graphically represented as a straight line). In any of these cases there is a key idea or a key technical decision—to follow one given technical path and not some other—that needs to be isolated and thoroughly understood, because it has a decisive effect on the final system.

The development of radically new systems. Radically new systems concepts are like inventions in ordinary engineering. Usually offering a substantial advance in overall performance, more than would be expected from a modest reproportioning of a known system, these clearly deserve special attention. On the other hand, in many cases it is impossible to predict accurately in advance of the development just what performance may be achievable in one or more of the critical elements of the new system. This leaves the systems engineer with a special problem in planning, which is usually addressed by establishing a minimum acceptable level of performance for the critical elements, with the rest of the system so arranged that whatever is realized beyond this level will appear as growth potential in the overall capabilities of the system. Thus definitive optimization studies may be postponed until the system is better understood.

Microwave-radio relay system. An illustration of this is furnished by the development of a microwave-radio relay system begun in the early 1940s and finished after the war. The possibility of using radio instead of wire lines for long-distance commercial communications had long been recognized. Beginning at quite low frequencies with Marconi, the transmission capability had crept up the radio spectrum to slightly beyond the high-frequency band, but the band width was still very small and most of the frequency spectrum in this range had to be conserved for vital services. Near the end of the 1930s, it was suddenly realized that many difficulties could be overcome by taking a large jump upward in the frequency spectrum. The system invention, then, consisted of a frequency jump of the order of 100 to 1 or more, from the high-frequency band (3–30 megahertz) to squarely in the middle of the so-called microwave range (about 4,000 megahertz). In this range, adequate frequency space was available and external interference comparatively low. Furthermore, at these wavelengths moderate-size antenas could be made highly directive, thus leading to efficient transmission links without mutual interference, even when they were physically close together and operated on the same wavelengths.

A systems analysis of radio waves

A preliminary systems analysis revealed that the microwave system itself was efficient only for rather broad transmission bands. Moreover, because a large part of the cost of the final system was for towers, access roads, power supplies, and other equipment, it was advantageous to provide numerous separate links at each site. Thus, everything pointed to a system of very high capacity, and high capacity became a necessary part of the total system concept. The advent of network television, which could readily be foreseen, provided an obvious justification for the provision of such broad bands and resultant high capacity. Growth potential was easy to achieve.

Nike Ajax missile system. The Nike Ajax missile system provides another example of the application of a radically new systems concept. The simple realization that the technology was available to provide a missile that could outmanoeuvre an enemy bomber taking evasive

action was perhaps the systems invention in this case. (Guided missiles had been thought of before, but only for use against simpler targets than a rapidly manoeuvring airplane.) In a more limited sense, however, the key idea in the overall systems concept was probably the decision to use a command-guidance system, as opposed, for example, to a homing system.

In the command system, both the radars used for tracking the aircraft and the intercepting missile and the computer that calculates how the missile should change course are on the ground. Such a system requires a minimum of control apparatus in the missile. It also allows the missile to follow computer-determined paths that are aerodynamically favourable. This was an especially important consideration at the time in maximizing the range achievable with available propulsion systems. It also allows, through the computer, maximum flexibility in dealing with evasive action by the target. On the other hand, adequate accuracy from the ground-tracking system becomes increasingly difficult in a command system as the range is increased, whereas a homing system is not so limited. Thus the adoption of the command system implied a belief that the ground radars would be accurate enough to provide satisfactory interception even at the limits of the expected field of fire. As the development turned out, tracking accuracies were more than adequate for the purpose, and the surplus provided growth potential toward still longer ranges and higher probabilities of interception. (In other circumstances, of course, a different choice might have been better.)

Long-term systems development. Thus far the description of systems engineering may seem to suggest that systems-engineering efforts are essentially episodic and that to the individual systems engineer successive studies may have no special connection with one another. In many situations, however, there are important elements of continuity in what the systems engineer actually does. He is likely, in fact, to work on a series of similar problems as part of a long-term effort. This happens frequently, for example, in the weapons-system field or in the design of refineries or other chemical-processing plants. In the case of telephone engineering, systems-engineering groups have been set up formally as permanent parts of the overall organizational structure, each group having cognizance over some wide area of telephone technology. Thus continuity in this case may extend over many years.

Permanent systems-engineering groups

Such elements of continuity have a major impact on the systems engineer's job. He can work more efficiently and expeditiously on each new task if he can draw upon relevant experience. More to the point, however, the efficient solution of any particular problem, using familiar tools, begins to be something that can be taken more or less for granted. Attention almost inevitably begins to centre on the relation these solutions have to general technical progress in the field. Systems engineering thus tends to shift away from operations research and toward long-range technical and economic planning.

Telephone engineering makes a particularly good example of this because of the importance of long-range forecasting for the telephone plant. One of the important responsibilities of the systems engineer in this field, for example, involves establishment of performance standards for new items in the plant. In most fields, the systems engineer's role in this respect is comparatively modest because performance standards represent value judgments established by others. In telephony, the systems aspects are more conspicuous because, at any given instant, the plant is a composite of new and old items, designed at different times in different ways to meet a variety of assumed operating conditions. The pieces must all work together. Any new apparatus or system must be compatible with what already exists. In addition, however, if an orderly evolution of the plant is to continue, the new apparatus must also be suited to the traffic and service needs of the future. Thus the systems engineer's responsibility, to provide a sort of long-term doctrine of performance standards, becomes very important.

Somewhat the same considerations affect the economics of new telephone systems. Here again a carefully developed forecast is essential. On the one hand, there are usually substantial economies of scale in communication apparatus. Thus it pays to take big steps in installing new equipment. On the other hand, capital charges on underutilized equipment are likely to be excessive. Because the successive steps of design, production, and installation of a complicated system may take years, and because meanwhile both technology and service demands may change materially, the situation may increase in complexity, and the systems engineer may find that the choice of the most inviting system is not an easy one to make. Indeed, the best choice may involve a calculated risk. He is responsible for judging the degree to which new possibilities should be exploited at any time. The principal result of assigning the telephone-systems engineer a continuing organizational position and function is to put him into much more intimate contact with the technical frontier in communications. His prime responsibility becomes that of monitoring the technical frontier to see what can be put to use in new operating systems.

Major technical advances may, of course, require many years between the original discovery or conception and the time when a practical design is feasible. The systems engineer keeps in touch with the work throughout this period. He can exercise a valuable if indirect influence on the investigation simply by noticing weaknesses and errors that need correction for a successful project. It frequently happens that a program of systematic measurements is called for before a new systems conception can be implemented, even when the basic conception is well established.

Competitive approaches to problems

In many cases, technology suggests two or more competitive approaches to the same problem. If a continuing systems-engineering organization exists, there is generally no need to make a premature choice between them. Rather, both lines can be followed until it is clear that one is superior or, perhaps, that each has its place. In telephony, for example, this was the case in the long-continued rivalry between microwave transmission systems and those based on transmissions by a coaxial cable.

New fields of application: future possibilities. Thus far, systems engineering has been dealt with in relation to two principal fields of application. One field is industry, in which the prospects of a further expansion of systems engineering appear bright. Existing applications furnish many good models, and it seems likely that such extensions can take place without raising many unusual problems. The other major field of application has been in military and space systems, and this may have been the principal force in shaping the systems-engineering field. Although techniques learned here are still valuable, interest in applications of systems ideas to military and space problems seems likely to decline as activities in both fields taper off somewhat. In their place may appear new applications of systems ideas in the nonmilitary side of government. There appear to be many such possibilities in the realm of worldwide basic social and economic problems. On the other hand, systems engineering as practiced in other contexts does not automatically transfer easily to this new environment. General interest in the subject dates, however, only from the rather late 1960s, so that the field is incompletely explored.

Systems approach in government

In one experiment in the conversion of military systems-engineering techniques, a U.S. state government contracted with four large aerospace companies (each of which had a substantial capability in systems engineering) to study the following four topics: (1) a statewide information-handling system, including a plan for implementation, (2) a program for the prevention and control of crime and delinquency, (3) a waste-management problem, and (4) a systems approach to basic transportation problems.

None of the four studies led to proposals that seemed attractive enough to be implemented by the state, and, in this respect, the experiment was a mild disappointment, although in view of the wide scope of the problems attacked and the limited effort called for by the study contracts, the result is not surprising. On the other hand, the

experiment was useful in advertising the possibilities of systems analysis as applied to civil problems and in illuminating difficulties that may be encountered in making such applications. The experiment stimulated interest in the civil uses of systems methods both inside and outside the United States. A survey made in 1967 listed more than a hundred studies then current. By the early 1970s there were several times as many.

The potential applications of the systems approach to governmental activities are so numerous and so varied, in both the developed and developing worlds, that an exhaustive catalog would be out of the question. Nevertheless, it may be worthwhile to list a few of the most conspicuous possibilities. The most obvious class is made up of massive engineering attacks on very broad socio-economic problems. These are the situations that seem to have most in common with the applications of systems methods in the weapons sphere. New transportation systems, comprehensive attacks on pollution, and radical reconstruction of urban areas belong here. A concrete issue is the problem of power-plant location, an urgent question in many advanced and developing countries. The systems overtones are obvious. Generating stations are customarily interconnected so that a new plant has an impact on the availability of power over a considerable region, and, of course, the effects of thermal and atmospheric pollution from a given plant may also be widespread.

Other applications of systems analysis in the social sphere tend generally to be smaller and more easily treated. One class consists of the extension of military budgeting and methods of financial control to the nonmilitary world. Another application has been the use of systems analysis to support the technical side of foreign-aid programs. Other fields include the possible application of specific items of new technology in such areas as crime detection, fire fighting, and traffic control. Still other current studies involve specific aspects of such subjects as housing and other types of building construction. Such studies attempt to be useful rather than broad or necessarily definitive for all time to come.

The applications of systems analysis in civil government obviously still have far to go before their potentialities are exhausted. On the other hand, there are many reasons why these potentialities can be realized only slowly, if at all. Some of them are related to the inherent difficulty of the problems presented—the wide range of both technical and social considerations that may enter certain decisions, for example. Others reflect some of the common characteristics of governmental structure, the necessary bureaucratization of functions, for example, or the frequent problem of overlapping jurisdiction. Still other problems reflect the fact that existing systems analysts are trained preponderantly in the physical sciences and engineering and thus may not be well matched to the socio-economic issues they are likely to confront, though most systems-analysis groups working in socio-economic questions try to balance their strength by adding appropriate missing skills. The most common problem, however, is probably simply the need to build up an adequate basis for mutual cooperation between systems analysts and government.

As such an evolution proceeds, there may be an increasing tendency for the individual systems analyst to become identified with the substantive area in which he works and to lose his special relation to systems analysis as a distinctive field. Thus, it may be the ultimate fate of systems analysis to disappear as a separate field and instead become an important constituent of the planning function required in many parts of modern society.

BIBLIOGRAPHY. The systems engineering field is so broad and so diffuse that one should not expect to find a complete treatment in any single work. A well-rounded picture requires several sources.

Because systems engineering and operations research overlap extensively, one possible source is simply a good account of operations research. The classic starting point here is PHILIP M. MORSE and GEORGE E. KIMBALL, *Methods of Operations Research*, rev. ed. (1951). The introductory discussion in this book is still useful for orientation, but the specific examples, dating mostly from World War II, are too limited to be of much value now. The book *Operations Research for Management*, vol. 1 by JOSEPH F. MCCLOSKEY and FLORENCE N. TREFETHEN (1954), and vol. 2 by MCCLOSKEY and J.M. COPPINGER (1956), on the other hand, give a broad range of examples, many of which are still of interest. Because, in addition, most of the examples are developed with a minimum of mathematics, the book is well suited to the general reader interested primarily in a broad understanding of the field.

Later books have tended to place more stress on mathematical techniques. Thus, *Introduction to Operations Research* by CHARLES WEST CHURCHMAN, RUSSELL L. ACKOFF, and E. LEONARD ARNOFF (1957), gives a fair survey of the field but is more useful for its description of operations research methodology. Recent books such as FREDERICK STANTON HILLIER and GERALD J. LIEBERMAN, *Introduction to Operations Research* (1967), put still more emphasis on methodology (and probabilistic techniques in particular) and are less suited to the general reader.

Among books more directly related to systems engineering, the earliest significant work is HARRY HERBERT GOODE and ROBERT ENGEL MACHOL, *System Engineering* (1957), followed by Machol's *System Engineering Handbook* (1965). These give substantial treatments of both the philosophy and methodology of systems engineering, and the second in particular is a mine of background information. On the other hand, both books are clearly directed primarily at large military systems and are less useful for other applications. A somewhat more recent book that also attempts to cover both philosophical and technical aspects of systems engineering is A.D. HALL, *A Methodology for Systems Engineering* (1962). It takes its inspiration largely from the author's experience with television and similar broad-band circuits. HAROLD CHESTNUT, *Systems Engineering Tools* (1965) and *Systems Engineering Methods* (1967), deal largely with methodology and overlap corresponding books in operations research extensively.

Among shorter works, an article by H.W. BODE, "The Systems Approach," in the NATIONAL ACADEMY OF SCIENCES, *Applied Science and Technological Progress* (1967), may be mentioned. It furnishes a broad discussion of the field along lines similar to those followed in the present treatment. Another article with the same title, by G.M. JENKINS, in the *J. Systems Engng.*, 1:3–49 (1969), furnishes a valuable elementary introduction to systems engineering, especially for cases in which technical uncertainties are not serious. C.J. HITCH, "Sub-Optimization in Operations Problems," *J. Ops. Res. Soc. Am.*, 1:87–99 (1953), is a classic.

(H.W.B.)

Szechwan

Szechwan (Si-chuan in Pin-yin romanization), the second largest province of China, is located in the Upper Yangtze Valley in the southwest part of the country. It covers an area of 219,700 square miles (569,000 square kilometres) and is bordered by the provinces of Kansu and Shensi to the north, Kweichow and Yunnan to the south, the Tibetan Autonomous Region to the west, Tsinghai to the northwest, and Hupeh and Hunan to the east. One of the most populous regions in China, in the early 1970s it had an estimated population of between 70,000,000 and 84,000,000. The name Szechwan means Four Streams and refers to the four main tributaries of the Yangtze River, which flows through the province. The capital, Ch'eng-tu, is located in the centre of the province.

From economic, political, geographical, and historical points of view, the heart and nerve centre of Szechwan is in the eastern basin area, commonly known as the Szechwan, or Red, Basin. Its mild and humid climate, fertile soil, and abundant mineral and forestry resources make it one of the most prosperous and economically self-sufficient regions of China. The area has been seen by some as China in a microcosm and is often viewed as a country within a country. The Chinese call the basin Tien Fu Chih Kuo, which literally means Heaven on Earth. (For an associated physical feature see YANGTZE RIVER; see also CHUNGKING.)

HISTORY

Early history. Apart from the Upper Huang Ho (Yellow River) Valley provinces, Szechwan was the first area of China to be settled by the Chinese, or Han, people. The first organized Chinese migration took place in the 5th century BC.

Szechwan was known as the Shu Pa territory during the Chou dynasty (1122 to 221 BC). During the Ch'in dynasty (221 to 206 BC) the territory was incorporated within the Ch'in Empire and began to assume considerable importance in China's national life. It was at this time that the Tu Chiang-yen irrigation system was built to control the Min Chiang and to water the Ch'eng-tu Plain. In the time of the Three Kingdoms Period (AD 220 to 264) the Szechwan region constituted the Shu Kingdom. From the end of this period until the 10th century, Szechwan was known by various names and administered through various political subdivisions. During the Sung and Southern Sung dynasties (AD 960 to 1279), it was known as Szechwan Lu ("Szechwan Province"). Szechwan was established as a province during the Ch'ing, or Manchu, dynasty (1644 to 1911).

The Chinese Republic. During the early years of the Chinese Republic (1911 to 1930) Szechwan suffered seriously from the feudal warlord system; at one time it was divided into as many as 17 independent military units, and not until 1935 was it unified under the Nationalist Party government. During the Sino-Japanese War of 1937 to 1945, there was a great influx of people and of new ideas from coastal China—a development that resulted in extensive economic development. A large number of factories, or trading posts, were moved from the coastal area into Szechwan, and a number of industrial centres were established, especially in Chungking and at Ch'eng-tu.

Because of its geographic isolation, inaccessibility, extensive area, large population, and virtual economic self-sufficiency, Szechwan has served periodically as a bastion in its own right. Because the area is easily defensible, geography has encouraged political separatism. During the war with Japan, the province served as the seat of the Nationalist government from 1938 to 1946; the Japanese were never able to penetrate the area.

THE NATURAL ENVIRONMENT

Relief. Szechwan is bordered on all sides by lofty highlands. To the north, the Tsinling Shan (Tsinling Mountains) extend from east to west and attain an elevation between 11,000 and 13,000 feet above sea level. The limestone Ta-pa Shan rise to approximately 9,000 feet on the northeast, while the Ta-lou Shan, a lower and less continuous mountain range with an average elevation of 5,000 to 7,000 feet, border the south. To the west, the Ta-hsüeh Shan—the Great Snowy Mountains or Azure Mountains of the Tibetan borderland—rise to an average elevation of 14,500 feet; to the east, the rugged Wu Shan, rising to about 6,500 feet, contain the spectacular Yangtze Gorges.

The Szechwan Basin

In general, the relief of the eastern region of the province is in sharp contrast to that of the west. An extensive depression—known as either the Szechwan or the Red Basin—and its peripheral highlands predominate in the east; the land slopes toward the centre of the basin from all directions. This basin was a gulf of the China Sea in the later Paleozoic Era from 225,000,000 to 570,000,000 years ago; most of it is underlain by soft sandstones and shales that range in colour from red to purple. Toward the west there is a generally horizontal rock structure, while to the east there is considerable discontinuity and deformation.

Within the basin, the surface is extremely uneven and gives a general appearance of badland topography. Numerous low, isolated, and rolling hills are interspersed with well-defined high ridges, flood plains, valley flats, and small, local basins. The most impressive portion of the basin's surface is the Ch'eng-tu Plain—the only large continuous tract of relatively flat land in the entire province.

The land forms of western Szechwan include a plateau in the north and mountains in the south. The northern area is part of the edge of the Plateau of Tibet, which consists of highlands over 12,000 feet above sea level and higher mountain ranges. There is also an extensive plateau and some swampland. To the south, the transverse mountain belt of eastern Tibet and western Yunnan Province rises to an average of between 9,000 and 10,000 feet above sea level. Trending from north to south is a series of parallel lofty ranges with narrow divides and canyons more than a mile deep. The Minya Konka in the Ta-hsüeh Shan, is the highest peak in the province; it rises to a height of 24,900 feet (7,590 metres).

Drainage. Seen from the air, the principal drainage pattern of the eastern section of the province has the appearance of a leaf with a network of veins. The Yangtze—flowing from east to west—is conspicuous as its midrib, and the main north and south tributaries appear as its branch veins. Especially important are the Chia-ling and Min river systems in the north and the Wu Chiang system in the southeast. The distribution of these veins is primarily concentrated in the upper, or northern, half of the leaf.

The name Szechwan (Four Streams) refers to the four main tributaries of the Yangtze. These are the Min, T'o, Chia-ling, and Fou rivers, which flow from north to south. Most of the major streams flow to the south, cutting steep gorges in the west or widening their valley floors in the soft sediments of the Szechwan Basin; they then empty into the Yangtze before it slices its precipitous gorge, 150 miles due east, through the Wu Shan below Wan-hsien. Within the basin most of the rivers are navigable and are a common means of transportation.

Soils. There are six major soil regions—three in the east and three in the west. In the east, they include the highly fertile, purple-brown forest soils for which the Red Basin is named. This group of soils rapidly absorbs and loses water; one evening's thick fog may cause the entire depth of the surface soil to be thoroughly moistened, while a half day's torrential rain may cause the soil to be completely saturated and washed away. The other eastern soils consist of the noncalcareous alluvium and rice-paddy soils of the Ch'eng-tu Plain and other river valleys and the yellow earths of the highlands and ridges. From the standpoint of agricultural land use, the alluvial soils are the most important group. They are very fertile and are formed mainly from the rich black soils that are washed down from the Tibetan borderlands. The yellow earths are usually gray brown in colour; they are generally less fertile and are agriculturally unimportant.

The three major groups of soils in the west are the degenerated chernozem (dark-coloured soils containing deep, rich humus) soils of the Sung-p'an Grassland, the alluvial soils of the numerous valleys, and the podzolized (leached), gray-brown soils of the mountain slopes.

The problem of soil creep

In Szechwan, a form of soil erosion known as soil creep has developed. On hillsides where the surface slopes are composed of smooth sandstones, the covering soil gradually slides downward under the influence of gravity. In many places the thin surface soils have been completely removed, leaving only bare rocks. When the surface rock is composed of comparatively rougher shales, the soil is less easily moved.

Climate. The eastern basin area and the lower western valleys are sheltered from cold polar air masses by the surrounding mountains. The climate is therefore milder than would be expected and is similar to that of the Yangtze Delta region. There are 350 frost-free days in the east, and the growing season lasts nearly all year round. In the west, the sheltering effect of the mountains is evident from the contrast between the perennially snow-capped peaks and the mild weather prevailing in the valleys beneath them. During the summer, in the month of July, the mean temperature is about 84° F (29° C) at Chungking in the southeast and less than 68° F (20° C) in most parts of the west. During the winter, the mean temperature in the west decreases northward from 54° F (12° C) in Hsi-ch'ang to 18 ° F (−8° C) in Ch'ien-ning.

The eastern rainy season begins in April and reaches its peak during July and August. Annual rainfall and precipitation measures about 40 inches annually. The east is noted for its frequent fogs, its many cloudy days, the relative absence of wind, and the high relative humidity. The extent to which the region is overcast with clouds is reflected in the saying "Szechwan dogs bark when they see the sun."

Precipitation is lower in the west than in the east. The average total of about 20 inches falls mainly during the summer, and there is heavy snowfall in the mountains during the winter.

Vegetation and animal life. There are four major vegetation regions—the pine–cypress–banyon–bamboo association of the basin area, the dense mixed association of coniferous and deciduous trees in the eastern highlands, the grasslands of the northwest, and the dense coniferous forests of the western highlands. Its great altitudinal differences, low latitudinal position, diversified topography, and high rainfall, make the area what has been called a paradise for botanists. Extensive forests grow on the upper slopes, and rich rhododenderons are found at higher elevations; arid vegetation prevails on many canyon floors.

One of the outstanding features of vegetation of Szechwan Province is its division into vertically differentiated zones. Cypress, palm, pine, bamboo, tung-oil, and citrus-fruit trees grow below 2,000 feet, while, between 2,000 and 5,000 feet, there are evergreen forests and oaks. From 5,000 to 8,000 feet, the vegetation is characterized by dense groves of mixed coniferous trees. Between 8,500 and 11,500 feet, there is a sub-Alpine zone of coniferous forest, while above 11,500 feet there are Alpine zones of scrub and meadow up to the snow line, which occurs at 16,000 feet.

One of the unique vegetational features is the presence of the dawn redwood, or *Metasequoia*—a tree previously believed to be extinct.

Two of the most interesting indigenous animal species are the *hsiung-mao* (the bear cat or panda) and the ling yang (a special species of antelope), both of which frequent the highlands of western Szechwan.

POPULATION

Population groups. *Ethnic composition and distribution.* Szechwan Province has one of the most diversified range of ethnic groups in all China. They include the Han, or Chinese, people, the Yi (Lolo), the Tibetans, the Miao, the Hui (Chinese Muslim), and the Ch'iang. Most of the Han people—who comprise about 96 percent of the population—live in the basin region of the east. The Yi reside in the Liang-shan Fi area of the southwest, while the Tibetans are distributed in the plateau region of the west. The Miao people live in the mountains of the southwest near Kweichow and Yunnan provinces. The Hui are concentrated in the Sung-p'an Grassland of the northwest and are also scattered in a number of districts in the east. The Ch'iang are concentrated in the Mao-wen area on both banks of the Min Chiang.

The minority peoples

The majority of the non-Chinese ethnic groups are fiercely independent and have maintained their traditional way of life. In most cases, they practice a mixture of agriculture, animal husbandry, and hunting.

Among the Chinese, there has been an influx of people from various neighbouring provinces, and particularly from Hupeh in the east and Shensi in the north. This immigration was especially intensified in the early part of the 18th century, as a result of the massacre of the Szechwanese by a local warlord. The immigrants brought with them agricultural techniques that are reflected in the heterogeneity of present cultivation patterns.

Linguistic groups. There are three major linguistic groups—the Chinese, who speak southern Mandarin; the Tibeto-Burman group including the Tibetans and the Yi; and the Muslims who speak southern Mandarin but use Turkish or Arabic in their religious services.

Religious affiliations. The Chinese practice a mixture of Confucianism, Buddhism, and Taoism. They do not maintain rigid boundaries in religious belief. The Tibetans follow a form of Buddhism in which the lamas are spiritual authorities. Many people in the northwest profess Islām, while some hill tribes of the southwest may be classified as animistic.

Patterns of human settlement. *Population density.* As one of the most densely populated provinces of China, Szechwan may be compared to the Yangtze Delta and the North China Plain. Given the larger estimated population of 84,000,000, it has an average density of over 717 persons per square mile. The population, however, is unevenly distributed. About 95 percent of the total population is concentrated in the eastern part of the province. The number of persons per square mile of cultivated land varies from about 26,000 persons in the Ch'eng-tu Plain to fewer than 130 persons in the west.

Rural settlement. As in most of China, about 85 percent of the population is rural. There are comparatively few large villages and nucleated hamlets, except for the provincial capitals. In the hilly regions, farmsteads are scattered through generally small and irregular terraced fields. In the Ch'eng-tu Plain, the larger field units are commonly square or oblong in shape, and the farmsteads are surrounded by groves of banyan, cypress, mimosa, palm, or bamboo.

Urban settlement. One of the outstanding features of urban settlement is the concentration of cities on river terraces. Along the Yangtze River from Feng-chieh to I-pin, there are no fewer than 15 urban centres, including the most populous cities and trading towns of the province. Because water transportation is vital, large cities are always found wherever two major streams converge. Chungking is located at the confluence of the Yangtze and the Chia-ling rivers, Lu-chou on the Yangtze and the T'o, I-pin on the Yangtze and the Min, and Lo-shan on the Ta-tu and the Min.

The relation between cities and rivers

The principal disadvantage of these urban sites is that their areas are limited by their locations, so that development is hindered. Because of the fluctuations of high and low water levels in most of the rivers of the basin, the hazards of flood are great. Many houses are perched on understructures of poles and posts.

Chungking, located in the southeast, is the largest city of the province and the most important trading, transportation, industrial, and cultural centre of southwestern China. Ch'eng-tu is located in the centre of the Ch'eng-tu Plain. It is the provincial capital, and Szechwan's second largest city. Other important urban centres in the east are Tzu-kung on the T'o Chiang, I-pin in the southeast, and Nan-ch'ung on the upper Chia-ling Chiang. The chief western cities are Ya-an on the Ching-yi Chiang, K'angting near the Ta-tu Ho, and Hsi-ch'ang on the An-ning Ho.

Most urban settlements give the appearance of being compactly built. Generally, the houses have only one story. There are no yards or sidewalks in front of the houses, which abut the street. The streets are narrow and in most cases not over ten to 12 feet wide; they are paved with limestone slabs. Like many other long established and unplanned urban communities in the Orient, most of the cities have no sharply delineated functional areas.

ADMINISTRATION AND SOCIAL CONDITIONS

Government structure. *Provincial administration.* Szechwan is one of the 21 provinces forming the first-order administrative units of China. In 1955, former Sikiang Province at the edge of the Plateau of Tibet was incorporated into Szechwan Province, doubling its area and making it the second largest province of China. The province is divided into three municipalities (*shih*) of Ch'eng-tu, Chungking, and Tzu-kung; 12 special districts (*chuan-ch'ü*); and three autonomous districts (*tzu-çhih-chou*).

Administrative subdivisions

The autonomous districts are the A-pa Tibetan Autonomous District with its headquarters at Ma-erh-K'ang; the Tibetan Kan-Tzu Tibetan Autonomous District with its capital at K'angting; and the Liang-shan Yi Autonomous District with its capital at Chao-chiao. As a rule, the districts are inhabited by sizable groups of non-Chinese peoples, and they represent little more than a symbolic cultural indulgence of local minorities. The actual control of the units is exercised by the central government at Ch'eng-tu. The ethnic groups, however, enjoy their own mode of life and preserve their cultural traditions with a minimum of interference by the Chinese-controlled provincial government.

Local administration. There are 181 counties (*hsien*) in the province. They are the most important administra-

tive units because the government exercises control over the rural masses through them. Normally, the county headquarters is not only a political centre but also an economic, cultural, transportation, and market focus of the county. It also represents the county's largest congregation of people.

Social conditions. *Education.* The educational system was restructured as a result of the Cultural Revolution of the late 1960s. It consists of preschool classes, or kindergartens; six-year primary schools; and six-year secondary schools. Vocational schools usually include work-study programs, but there are also several full-time vocational schools with two- or three-year courses.

Ch'eng-tu and Chungking are centres of higher education. Institutions include Chungking University, Szechwan University, Che'eng-tu Technical University, Szechwan Agricultural Institute, Southwestern Institute of Law, Ch'eng-tu Chinese Medical College, Szechwan Medical College, and various engineering institutes.

Health. Because of the warm and wet climate of most of the province, respiratory ailments are a major health problem. Despite the rich agricultural production of the region, malnutrition is fairly common, although famine is rare.

Welfare. Because of the severe pressure of the people on the land, the Szechwanese farmer must work extremely hard to eke out a living. The farmers of the Ch'eng-tu Plain are the most prosperous and have the highest standard of living. Rural life is harder in the hills surrounding the basin, and the standard of living is considerably lower in the west, where pastoral activities predominate. In the western mountains, many of the people practice transhumance—*i.e.*, seasonal migration from the lowlands to the highlands in search of pasturage.

THE ECONOMY

Szechwan Province has a varied economy based on agriculture, valuable forestry products, mineral deposits, and diversified branches of industry.

Agriculture. Approximately 85 percent of the population of the province earn their livelihood from agriculture, and most of the provincial exports are agricultural products. Cultivation is characterized by the diversity of crops, intensive land use, the extensive practice of terracing, the use of *tung-shui-t'ien* (or fields that are flooded in winter in order to store water), the cultivation of *tsai-seng-tao* (or "rebirth" rice), and the special methods of soil culture, fertilization, composting, and crop rotation.

Agricultural practices

Agricultural products. Crops range from those of subtropical climates to those of the cool temperate zone. Although Szechwan is generally classified as a rice region, it is also a prominent producer of such crops as corn, sweet potatoes, wheat, rapeseed (the oil seed of the rape plant), kaoliang (drought-resistant sorghum), barley, soybeans, and millet. Tropical fruits—such as litchi (the oval fruit of a soapberry tree) and citrus—grow together with the apples and pears of cool temperate climates.

Szechwan leads all of China in production of rice, corn, sweet potatoes, and rapeseed and ranks second in the production of barley and soybeans. Other principal economic crops include sugarcane, cotton, tobacco, silk, hemp, and tea. Szechwan has a long history of silk production; its hemp and other fiber crops normally occupy second place in the total production of the nation.

Livestock. Szechwan leads the nation in the total number of its cattle and pigs. It is the only region in China in which both water buffalo of South China and oxen of North China are found together. Pig bristles from Szechwan are of international repute and have been an important item of foreign trade for years.

About half of the inhabitants of the west are pastoral. Their animals include cattle, sheep, horses, donkeys, and yaks (long-haired oxen from Tibet). The most important animal products are butter, cheese, fur, and hides.

The "land of 1,000,000 steps"

Terracing and irrigation. The basin area of eastern Szechwan is extensively terraced and is often known as a "land of 1,000,000 steps." The terraces are of varying dimensions. They are commonly long narrow strips of land and frequently have rather steep slopes. They are easy to construct because the bedrock is soft and weathers easily. Even 45 degree slopes have tiny steps of terraced land.

A special landscape feature of the eastern basin is the *tung-shui-t'ien* (literally "winter water-storage field") system. In spite of the favourable climate and the fertile soil, large tracts of terraced fields—estimated to amount to 25 to 50 percent of the total area of rice fields under cultivation—are left fallow during the winter season. They are used for the storage of water that will be needed in the paddy fields the following spring; from the air they give the appearance of a mosaic of broken mirrors.

Irrigation is widely practiced in the terraced fields and numerous methods and devices are employed. Among the most spectacular is the Tu Chiang-yen system of the Ch'eng-tu Plain, which captures the torrential flow of the Min Chiang and guides it through an artificial multiplication of channels into numerous distributaries along the gently graded plain. The river level is kept constant by the removal of excess silt through annual dredging. The system is not only the oldest but also the most successful and easily maintained irrigation system in China. It has freed the Ch'eng-tu Plain from the hazard of floods and droughts and insured the unbroken agricultural prosperity of the basin. Since 1950, the Szechwan provincial government has initiated numerous engineering improvements of the project.

Forestry. Valuable forests are located in the east, on the peripheral highlands which surround the basin area, and on the numerous hills within the basin. Because of its high elevation and plentiful humidity, Szechwan is second only to Manchuria as a lumber region. It still has much of its original forest cover, with a predominance of spruce, fir, and birch trees. The most important products from the forests are tung oil, white wax, various kinds of herbs, and some animal products, including antelope horn and *hsi-hsiang* (a special perfume extracted from deer secretions).

Mining and quarrying. Mineral deposits are abundant and varied. They include both metallic and nonmetallic deposits, such as those of iron, copper, gold, silver, aluminum, salt, coal, petroleum, asbestos, and marble. Among them, salt, coal, and iron were by far the most important in the early 1970s. The production of brine salt is the most extensive mining activity; the province is the only large producer of this commodity in China.

Petroleum and natural gas are often located together and are widely spread throughout the province, especially in the Tzu-kung area. Natural gas has been used for centuries in the production of brine salt. There are also more natural gas deposits being exploited in the Pa-hsien area and elsewhere.

Most coal fields are located in the eastern and southern mountain areas; those of the Hua-yang Shan area are the richest, and production there is the largest. The most important iron deposit is in the Ch'i-chiang area, and those of the western plateau area are of high quality. Placer gold along the Chin-sha Chiang (Gold Sand River) is important. Other valuable minerals include tin, zinc, and sulfur.

Industries. There has been considerable industrial development since the 1950s and Szechwan has become the most industrialized province of southwestern China. The most important industries include iron and copper smelting, the production of machinery and electric power, coal mining, petroleum refining, and the manufacture and processing of chemicals, textiles, and food. Chungking is the most important industrial centre, and other industrial areas include those centred on Ch'eng-tu, Nei-chiang, I-pin, and Lo-shan.

Szechwan is also known for its cottage industries. These include the handweaving of cloth, the production of silk products, silversmithing and coppersmithing, and embroidery.

TRANSPORTATION

Overall problems. Among the problems facing the province, none is more important and more acute than

the problem of transportation. For centuries, travel into the province has been extremely difficult; the main entrances to it were the dangerous Yangtze Gorges in the east, the treacherous Chang Tao road across the mountains in the north, and the deep canyons and swift currents of the Ta-tu Ho and Chin-sha Chiang in the west. Since the 1950s, both the central government and the provincial government have made great efforts to improve transportation. Numerous dangerous shoals along the Yangtze Gorges routes have been destroyed, a railway has been built across the northern mountains, and steel bridges have been constructed over the Ta-tu Ho and other rivers in the west.

Component systems. *Inland waterways.* Water routes are the most important means of transportation. There are about 300 streams in the province, with a total length of more than 8,000 miles. The mighty Yangtze River, the most significant, traverses the entire width of the basin from the southwest to the northeast. It is the spinal cord of the river transportation system; steamboats of 1,000-ton capacity can negotiate its waters between the gorge section and Chungking during the high-water season. Smaller steamers ply the river between Chungking and I-pin throughout the year and travel on various tributaries, such as the Chia-ling, the Min, and the T'o.

In the west, water transport is difficult and limited except for the lower reaches of the An-ning Ho and the Ta-tu Ho.

Railways. Railways are important for the transport of bulky products. Until the 1950s, only a small railroad that ran for 47 miles between Hsiu-shan and Pai-miao-tzu near Chungking served the province. Since then, facilities have been greatly improved. The Ch'eng-tu–Pao-chi railroad—which crossed the lofty Tsinling Shan for the first time in Chinese history—links with the principal east–west Lunghai Railway. Szechwan is thus connected to both northwest and coastal China. The Ch'eng-tu–Chungking railroad, completed in the 1950s, links the rich Ch'eng-tu Plain with the industrial and navigation hub of the Yangtze River at Chungking. A Szechwan–Kweichow railway is planned to connect the province with its southern neighbours.

Roads. Because of the thoroughly dissected terrain and the easily weathered rock structures, the building of highways is attended with serious difficulties. The constant threat of landslides, the presence of numerous steep slopes and hairpin turns, and the necessity of constructing many solid embankments make road construction and maintenance costly and hazardous.

Principal highways

Ch'eng-tu and Chungking form the two principal highway focuses of the province; Nan-ch'ung, Ya-an, and San-t'ai are secondary centres. Major highway routes that connect with bordering provinces run to Shensi in the north, Hupeh in the east, Kweichow and Yunnan in the south, and Tibet in the west. The Ch'eng-tu–A-pa–Lan-chou highway and the Ch'eng-tu–Tibet routes cross lofty and rugged mountains, broad deep valleys, and vast high plateaus. Within the province, the most important highways are the Ch'eng-tu–Chungking route and the Ch'eng-tu–Wan-hsien-shih route.

In the west, Ya-an is the only major highway centre. From there routes radiate into Ch'eng-tu in the east, Tsinghai Province in the north, Yunnan Province in the south, and Tibet in the west.

Air services. The Shanghai–Ch'eng-tu airline, initiated in 1937, was the province's first commercial air service. Since then, commercial flying has grown steadily. From Chungking, the principal air-transportation centre, there are regular flights to Han-k'ou and Wu-chiang in the east, Sian in the north, K'un-ming in the south, and Ch'eng-tu in the west. The airlines engage principally in passenger and mail service. Freight capacity is limited.

CULTURAL LIFE

Village fairs

Rural culture. The village fair is an important institution in the rural life of eastern Szechwan. The fairs are widespread and colourful and attract great interest. Because of the isolated character of the settlements in many parts of the province, these fairs fulfill a pressing need. In many cases, the village fair—locally known as *chang* meaning "enclosure"—is the only place where a farmer can buy and sell his products, and is, thus, essential to the rural economy.

Urban culture. Ch'eng-tu has always played a vital role in the cultural and intellectual life of the Szechwanese people. Some of the streets are lined with two-story buildings that boast elaborate carvings. The city is a haven for intellectuals and scholars, and—with its heavy traffic, rich night life, and luxurious suroundings—is sometimes called the "Little Paris" of China.

Architecture. The unique form of architecture of the eastern basin is characterized by projecting eaves, gracefully curved roofs, and rich, elaborate roof ornaments. As there is little wind and practically no snow in the basin, these fragile and extraordinarily beautiful structures and decorations can safely be constructed. The frequent misty rains make it necessary to project the roof eaves over the walls to protect them from the rain.

BIBLIOGRAPHY. J. LOSSING BUCK, *An Agricultural Survey of Szechuan Province, China* (1943), an authoritative work that contains excellent descriptions and evaluations of the province's agricultural conditions; ALEXANDER HOSIE, *Szechuan: Its Products, Industries, and Resources* (1922), a classic work on the economic products of Szechwan; CHARLES Y. HU, *The Agricultural and Forestry Land-Use of Szechuan Basin* (1946), an exhaustive and detailed work, based on firsthand knowledge; FRED O. JONES, "Tukiangyien: China's Ancient Irrigation System," *Geogrl. Rev.*, 44:543–559 (1954), an interesting and detailed description of the irrigation works of the Cheng-tu Plain; J.S. LEE, *The Geology of China* (1939), contains a summary of geological works on China, including Szechwan Province; H.L. RICHARDSON, *Soils and Agriculture of Szechuan* (1942), an authoritative work based on field surveys and investigations of the soils and agriculture of the province; JOSEPH E. SPENCER, "Changing Chungking: The Rebuilding of an Old Chinese City," *Geogrl. Rev.*, 29:46–60 (1939), an accurate description of the city based on first-hand observation; H.B. WHITTINGTON "The Physiographical History of Western Szechuan," *Geol. Mag.* 83:141–146 (1946), an extensive discussion of land forms in western Szechwan; H.E. WILSON, *A Naturalist in Western China* (1913), a valuable volume on the vegetation of Szechwan based on many years of field observation.

(C.Hu./Y.-G.G.H.)

Ṭabarī, aṭ-

Abū Jaʿfar Muḥammad ibn Jarīr aṭ-Ṭabarī, whose academic endeavours spanned most of the Muslim sciences of his day, is remembered today chiefly for his enormous compendia of early Islāmic history and Qur'ānic exegesis. Aṭ-Ṭabarī was born, probably in 839 (AH 224/225), at Āmol in Ṭabaristān, now in Iran. He demonstrated a precocious intellect and journeyed from his native town to study in the major centres of learning in Iraq, Syria, and Egypt. Over the course of many years he collected oral and written material from numerous scholars and libraries for his later work. Aṭ-Ṭabarī enjoyed sufficient financial independence to enable him to devote the latter part of his life to teaching and writing in Baghdad, the capital of the ʿAbbāsid caliphate, where he died in 923 (AH 310). The times in which he lived were marked by political disorder, social crisis, and philosophical-theological controversy. Discontent, of diverse cause and circumstance, brought open rebellion to the very heart of the caliph's empire, and, like all movements of socio-economic origin in medieval Islām, sought legitimacy in religious expression directed against the official credo of Sunnī (major Muslim sectarian division) orthodoxy. Aṭ-Ṭabarī rejected out of hand the extreme theological positions of these opposition movements but at the same time he also retreated from the embrace of the ultra-orthodox Sunnī faction, the Ḥanbalites (a major school of Islāmic law), represented most powerfully in the capital itself. An independent within orthodox ranks, he established his own school of jurisprudence, which did not long survive his own death. He nevertheless made a distinct contribution to the consolidation of Sunnī thought during the 9th century. What aṭ-Ṭabarī accomplished for historical and Qur'ānic studies consisted less in the discovery and initial recording of material than in the sifting and re-organiza-

tion of it. His achievement was to condense the vast wealth of exegetical and historical erudition of the preceding generations of Muslim scholars (many of whose works are not extant in their original form) and to lay the foundations for both Qur'ānic and historical sciences.

Major works

His life's labour began with the *Qur'ān Commentary* and was followed by the *History of Prophets and Kings* (*Ta'rīkh ar-Rusūl wa al-Mulūk*). Aṭ-Ṭabarī's *History* became so popular that the Sāmānid prince Manṣūr ibn Nūḥ had it translated into Persian (*c.* 963).

In the *Commentary*, aṭ-Ṭabarī's method of composition was to follow the Qur'ān text word by word, juxtaposing all of the juridical, lexicographical, and historical explanations transmitted in reports from the Prophet Muḥammad, his companions, and their followers. To each report (Ḥadīth) was affixed a chain of "transmitters" (*isnād*) purporting to go back to the original informant. Divergent reports were seldom reconciled, the scholar's only critical tool being his judgment as to the soundness of the *isnād* and not of the content of the Ḥadīth. Thus a plurality of interpretation was admitted on principle.

The *History* commenced with the Creation followed by accounts regarding the patriarchs, prophets, and rulers of antiquity. The history of the Sāsānian kings came next. For the period of the Prophet's life, aṭ-Ṭabarī drew upon the extensive researches of 8th-century Medinan scholars. Although pre-Islāmic influences are evident in their works, the Medinan perspective of Muslim history evolved as a theocentric (god-centred) universal history of prophecy culminating in the career of Muḥammad and not as a continuum of tribal wars and values.

The sources for aṭ-Ṭabarī's *History* covering the years from the Prophet's death to the fall of the Umayyad dynasty (AD 661–750) were short monographs, each treating a major event or the circumstances attending the death of an important person. Aṭ-Ṭabarī supplemented this material with historical reports embodied in works on genealogy, poetry, and tribal affairs. Further, details of the early 'Abbāsid period were available to him in a few histories of the caliphs that unfortunately have come down to us only in the fragments preserved by aṭ-Ṭabarī. Almost all of these accounts reflected an Iraqi perspective of the community; coupled with this is aṭ-Ṭabarī's scant attention to affairs in Egypt, North Africa, and Muslim Spain so that his *History* does not have the secular "universal" outlook sometimes attributed to it. From the beginning of the Muslim era (dated from 622, the date of the *hijrah*—the Prophet Muḥammad's migration from Mecca to Medina), the *History* is arranged as a set of annals according to the years after the *hijrah*. It terminates in the year 915.

Views of history

Obviously, aṭ-Ṭabarī could not sustain his preference for reports originating with the Prophet and the pious scholars of the early community known as *as-salaf*. His judgment of a report's reliability was now based upon the largely theoretical criterion that it should originate either with an eyewitness or a contemporary informant. This posed a problem for which aṭ-Ṭabarī had no practical solution. Oftentimes he placed separate accounts of an event side by side without editorial comment. He saw no relevance in searching for the nature and causes of events, for any ultimate explanation lay beyond history itself and was known to God alone. Prophetic tradition, like the Qur'ān, provided positive commands and injunctions from God. History pointed to the consequences of heeding or ignoring Him. For aṭ-Ṭabarī, therefore, history was the Divine Will teaching by example.

BIBLIOGRAPHY. FRANZ ROSENTHAL, *A History of Muslim Historiography*, 2nd ed. (1968); O. LOTH, "Tabari's Korankommentar," *Zeitschrift der deutschen Morgenlandischen Gesellschaft*, 35:588–628 (1881). A French translation of the abridged Persian version was made by H. ZOTENBERG, entitled *Chronique de Tabari* (reprinted 1958). E. MARIN translated a portion of the *History* into English, entitled *The Reign of al-Mu'tasim* (1951). The *Commentary* has been edited by M.M. and A.M. SHAKIR, 13 vol. (1954–) under the title *Tafsir al-Tabari*. M.J. DE GOEJE *et al.* edited the *History*, known as the *Annals*, 15 volumes with index and glossary (1879–1901, reprinted 1964).

(D.Wa.)

Table Tennis

Table tennis, an indoor ball game similar in principle to lawn tennis, is played on a flat table divided into two equal courts by a net fixed across its width at the middle. The lightweight hollow ball is made of celluloid or a similar plastic and is propelled back and forth across the net by small rackets (bats or paddles) held by the players. The game is popular all over the world. In most countries it is very highly organized as a competitive sport, especially in Europe and Asia, particularly in China and Japan.

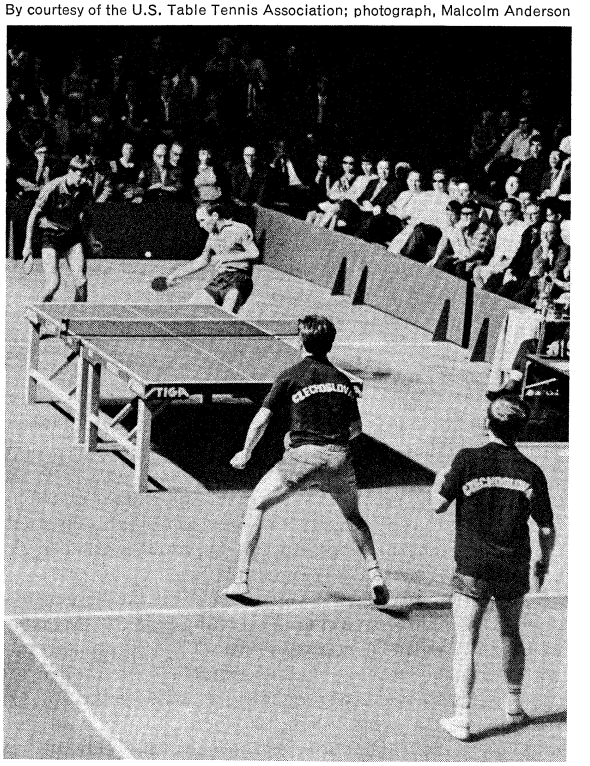
History. Invented in England in the early days of the 20th century, the game was originally called Ping-Pong, which was a trade name. The name table tennis was adopted in 1921–22 when the old Ping-Pong Association formed in 1902 was revived. The original association had broken up about 1905, though apparently the game continued to be played in parts of England outside London, and by the 1920s was being played in many countries. Led by representatives of Germany, Hungary, and England, the Federation Internationale de Tennis de Table (International Table Tennis Federation) was founded in 1926, the founder members being England, Sweden, Hungary, India, Denmark, Germany, Czechoslovakia, Austria, and Wales. By the 1970s more than 90 national associations were members and there were six Continental Federations: Europe, Asia, Africa, North America, South America, and Oceania.

First world championships

The first world championships were held in London in 1927, and from then until 1939 the game was dominated by players from Central Europe, the men's team event being won nine times by Hungary and twice by Czechoslovakia. There was no competition during World War II, but from 1946 to 1953 Czechoslovakia won four times and Hungary twice. These two countries produced some of the greatest players in the game. Victor Barna (five times world singles champion), Laszlo Bellak (twice a finalist), Mike Szabados (once winner and twice finalist) all came from Hungary, as did Ferenc Sido, winner in 1953. From Czechoslovakia came Bohumil Vana, twice world champion and twice a finalist, and Ivan Andreadis, who was a great stylist and four times world doubles champion.

Another Central European, recognized as one of the

By courtesy of the U.S. Table Tennis Association; photograph, Malcolm Anderson



K. Johansson and H. Alser of Sweden, world men's doubles champions in 1967 and 1969, face opponents J. Stanek and V. Miko of Czechoslovakia in a Vanderbilt Cup match, New York, 1970.

greatest champions of all time, was Richard Bergmann of Austria, who won the world singles title four times (twice as a naturalized English citizen). In the women's team event, Central Europe has also been prominent. Of the 16 championships played between 1934 and 1956, Romania and Czechoslovakia won eight times. Great women players have been Maria Mednyanszky (Hungary) who won the ladies singles title each year in the first five years, Gizi Farkas (Hungary) who won the title three years, and Angelica Rozeanu (Romania) with the record of six world singles titles.

The 1953–54 season saw the emergence of Asia as a breeding ground of champions, and from that time the men's team event has been won either by Japan or China while the women's event also has been dominated, though to a lesser extent, by Japan and China. Japan has produced many great players, in particular Ichiro Ogimura and Toshiaki Tanaka, both twice world champions, and China, too, had a three times consecutive winner in Chuang Tse-tung.

At world championships, held biennially since 1957, players compete for:

The Swaythling Cup	Men's team event. The best of nine singles matches.
The Corbillon Cup	Women's team event. The best of four singles and one doubles matches.
St. Bride's Vase	Men's individual singles.
G. Geist Prize	Women's individual singles.
Iran Cup	Men's doubles championships.
W.J. Pope Trophy	Women's doubles championships.
Heydusek Prize	Mixed doubles championships.

For world championship winners see SPORTING RECORD in the *Ready Reference and Index*.

The game. Table tennis equipment is relatively simple and inexpensive. The table is rectangular 9 ft by 5 ft (2.74 m by 1.52 m), its upper surface a horizontal plane 30 in (76 cm) above the floor. The net is 6 ft (1.83 m) long and its upper edge along the whole length 6 in (15.25 cm) above the playing surface. The ball, which is spherical, hollow, and made of white celluloid or, since 1969, a similar plastic, weighs about .88 oz (2.5 gm) and has a diameter of about 1.5 in (3.81 cm).

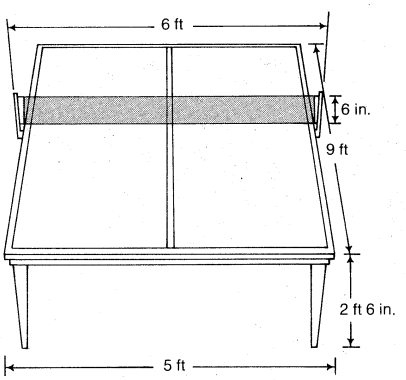


Table tennis table.

The racket or bat is of wood, flat and rigid, and may be either plain uncovered wood, or wood covered with a thin layer of ordinary stippled or pimpled rubber, which may be laid over a thin layer of sponge rubber and may have the pimples reversed; whatever combination is used, the colours on both sides must match and the total thickness must not exceed 0.37 in (4 mm).

Scoring

A match consists of the best of three or the best of five games, each game as in handball being won by the player who first reaches 21 points, or who after 20 points each, wins two clear points ahead. A point is scored when the server fails to make a good service, when either player fails to make a good return, or when either player commits a specified infraction; *e.g.*, touches the playing surface with his free hand while the ball is in play. Service changes hands after every five points until 20-all is reached, when it changes after every subsequent point.

The serve is made from behind the end of the table, the server projecting the ball upward from the palm of his free hand and striking it as it descends so that it first touches his own court and then, passing over the net, touches his opponent's court. In serving no spin may be imparted to the ball by the fingers. This was not always so. Finger spin, especially in the U.S., reached a stage where the experts could produce untakable services and the game became farcical. Finger spin was universally banned in 1937.

Interest to the spectator lies in observing the ability of one player to defeat his opponent by out-manoeuvring him by well-thought-out strategy and by superior fitness, footwork, and stroke repertoire. Increasing the speed of the game, slowing it down, varying the direction of or imparting different spin or pace to the ball, and employing gentle drop shots over the net when the opponent is out of position, are some of the tactics that may be used to support the strategy planned.

Slow or defensive play at one time reached the stage that, at the 1936 world championships in Prague, an hour was needed to decide a single point. Play is now restricted. If a game is unfinished 15 minutes after it has begun, the rest of that game and the remaining games of the match proceed under the Expedite System. Thereafter if the service and 12 following strokes of the server are returned by the receiver, the server loses the point. The service changes after each point.

Variants

Table tennis may be played as a game of singles with one player at each end of the table, or as doubles with two players at each end who may be both men or both women or one of each. Worldwide, the women's game is comparable in organization to the men's, and women take part in world championships and all other organized events. Table tennis as well as being fully organized is also extremely popular as a recreational game and is so played in all types of sports clubs, social clubs, game rooms, in the home, and even out-of-doors when conditions are reasonably calm.

BIBLIOGRAPHY. A. PARK, *Ping Pong* (1902), the first instructional book to give an analysis of the game, now rather obsolete; I.G.S. MONTAGUE, *Table Tennis*, rev. ed. (1951), the first book to introduce cine strip photos showing strokes from preparation to final execution; R. BERGMANN, *Twenty-One Up* (1950), highlights of the author's career as winner of the World Singles Championship, with instructions based on his own experiences; V. BARNA, *How to Win at Table Tennis* (1957), an instructional book, the first part for beginners, the second part for players with ambitions to become champions; and *Table Tennis Today* (1962), on the various methods of stroke production, instructions on how to be successful in tournament play, and descriptions of world champions and their games; S. SCHIFF, *Table Tennis Comes of Age* (1949), a discussion of why this game is so much more popular in other parts of the world than in the U.S. and instructional advice; D. MILES, *The Game of Table Tennis* (1968), hints on how to play based on championship experience; J. LEACH, *Better Table Tennis* (1969), instruction based on photography. Handbooks of official rules and regulations may be obtained from the International Table Tennis Federation in English, French, and German. National handbooks can be obtained from the various national Table Tennis Associations.

(V.B.)

Tacitus, Cornelius

Renowned as a public orator and high official in Rome's imperial administration, Tacitus achieved lasting fame with the historical writing of his later years, in which his political experience and literary skill combine to describe the character of 1st century Roman rule. Though soundly based, this work was often adapted, even distorted, to fit his own view of the empire, which he saw as a study in the pathology of tyranny, where power corrupted and impotence had led to degeneration. Tacitus exploited a brilliant style to emphasize, as well as to record, his interpretation. Hence, his descriptions of policy and persons, especially of the imperial court, are often highly ironic. The political tone did not suit later Roman sentiment, though he was still read for his Latin style. When statecraft again became a subject of interest in the Renaissance, his work was re-examined, as it was, once again, after the French

Revolution. In modern scholarship Tacitus' writings are studied seriously—with critical reservation—to reconstruct the early history of the Roman Empire. On the literary side it is appreciated as a stylistic masterpiece.

Family and early career. Tacitus was born *c.* AD 56, perhaps in north Italy (Cisalpine Gaul) or, more probably, in southern Gaul (Gallia Narbonensis, or present southeastern France). Nothing is known of his parentage. Though Cornelius was the name of a noble Roman family, there is no proof that he was descended from the Roman aristocracy; provincial families often took the name of the governor who had given them Roman citizenship. In any event he grew up in comfortable circumstances, enjoyed a good education, and found the way open to a public career.

Tacitus studied rhetoric, which provided a general literary education including the practice of prose composition, learning the concentrated and pointed style of the 1st century. This training was a systematic and fair preparation for administrative office. Although often artificial, this course of study would, along with conscious effort, enable a talented writer to express himself well. Tacitus studied to be an advocate at law under two leading orators, Marcus Aper and Julius Secundus; then he began his career with a "virgintivirate" (one of 20 appointments to minor magistracies) and a military tribunate (on the staff of a legion). In so doing he followed the traditional course for young men of social standing, and it may be assumed that he displayed his ability and promise at each of these early stages of his career.

Marriage

In 77 Tacitus married the daughter of Gnaeus Julius Agricola. Agricola had risen in the imperial service to the consulship, in 77 or 78, and he would later enhance his reputation as governor of Britain. Tacitus appears to have made his own mark socially and was making much progress toward public distinction; he would obviously benefit from Agricola's political connections. Moving through the regular stages, he gained the quaestorship (often a responsible provincial post), probably in 81; then in 88 he became a praetor (a post with legal jurisdiction) and a member of the priestly college that kept the Sibylline Books of prophecy and supervised foreign-cult practice. After this it may be assumed that he held a senior provincial post, normally in command of a legion, for four years. When he returned to Rome, he observed firsthand the last years of the emperor Domitian's savage oppression of the Roman aristocracy. By 93 Agricola was dead, but by this time Tacitus had achieved distinction on his own in the imperial service. In 97, under the emperor Nerva, he rose to the consulship and delivered the funeral oration for Verginius Rufus, a famous soldier who had refused to compete for power in 68/69 after Nero's death. This distinction not only reflected his reputation as an orator but his moral authority and official dignity as well. For an ambitious man there remained only appointment to a high provincial governorship, and Tacitus had to await his turn; in the meantime he turned to literary composition.

First literary works

In 98 Tacitus wrote two works: *De vita Julii Agricolae* and *De origine et situ Germanorum* (the *Germania*), both reflecting his personal interests. The *Agricola* is a biographical account of his father-in-law's career, with special reference to the governorship of Britain (78–84) and the later years under Domitian. It is laudatory yet circumstantial in its description, and it gives a balanced political judgment. Agricola (as Tacitus depicts him) was a man of true Roman virtue who defended the empire without either reckless defiance or craven servility toward an evil emperor, and he recognized that Rome was greater than its rulers. One may sense the tone of administrative apologia, for the fact that the political martyrs of Domitian's reign were now being glorified may have tarnished (if only by implication) the reputation of those who had advanced their careers during this period. And both Tacitus and Agricola had done well.

The next work, the *Germania*, is another descriptive piece, this time of the Roman frontier on the Rhine. Tacitus emphasizes the simple virtue as well as the primitive vices of the Germanic tribes, in contrast to the moral laxity of contemporary Rome, and the threat that these tribes, if they acted together, could present to Roman Gaul. Here his writing goes beyond geography to political ethnography. The work is hardly, as has been argued, a pamphlet about Roman strategy in the West: it gives an administrator's appreciation of the German situation, and to this extent the work serves as a historical introduction to the Germans.

Ancient literature had already provided models of style for biography, geographical description, and ethnographical study. Tacitus knew these conventions and adapted them to his new writing. What, then, of his celebrated oratory? He still practiced advocacy at law—in 100 he, along with Pliny the Younger, successfully prosecuted Marius Priscus, a proconsul in Africa, for extortion—but he felt that oratory had lost much of its political spirit and its practitioners were deficient in skill. This decline of oratory seems to provide the setting for his *Dialogus de oratoribus*. The work refers back to his youth, introducing his teachers Marcus Aper and Julius Secundus. It has been dated as early as *c.* 80, chiefly because it is more Ciceronian in style than his other writing. But its style arises from its form and subject matter and does not point to an early stage of stylistic development. The date lies between 98 and 102; the theme fits this period. Tacitus compares oratory with poetry as a way of literary life, marking the decline of oratory in public affairs: the Roman Republic had given scope for true eloquence; the empire limited its inspiration. The work reflects his mood at the time he turned from oratory to history.

Historical writing. There were historians of imperial Rome before Tacitus, notably Aufidius Bassus, who recorded events from the rise of Augustus to the reign of Claudius, and Pliny the Elder, who continued this work (*a fine Aufidii Bassi*) to the time of Vespasian. In taking up history Tacitus joined the line of succession of those who described and interpreted their own period, and Tacitus took up the story from the political situation that followed Nero's death to the close of the Flavian dynasty.

The *Histories*

The *Historiae* began at January 1, 69, with Galba in power and proceeded to the death of Domitian, in 96. The work contained 12 or 14 books (it is known only that the *Histories* and *Annals*, both now incomplete, totalled 30 books). To judge from the younger Pliny's references, several books were ready by 105, the writing well advanced by 107, and the work finished by 109. Only books i–iv and part of book v, for the years 69–70 are extant. They cover the fall of Galba and Piso before Otho (book i); Vespasian's position in the East and Otho's suicide, making way for Vitellius (book ii); the defeat of Vitellius by the Danubian legions on Vespasian's side (book iii); and the opening of Vespasian's reign (books iv–v). This text represents a small part of what must have been a brilliant as well as systematic account of the critical Flavian period in Roman history, especially where Tacitus wrote with firsthand knowledge of provincial conditions in the West and of Domitian's last years in Rome. The narrative as it now exists, with its magnificent introduction, is a powerfully sustained piece of writing that, for all the emphasis and colour of its prose, is perfectly appropriate for describing the closely knit set of events during the civil war of 69. Tacitus found the opening of the *Histories* well within the scope of his political and military experience and his literary training; he displays his firm control of the incisive Latin style of his time.

This was only the first stage of Tacitus' historical work. As he approached the reign of Domitian, he faced a Roman policy that, except in provincial and frontier affairs, was less coherent and predictable. It called for sharper analysis, which he often met with bitterness, anger, and pointed irony. Domitian's later despotism outraged the aristocratic tradition. It is not known, and it is the most serious gap, how Tacitus finally handled in detail Domitian's reputation. Perhaps his picture of the emperor Tiberius in the *Annals* owed something to his exercise on Domitian.

It is necessary to keep the dating of Tacitus' work in mind. He had won distinction under Nerva and enjoyed

the effects of liberal policy; at the same time he had lived through the crisis of imperial policy that occurred when Nerva and Trajan came to the succession. Under Trajan he retained his place in public affairs, and in 112–113 he crowned his administrative career with the proconsulate of Asia, the top provincial governorship. His personal career had revealed to him, at court and in administration, the play of power that lay behind the imperial facade of rule. He was especially familiar with the effect of dynastic control, which tended to corrupt the rulers, as it had in the period from Vespasian to Domitian, and to reduce the supporting nobles to servility, while only military revolt within Rome or from the frontier legions could change the situation—as it had done at the end of Nero's reign.

From what can be reconstructed from his personal career along with the implications of his subsequent historical thought, it is possible to mark an intellectual turning point in his life after which he began to probe deeper into the nature of the Roman Empire. Although in the *Agricola* he had lightly promised to continue his writing from the Flavian years into the new regime, he now moved not forward but backward. He was no longer content to record the present but to interpret the political burden of the past from the time when Tiberius consolidated Augustus' policy of imperial government. It was for these reasons that he wrote the *Annals*.

The *Annals*

The *Annals* (*ab excessu divi Augusti*), following the traditional form of yearly narrative with literary elaboration on the significant events, covered the period of the Julio-Claudian dynasty from the death of Augustus and the accession of Tiberius, in 14, to the end of Nero's reign, in 68. The work contained 18 or 16 books. There is some evidence that he organized his books into groups of six: books i–vi on Tiberius (subdivided into books i–iii on Tiberius and Germanicus, books iv–vi on Sejanus and Tiberius), and books vii–xii on Caligula (Gaius) and Claudius. To judge from references to Parthia that reflect its power before Trajan's invasion, Tacitus had written most of books i–vi by 114. If a reference in book ii, 61 to Mare Rubrum ("Red Sea") as the Roman frontier means the Persian Gulf, he was touching up these books in 116–117, after Trajan's victory and before Hadrian's withdrawal; but the term refers rather to the Red Sea as we know it, after the annexation of Arabia in 105–106. In any event he probably completed the work early in Hadrian's reign. Only books i–iv, part of book v, most of book vi (treating the years 14–29 and 31–37 under Tiberius), and books xi–xvi, incomplete at beginning and end (on Claudius from 47 to 51 and Nero from 51 to 66), now exist.

In casting back to the early empire Tacitus did not wish necessarily to supersede his predecessors in the field, whose systematic recording he seemed to respect, judging from the use he made of their subject matter. His prime purpose was to reinterpret critically the Julio-Claudian dynasty, when imperial rule developed a central control that, even after the complex military coup d'etat of 68–69, would continue under the Flavians. In effect, the *Annals* represents a diagnosis in narrative form of the decline of Roman political freedom, written to explain the condition of the empire he had already described in the *Histories*. Tacitus viewed the first imperial century as an entity. There was (in his eyes) a comparison to be made, for example, between the personal conduct of Tiberius and that of Domitian, not that they were the same kind of men but that they were corrupted by similar conditions of dynastic power. Yet he did not begin with Augustus, except by cold reference to his memory. The modern world tends to think of Augustus as the founder of the empire. The Romans—one may cite Appian and Florus alongside Tacitus—regarded him, at least during the first part of his career, as the last of the warlords who had dominated the republic.

Criticism of the imperial system

In opening the *Annals* Tacitus accepts the necessity of strong, periodic power in Roman government, providing it allowed the rise of fresh talent to take over control. That was the aristocratic attitude toward political freedom. But to secure the continuity of personal authority by dynastic convention, regardless of the qualifications for rule—he could quote the examples of Caligula, Nero, and Domitian—was to subvert the Roman tradition and corrupt public morality. If Augustus began as a warlord, he ended by establishing a dynasty. But the decisive point toward continuing a tyrannical dynasty was Tiberius' accession. One may, indeed, believe that Tiberius was prompted to assume imperial power because he was anxious about the military situation on the Roman frontier; but Tacitus had no doubts about the security of the Roman position, and he considered the hesitation Tiberius displayed on taking power to be hypocritical; hence, the historical irony, in interpretation and style, of his first six books. Here, perhaps, Tacitus had some support for his interpretation. A strong, dour soldier and a suspicious man, Tiberius had little to say in his court circle about public affairs. On his death he was blamed for never saying what he thought nor meaning what he said, and Tacitus elaborated this impression. His criticism of dynastic power also stressed the effect of personality: if Tiberius was false, Claudius was weak, Nero was not only unstable but evil, and the imperial wives were dangerous. With regard to provincial administration, he knew that he could take its regular character for granted, in the earlier period as well as his own.

Sources

For the period from Augustus to Vespasian, Tacitus was able to draw upon earlier histories and monographs that contained material from the public records, official reports, and contemporary comment. It has been noted that the work of Aufidius Bassus and its continuation by the elder Pliny covered these years; both historians also treated the German wars. Among other sources Tacitus consulted Servilius Nonianus (on Tiberius), Cluvius Rufus and Fabius Rusticus (on Nero), and Vipstanus Messalla (on the year 69), who are only names to us. He also turned, as far as he felt necessary, to the Senate's records, the official journal, and such firsthand information as a speech of Claudius, the personal memoirs of the younger Agrippina, and the military memoirs of the general Corbulo. For Vespasian's later years and the reigns of Titus and Domitian, he must have worked more closely from official records and reports, applying the impressions of his own and his friends' knowledge.

In the light of his administrative and political experience, Tacitus in the *Histories* was able to interpret the historical evidence for the Flavian period more or less directly. Yet contemporary writing may lack perspective. He recognized this problem when, in the *Annals*, he revived the study of the Julio-Claudian dynasty. But to go back a century raises additional problems of historical method. Tacitus first had to determine the factual reliability and political attitude of his authorities and then to adjust his own general conception of the empire, in case it was anachronistic, to the earlier conditions. The strengh of his conviction limited his judgment at both points. He underplayed the effect of immediate circumstances and overplayed the personal factor, a tendency that influenced his use of the historical sources. In particular Tiberius, who in spite of his political ineptness struggled with real difficulties, suffered in reputation from this treatment. But Tacitus did not spare any man in power. He controls the performance of his characters; it is magnificent writing, but it is not necessarily strict history.

Literary style. Since he was a conscious literary stylist, both his thought and his manner of expression gave life to his work. Greek historiography had earlier defined ways of depicting history: one could analyze events in plain terms, set the scene with personalities, or heighten the dramatic appeal of human action. Each method had its technique, and the greater writer could combine elements from all three. The Roman "annalistic" form, after years of development, allowed this varied play of style in significant episodes. Tacitus knew the techniques and controlled them for his political interpretations; as a model he had studied the early Roman historiographer Sallust.

It is finally his masterly handling of literary Latin that impresses the reader. He wrote in the grand style, helped by the solemn and poetic usage of the Roman tradition, and he exploits the Latin qualities of strength, rhythm, and colour. His style, like his thought, avoids artificial

smoothness. His writing is concise, breaking any easy balance of sentences, depending for emphasis on word order and syntactical variation and striking hard where the subject matter calls for a formidable impact. He is most pointed on the theme of Tiberius, but his technique here is only a concentrated form of the stylistic force that can be found throughout his narrative.

Tacitus' work did not provide an easy source for summaries of early imperial history, nor (one may guess) was his political attitude popular in the ruling circles; but he was read and his text copied until in the 4th century Ammianus Marcellinus continued his work and followed his style. A thousand years after his own day it was the Renaissance that welcomed him for his illustration of statecraft, as did the political thinkers of the French Revolution, who enjoyed his condemnation of dynastic tyranny. More soberly modern scholarship has probed his evidence and appreciated his Latin, and, in light of contemporary epigraphic and archaeological study, one may set Tacitus' political pattern within the wider perspective of the Roman Empire.

BIBLIOGRAPHY

Texts: C.D. FISHER and H. FURNEAUX, 3 vol., "Oxford Classical Library" (1900–10); E. KOESTERMANN, 3 vol., "Teubner Series" (1960–62); W. PETERSON *et al.*, "Loeb Series," with Eng. trans. (1914–37).

Commentaries: Agricola: R.M. OGILVIE and I.A. RICHMOND (1967). *Germania:* J.G.C. ANDERSON (1938). *Dialogus:* W. PETERSON (1893). *Histories:* W.A. SPOONER (1891). *Annals:* Books i-vi, H. FURNEAUX (1896); books xi-xvi, H. FURNEAUX, rev. ed. by H.F. PELHAM and C.D. FISHER (1907).

Translations: A.J. CHURCH and W.J. BRODRIBB, 3 vol. (1869–72). *Annals and Histories:* G.G. RAMSAY, 2 vol. (1904–15). *Histories and minor works:* W.H. FYFE, 2 vol. (1908–12).

On the historical material: M.P. CHARLESWORTH *et al.* in the *Cambridge Ancient History*, vol. 10, ch. 19–25 (1934), still the best treatment (especially on Tiberius); for provincial background, M.I. ROSTOVTSEV, *The Social and Economic History of the Roman Empire*, 2nd ed., rev. by P.M. FRASER, 2 vol. (1957).

Literary discussion: E. LOFSTEDT, *Romare* (1956; Eng. trans., *Roman Literary Portraits*, ch. 7–8, 1958), on style and language.

General: G. BOISSIER, *Tacite* (1903; Eng. trans., *Tacitus and other Roman Studies*, 1906); R. SYME, *Tacitus*, 2 vol. (1958), the most comprehensive and sensitive study of Tacitus' life and work. For a critical survey (with bibliography) of Tacitus' place in Roman historiography, see A.H. MCDONALD, "The Roman Historians," in M. PLATNAUER (ed.), *Fifty Years (and Twelve) of Classical Scholarship*, 2nd ed. (1968).

(A.H.McD.)

Tadzhik Soviet Socialist Republic

The Tadzhik (or Tajik) Soviet Socialist Republic (Tadzhik S.S.R., or Tadzhikistan, in the Russian spelling; as transliterated from the Tadzhik language in its own modified Cyrillic script, the name is Tojikiston), lying in the mountainous heart of Asia, is one of the 15 constituent union republics of the Soviet Union. It is in the extreme south, bounded by China to the east, Afghanistan to the south, and the Uzbek and Kirgiz S.S.R.'s to the west and north. Covering 55,250 square miles (143,100 square kilometres), it was the home by 1975 of almost 3,400,000 people. The capital is Dushanbe, known from 1929 till 1961 as Stalinabad. The republic includes the Gorno-Badakhshan Autonomous Oblast, with its capital at Khorog. Tadzhikistan was first constituted as an autonomous republic within the Uzbek S.S.R. in 1924; it became a union republic in 1929.

As in the other Central Asian republics, an economic and social transformation has taken place during the Soviet period: industry now provides a large share of the republic's income, although cotton growing is still a mainstay of the economy. This change has been accompanied by considerable immigration and a high birth rate. Much of the old way of life nevertheless survives, and the republic's ancient cultural traditions are preserved.

For related information see also PAMIR MOUNTAIN AREA; TIEN SHAN (MOUNTAINS); INNER ASIA, HISTORY OF; CENTRAL ASIAN PEOPLES, ARTS OF; INDO-IRANIAN LANGUAGES: *Iranian languages*; SOVIET UNION; RUSSIA AND THE SOVIET UNION, HISTORY OF.

THE LAND

Relief and drainage patterns. The Tadzhik S.S.R. is a mountainous country, more than half its territory lying above 10,000 feet (3,000 metres). The western Tien Shan reaches into the north, with the Kuramin and Mogoltau ranges. In the central portion of the republic are found the massive ranges of the southern Tien Shan: the Turkistan Range and the slightly lower Zeravshan, Gissar (Hissar), and, in part, Alay ranges. The southeast contains the ice-clad peaks of the Pamir–Alay mountain system, which is made up of the Pamir Upland and, extending westward, the Peter I and Darvaz ranges. The Soviet Union's highest peaks—Lenin Peak (23,405 feet [7,134 metres]) and Communism Peak (24,590 feet [7,495 metres])—are found in the northern portion of the Pamirs. The relief of the southwestern portion of the republic, lying between the Gissar range and the Pyandzh (Panj) River (Ab-e Panj in Afghanistan), is made up of a series of transverse ranges interspersed by wide river valleys and structural depressions.

The valleys, although important on Tadzhikistan's human geography, comprise less than 10 percent of the total area of the republic; the largest are the western portion of the Fergana (Fergnana) Valley in the north and the Gissar, Vakhsh, Yavansu, Obikiik, Lower Kafirnigan (Kafirnihan), and Pyandzh valleys to the south. The whole country is in an active seismic belt, and severe earthquakes are common.

Rivers

The republic has a dense river network, the largest courses being those of the Syrdarya and the Amu Darya, together with their tributaries, notably the Vakhsh and the Kafirnigan. Upstream from the Vakhsh–Amu Darya junction, the latter river is known as the Pyandzh and forms a good portion of the southern boundary of the republic. Other rivers include the Zeravshan (Zarafshān), which runs through the central part of the republic. Most of the rivers drain into the Aral Sea Basin, although a few of the small streams running down from the Pamirs empty into Lake Kara-Kul, which has no outlet. The rivers have two high-water periods each year: in the spring, when the mountain snows melt and rainfall occurs, and in the summer, when the glaciers begin to melt. The summer flow is particularly helpful for irrigation purposes.

Most of the few lakes in Tadzhikistan are found in the Pamir region. Lake Kara-Kul, salt and devoid of life, lying at an altitude of about 13,000 feet, is the largest; Lake Sarez was formed in 1911 during an earthquake, when a colossal landslide dammed the Murgab River. The Zeravshan range contains Iskanderkul, of glacial origin like most Tadzhikistan lakes.

Climate and plant and animal life. The climate of Tadzhikistan is sharply continental and changes with altitude. In the subtropical valley areas, summer is very hot and dry; the mean July temperature is 81.3° F (27.4° C) in Leninabad and 86.5° F (30.3° C) in Kulyab, farther south. The corresponding January figures are 29.7° F (−0.9° C) and 36.1° F (2.3° C), respectively. In very severe winters, the mercury has been known to drop to −4° F (−20° C), and even lower. Annual precipitation is slight, ranging between six and 10 inches (150 and 250 millimetres), but is higher in the Gissar Valley. In the highlands the situation is different; the mean January temperature in Murgab is −3.3° F (−19.6° C), and the temperature sometimes drops to −51° F (−46° C), with the record standing at −81° F (−63° C). Precipitation is scant, barely reaching two to three inches (60 to 80 millimetres) a year, mostly falling in summer.

Precipitation

The long, wide valleys that separate the ranges of the Tien Shan and the Pamir-Alay join in the Turgay Valley in the southwest, but in the east the high mountain barriers form dead ends. As a result, moist air masses from the west move up the valleys and suddenly reach low temperature areas, with heavy precipitation (30–60 inches [800–1,500 millimetres] or more annually), falling

mostly as snow. Not all this snow melts in the summer heat, and there are large ice fields in the higher regions.

This topographic and climatic variation has given Tadzhikistan an extremely varied plant life, with more than 5,000 kinds of flowers alone. Generally, grasses, bushes, and a variety of shrubs predominate. Animal life, too, is rich and varied. The great gray lizard, jerboa, and gopher are typical inhabitants of the desert regions. In wooded areas deer and tigers are still sometimes seen, and jackals and wildcats live and hunt in reed thickets. In the mountains, the brown bear is found at the lower levels, and the goat and the golden eagle frequent higher altitudes.

The human imprint. The soils of the republic are mostly gray soils, poor in organic humus but rich in mineral nutrients. In unirrigated areas, there are dark gray steppe soils and brown mountain loess. About two-thirds of the surface is covered by sand, shingle, scree, bare rock, and permanent snow and ice. Those desert and semidesert areas suitable for irrigated farming have been turned into flourishing oases, with cotton plantations, gardens, and vineyards, by means of an extensive irrigation network. Population density is also high in the large villages strung out in clusters along the foothill regions.

The Pamir region

The landscapes of the Pamir region are quite distinctive, especially in the east, where there is sparse vegetation and the underbrush, known as *teresken*, supplies the only fuel. The high-altitude semidesert regions manage to provide year-round food for sheep. There are also narrow valleys with small villages (*qishlaq*s) surrounded by apple orchards, apricot trees, mulberry groves, and small cultivated fields.

Tadzhikistan may be divided into a number of regions: Northern Tadzhikistan, the southern boundary of which is the northern slope of the Turkistan Range; Central Tadzhikistan, or Kukhistan (*kūhistān*, "mountain country"), containing the Turkistan, Zeravshan, and Gissar ranges; low-lying Southwestern Tadzhikistan, containing two-thirds of the country's people and stretching to the south of the Gissar Range; and the Pamirs. The last may be subdivided into the East Pamirs—which, in its structure, resembles Tibet—and the deeply cleft West Pamirs.

THE PEOPLE

National origins and demographic trends. The name Tadzhik (Tajik) did not acquire its present definition till very recent times. Not unil the 1920s did an official Tadzhik administrative unit come into being under that name. The area's population is ethnically extremely mixed, as it has been for centuries. At the census of 1970, which showed a total population of 2,900,000, only 56.2 percent was Tadzhik.

On the basis of language, customs, and other traits, the Tadzhiks can be subdivided into a number of distinct groups. The Pamir Tadzhiks include minority people speaking Wakhī, Shughnī, Rōshānī, Khufī, Yāzgulāmī, Ishkashmī, and Bartang, grouped in the Gorno-Badakhshan Autonomous Oblast. Another distinct group is formed by the Yaghnābīs, direct descendants of the ancient Sogdians, who live in the Zeravshan River basin.

So closely are the Tadzhiks mixed with their Muslim coreligionists the Uzbeks that the Soviet partition of the area in 1924 failed to segregate the two nationalities with any degree of thoroughness. The 665,662 Uzbeks constituted the second largest ethnic group in Tadzhikistan in 1970, making up nearly 23 percent of the population. Russian immigrants and descendants of earlier Russian immigrants formed almost 12 percent of the population. Others, in order of size, were Tatars, Germans, Kirgiz, Ukrainians, and Jews, altogether supplying an additional 6.5 percent.

The density of population is low, but the average figure —some 61 persons per square mile (24 per square kilometre)—conceals a variation between the valleys (up to five times the average figure) and the almost empty eastern mountain regions.

Like the other Central Asian republics, Tadzhikistan has a high rate of population increase. The natural population growth rate is three times the overall Soviet figure and is even higher in the Gorno-Badakhshan region. Half the population is under 20 years of age, and children under nine constitute more than 30 percent.

The population is made up of collective farmers (54 percent), workers (31 percent), and white-collar employees (about 16 percent). Women constitute almost 40 percent of the industrial and office employees.

Patterns of settlement. The Tadzhik S.S.R. contains almost 20 cities, together with about 50 other settlements classified as urban. The two largest cities are Dushanbe, with a population of some 436,000 in 1975—the Tadzhik portion of which had evidently grown little, if any, above the 26.2 percent it formed in 1970—and Leninabad (formerly Khojand), with more than 100,000. Russians dominate Dushanbe's ethnic mixture (1970), constituting 42 percent of the inhabitants; Uzbeks make up 10.3 percent. The more important of the smaller urban centres are Kulyab, Kurgan-Tyube, and Ura-Tyube. Kayrakkum, Ordzhonikidzeabad, Kalininabad, Nurek, and Regar are new creations. The combined urban population of the republic exceeded 1,240,000 by the mid-1970s.

Urban life

The larger part of the population, however, continues to live in the rural *qishlaq*s. Such a settlement usually consists of 200 to 700 one-family houses built along the banks of a river or an irrigation ditch. Traditionally, the house is surrounded by a mud fence. Many houses have flat roofs, and every household has an orchard or vineyard. In the mountains, the *qishlaq*s, sited in narrow valleys, form smaller settlements, usually 15 to 20 households. On the steep slopes, the flat roof of one house often serves as the yard for the house above it. There are schools, clinics, shops, and cultural centres in most *qishlaq*s. Settlements in newly developed areas have a more urban aspect; they cluster together and often include two-story houses and a community service centre.

THE ECONOMY

Tadzhikistan is a major producer of long-staple cotton, accounting for more than 10 percent of the Soviet cotton output. The main industries are nonferrous metallurgy, cotton processing, oil production, fruit canning, and wine making. Capital investment in the republic's economy has grown substantially, more than doubling during the decade of the 1960s.

Resources. The republic is rich in mineral deposits, including iron, lead, zinc, antimony, mercury, gold, tin, and tungsten. Nonmetals include common salt, carbonates, fluorite, arsenic, quartz sand, asbestos, and precious and semiprecious stones. The major mineral fuel is coal. There are also considerable deposits of petroleum and natural gas. Tadzhikistan's natural wealth also includes medicinal mineral springs. Finally, there is great potential electric power in the republic's rivers, most of it concentrated in the south.

Industry. Although 62.9 percent of the republic's entire population and 74.5 percent of the Tadzhiks themselves remained in the countryside in 1970, Soviet authorities classify Tadzhikistan as an industrial-agrarian republic, with both sectors versatile.

The chief mining and ore-dressing area is in the north. Oil extraction and coal mining are among the oldest industries in the republic, the latter the more significant. Gas extraction began in the Kyzyl-Tumshuk and Komsomol fields (the latter near Dushanbe) in the mid-1960s, but, because of its small scale, local supplies have to be augmented by natural gas from Afghanistan, delivered by means of a pipeline built in 1974. The power industry has been developing rapidly. Major power stations include the Kayrakkum on the Syrdarya and the Golovnaya on the Vakhsh, with capacities of 126,000 and 210,000 kilowatts, respectively. The Nurek station on the Vakhsh, which became operational in 1972, will eventually have a capacity of 2,700,000 kilowatts. An important thermal power station operates near Dushanbe.

Electric power

Light industry includes a dozen cotton-cleaning mills and two silk factories. The Dushanbe textile mill complex is the republic's largest enterprise in the cotton industry. Other branches of light industry include manufacture of knitted goods and footwear, sewing, and tanning. The

Soviet Union's largest carpet-making factory is in Kayrakkum. The food industry concentrates on processing local agricultural products, which include fruit, grapes, various natural oils, and tobacco. The republic is the main Soviet producer of geranium oil, used in perfumery.

Engineering and metalworking experienced a 15-fold increase during the 1950s and '60s; products include looms, power transformers, cables, and agricultural and household equipment. The chemical industry includes a nitrogen fertilizer plant in Kalininabad (built in 1967), and an electrochemical complex is under development at Yavan. An aluminum plant at Regar began operations in 1975 upon delivery of the first power from the Nurek power station.

Agriculture. Cotton growing is the leading branch of Tadzhik agriculture. Other important branches are cattle raising and the growing of fruit, grain, and vegetables; sericulture is also widespread, except in the Pamirs. Wheat and barley are the predominant grains, and rice cultivation has been expanded. Horticulture has been significant in Tadzhikistan since ancient times: apricots, pears, apples, plums, quinces, cherries, pomegranates, figs, and nuts are all grown. Dried apricots, nuts (particularly almonds), and grapes are important exports. There are almost 1,000,000 long-horned cattle in the republic and nearly 3,000,000 sheep (Gissar sheep) and goats.

The role of irrigation

Agriculture in Tadzhikistan would be severely limited without extensive irrigation. Emulating the technique widely practiced in Central Asia since ancient times, two main canals, the Vakhsh and the Gissar, were built at the end of the 1930s, followed by two joint Tadzhik–Uzbek projects, the Great Fergana and North Fergana canals. After World War II the network was expanded by construction of the Dalverzin and Parkhar–Chubek irrigation systems; the Muminabad, Kattasay, and Selbur reservoirs; the Hungry Plains irrigation system; and a tunnel through which water is brought from the Vakhsh to the Yavansus Valley.

Transportation. The railway net in Tadzhikistan is not extensive, and about two-thirds of it is narrow-gauge. Standard-gauge railways cross the northern part of the republic and enter the Gissar Valley in the south. Narrow-gauge railways connect Dushanbe with the Vakhsh Valley and the Kulyab district. A standard-gauge railway line 160 miles long, completed in the early 1970s, connects the power and industrial complex at Yavan in the south of the republic with other regions. Just under half of the total freight turnover is handled by rail, the rest by motor transport. About half of the highways are surfaced. Regular air service connects Tadzhikistan with Moscow, Tashkent, and other Soviet cities, and there is much internal air traffic. Dushanbe is the air-transport centre, and there is an airport in Leninabad. A small amount of freight is carried by water, mainly between the river port of Termez (in the Uzbek S.S.R.) and Faizal-Kala (160 miles up the Amu Darya).

ADMINISTRATIVE AND SOCIAL CONDITIONS

Constitutional framework. The constitution of the Tadzhik S.S.R., adopted on March 1, 1937, states that the republic has all the rights of a sovereign state, including its national emblem and anthem. In practice, this republic, like the other 13 non-Russian union republics, does not possess and cannot exercise independence from Moscow in any vital matter, whether economic, social, cultural, political, or military. The republic has a Supreme Soviet, selected every four years, which ordinarily meets briefly twice each year to provide *pro forma* sanction to acts and decisions taken elsewhere. The Council of Ministers is the administrative organ and, though formally accountable to the republic's Supreme Soviet, takes its direction from executive bodies in Moscow. Local administration is carried out by regional, district, city, and village bodies, selected from a single list of candidates every two years. Judges are selected in a similar manner every five years; the courts include the Supreme Court of the Tadzhik S.S.R., the court of the Gorno-Badakhshan Autonomous Oblast, and district and city courts.

Supreme Soviet and Council of Ministers

The Communist Party of Tadzhikistan, with more than 94,200 members and candidates in 1975, is the true centre of power in the republic, but it, too, takes its direction in all respects from the party headquarters in the Russian S.F.S.R. Though Tadzhik membership in the union-wide party organization had grown from a mere 234 in 1925 to 57,271 by 1972, the proportion of Tadzhik party members to Tadzhik population remains lower than that of any other union republic eponymous nationality except the Moldavians. Within the Communist Party of Tadzhikistan, Tadzhiks have also remained in minority status; in 1962 Tadzhiks constituted 45.5 percent of the members, compared with 24.2 percent Russians, while their percentages in the total population (1959) were 53 and 13.3, respectively. The Young Communist League (Komsomol), another party organization, had a membership in 1975 of about 305,000. Tadzhik trade unions, representing party and state rather than workers' interests, have about 750,000 members of various nationalities.

Social conditions. The republic constitution says that all inhabitants, regardless of nationality or race, have equal right to employment, vacations, education, and social security on retirement or during disablement. In actuality, considerable inequality exists in the realization of such rights; the educational level, for example, of different ethnic groups varies greatly. Tadzhiks awarded college degrees grew from 0.8 percent in 1959 to 2.1 percent in 1970, while Russians rose from 4.4 to 6.2 percent. Tatars living in Tadzhikistan also surpassed their host Tadzhiks substantially.

Education

Nevertheless, the educational level of Tadzhiks has risen greatly in the 20th century, initially under the impetus supplied by the Jadid (New Method) schools in Bukhara, Samarkand, and elsewhere, opened or taught by intellectual leaders such as Shakuri, Behbudi, and Ayni. In the late 19th century and early 20th there were also a few tsarist schools offering secular learning to indigenous pupils. The educational establishment was dominated until the 1920s, however, by the network of Muslim *mekteb*s and *medrese*s for which Central Asia was well known.

Literacy for Tadzhikistan's inhabitants between the ages of nine and 49 in 1970, according to official statistics, reached 99.6 percent, though there remained many Tadzhiks over age 49 who had not finished a grade school or high school education. Progress in schooling among younger groups has been great and will no doubt in a few decades offset the limitations in the education of older adults.

There were in the mid-1970s about 3,200 schools attended by some 900,000 children; about 40 specialized secondary schools enrolling 37,400 students; and nine institutions of higher learning had a combined registration surpassing 47,000. There are about 60 research institutes, 16 of them under the auspices of the Tadzhik Academy of Sciences, founded in 1951. This emphasis on education and research keeps Tadzhikistan supplied with young specialists. The education of women lags far behind that of men (0.7 percent of Tadzhik women in the republic held college degrees in 1970). The chronic Central Asian problem of placing indigenous graduates in employment commensurate with their advanced training besets the Tadzhiks, for most urban and industrial jobs are held by outsiders such as Russians, Tatars, and Ukrainians.

There is an extensive modern medical system, with 96 hospital or clinic beds and 19 doctors per 10,000 people. Housing construction and community services are also developing. Electricity has been brought to every settlement, and the gas network is being extended. Every year about 20,000 modern flats are turned over to tenants, and living conditions generally are improving. These amenities, however, benefit principally the urban centres, with their large non-Tadzhik populations.

CULTURAL LIFE

Tadzhikistan has an ancient culture, and over the centuries it has given the world poets, philosophers, and scientists. Many of the popular traditions and customs have

Festivals

been retained, including the costumes worn by both men and women and such ancient festivals as the New Year celebration of the tillers of the land, known as Navrūz, which takes place on March 21, the period of the vernal equinox. A festival celebrated after the gathering of the cotton crop is another. These festivals are colourful affairs, incorporating horse races and horsemanship and wrestling contests. The Muslims (mostly Sunnī) continue regular mosque worship and observe religious holidays where possible, though religion is actively persecuted.

The national cuisine includes flat cakes, the equivalent of bread, made in special ovens; *manti*, a kind of large dumpling; *laghmān*, a stew of chopped meat and noodles; *kaurdak*, fried meat and potatoes; and fermented milk. Fruit is an important dietary item, as are black and green tea.

Languages

The principal language of Tadzhikistan, Tadzhik (known to its speakers as Tojikī), with distinct northern and southern dialects, belongs to the southwest group of Iranian languages, in the Indo-European family; it is very close to the language used widely in neighbouring Afghanistan, where it is known as Farsī. The language of the Pamir Tadzhiks belongs to the east Iranian group. Tadzhik was formerly written in a modified Arabic and later in the Roman alphabet but since 1940 has used a modified Cyrillic script. Writers from this region have made notable contributions to Persian literature since the 10th century AD, and a vigorous folk literature is still being collected.

A number of Tadzhik poets and novelists have achieved fame in the Soviet Union and beyond. They include Abdalrauf Fitrat, whose *Last Judgment* has been reprinted many times in Tadzhik, Russian, and Uzbek, and Sadriddin Ayni, known for his novels *Slaves* and *Dokhunda* and for his autobiography, *Bukhara*. Abū ol-Qāsem Lāhūtī's poems *We Won* and *Crown and Banner*, and Mirzo Tursunzode's poem *Hasan the Arab* respond to the changes of the Soviet period; the latter's lyric cycle *The Voice of Asia* won him the Lenin Prize in 1960, followed by a number of other major awards noting his contribution to Soviet Tadzhik poetry and his influence on younger poets.

For centuries, songs, dances, and theatrical and circus performances have been a distinctive popular tradition in the region: singers, tightrope walkers, conjurers, musicians, and other performers continue to entertain audiences. In addition, festivals are now held for amateurs working in both traditional and modern fields.

The Tadzhik National Theatre, which was established in 1929, has nine separate theatres presenting opera, ballet, musical comedy, and puppetry. The cinema is developing, with Tadzhik studios producing feature films and documentaries as well as dubbing films from elsewhere. Radio and television services are increasing, and Dushanbe has had a television centre since 1960.

PROSPECTS

The Tadzhik republic has much potential for further economic development and associated improvements in the material and cultural life of its people. Industrial development will continue, a major role being played by electrical power production. This latter trend will include the completion of the final units of the Nurek hydroelectric station complex, which will enable additional capacities in the metallurgical and electrochemical industries to come into operation.

The Tadzhik people are among the most outward-looking of Central Asia, in part because so many of them live outside the Soviet Union (mainly in Afghanistan) and also because of affinities strongly felt with the people of Iran. The Soviet Tadzhiks' intellectual rivalry with their closest neighbours, the Turkic Uzbeks, has been encouraged by the Soviet authorities in order, among other things, to maintain the separateness of the Central Asian nationalities. These factors support Tadzhik national consciousness as the group continues to modernize its economy and society. Handicaps include lack of national sovereignty, which prevents Tadzhiks from taking ultimate responsibility for important developments affecting their future. Pressing problems remain in the fields of education, women's rights, and intellectual life, among others. Book publishing (per capita) in Tadzhik declined noticeably on the average, between 1959 and 1970, gains in influential Communist Party membership and posts are coming very slowly, and unemployment of educated Tadzhiks has become evident, though most Tadzhiks remain on farms or in country villages.

Despite these difficulties, the Tadzhiks should grow in strength and confidence as their population rapidly increases and their educational standing generally improves. Where the new generations of educated Iranian Central Asians will find suitable outlets, either in the region or outside it, for their talents and skills is a question without obvious answers in the mid-1970s.

BIBLIOGRAPHY. Various aspects of Tadzhik life are dealt with in the following works: JIRI BECKA, "Tajik Literature from the 16th Century to the Present," in JAN RYPKA, *History of Iranian Literature* (1968); TERESA RAKOWSKA-HARMSTONE, *Russia and Nationalism in Central Asia: The Case of Tadzhikistan* (1970); LAWRENCE KRADER, *Peoples of Central Asia*, 3rd ed. (1971); and EDWARD ALLWORTH (ed.), *The Nationality Question in Soviet Central Asia* (1973).

(A.I.I./E.Al.)

Tagus River

The Tagus (Spanish Tajo, Portuguese Tejo) is one of the great rivers of Europe. The longest river of the harsh and arid Iberian Peninsula, it has been of vital importance to the economic and social development of two nations. Its total length is 626 miles (1,007 kilometres), made up of 488 miles across Spain, a 27-mile section as the frontier between Spain and Portugal, and a final 111-mile stretch in Portugal until it reaches the Atlantic at Lisbon. Its drainage basin of 31,505 square miles (81,600 square kilometres) covers the heart of Portugal and Spain and is only slightly smaller than that of the Ebro, to the northeast.

Geographers and naturalists have been associated with the study of the river since the time of the Greeks and Romans. During the reign of the Spanish king Philip II in the 16th century, studies were made to determine the extent of the river's navigability for use during times of war. Philip III and IV continued the studies, sending various expeditions from Toledo to Portugal. The Tagus was found unsuitable, however, for military logistic purposes.

In 1828 a court navigator, Francisco Javier Cavanas, following the work of a predecessor, Juan Bautista Antonelli, two and a half centuries earlier, investigated the possibility of making the river a trade artery between Aranjuez and Portugal. It proved too expensive and, until the present day, only the lower part of the Tagus is used for navigation of any importance. By the 1970s there was an official Spanish agency, the Confederación Hidrográfica del Tajo, specializing in scientific study of the Tagus. Several Spanish ministries had departments concerned with the river. The interest of such private bodies at the Sociedad Española de Ciencias Naturales, the Real Sociedad Geográfica, and the Instituto Celestino Mutis further testified to the importance of the river.

The course of the river

The Tagus rises in the Sierra de Albarracín, near Casas de Fuente García (Teruel province) in eastern Spain, at an altitude of 5,216 feet (1,590 metres). The young stream cuts into the limestone rocks and flows through narrow, sinuous valleys with deep canyons and abundant ravines in a generally southwestward direction. Near Trillo (Guadalajara province) it runs through more solid soil and rock, and its course becomes generally more peaceful. Just before the town of Bolarque it is held back by the important dams of Entrepeñas and Buendía, forming an artificial lake known as the Sea of Castile, which covers an area of 51 square miles. Again flowing southwestward, the river irrigates the fertile lands of Aranjuez (famous for its strawberries), the area around Toledo (the beautiful Roman, Gothic, Jewish, Arab, and Christian city), and Talavera de la Reina until it reaches the province of Cáceres. Here it again flows through narrow,

steep-edged trenches (among them, Salto del Corzo, del Gitano, and del Macho) formed by quarzites and shales. In Puente de Alcántara a dam has formed the greatest man-made lake in contemporary Europe, with a length of 57 miles and with a capacity of 112,000,000 cubic feet. Less than a mile downstream the river becomes the frontier between Spain and Portugal for 27 miles, entering into Portugal at Beira and the Sierra de Carbajo, and then crossing the fertile region of Abrantes. At Vila Franca de Xira it forms the Tagus (or Lisbon) Estuary, emptying at this point into the Atlantic Ocean.

Centuries ago, as one can read in Pliny the Elder (1st century AD), the Iberian Peninsula was covered with luxuriant woods, but the destructive effects of constant wars changed the landscape into semi-arid lands. The Tagus Valley area was further damaged as recently as 1936 by battles fought there during the Spanish Civil War. As a result, all the present efforts of the public and private bodies associated with the river are dedicated to increasing reforestation, land irrigation, and the creation of hydroelectric power.

In the highlands of the Tagus Basin, conifer trees are numerous, with a well-developed timber industry. Near the smaller streams poplars and willows also grow. About a third of the cultivated area is covered by cereal farms and everywhere olive trees and vineyards are found. In Extremadura, in the west of Spain, the country is semi-arid; only oaks, cork trees, and holm oaks break the monotony of a rolling and rocky landscape.

Plant and animal life

Animal life near the river is abundant and diverse, with both European and North African species, often the subject of special attention by naturalists and conservationists. Royal carp, luces, and black-bass fishing is possible in the artificial lakes of Entrepeñas and Buendía, and trout, barbels, and many other kinds of fish may be captured in the Tagus and its tributaries. There is very important big game hunting in Gredos, noted for its *capra hispánica*, or Spanish ibex, and in the ridges of Cuenca and Guadalajara fallow deer and chamoises are found.

The hydroelectric and irrigational utilization of the Tagus had made great strides forward by the decade of the 1970s. In the whole drainage basin there were 62 dams with a total capacity of about 2½ cubic miles. The most important are: Alcántara; Buendía; Valdecañas; Gabriel y Galán; Entrepeñas. The installed power is of the order of 1,200,000 kilowatts. (J.B.Ro.)

Tai Languages

The name Tai denotes a family of closely related languages, of which the Thai, or Siamese, language of Thailand is the most important member. Because the word Thai has been designated as the official name of the language of Thailand, it would be confusing to use it for the various other languages of the family as well. Tai is therefore used to refer to the entire group.

Geographical distribution

Spoken in Thailand, Laos, Burma, Assam in northeast India, North Vietnam, and the southwestern part of China, the Tai languages together form an important group of languages in Southeast Asia. In some countries they are known by different tribal names or by designations used by other peoples; *e.g.*, there is Shan in Burma; Pai-i in Yunnan, China; Chuang-chia in Kwangsi, China; Chung-chia, Dioi, Jui, Yoi, Yay, or Pu-yi in Kweichow, China; Tho, Nung, White Tai, Black Tai, Red Tai, etc. in North Vietnam; and Khün, Lü, etc. in Thailand and Laos. Ahom, an extinct language once spoken in Assam, has a considerable amount of literature. The Tai languages are divided into three linguistic groups—the Southwestern, the Central, and the Northern. Thai and Lao, the official languages of Thailand and Laos, respectively, are the best known of the languages.

The number of Tai speakers is estimated at 30,000,000 in *Ethnic Groups of Mainland Southeast Asia* (1964). Of these, 18,000,000 are in Thailand, 9,000,000 in China, and approximately 1,000,000 each in Laos, North Vietnam, and Burma. There are tremendous variations between several estimates, and these figures may serve as only rough indications of the Tai populations.

Relationship to other languages. The Tai languages are traditionally assumed to be related to the Sino-Tibetan family of languages with which they share similar phonological structures, tone systems, and some lexical items; the relationship, however, has never been definitely established. The Tai tongues are related to the Kam-Sui languages of Kweichow, China, and to the Bê language in Hainan; but because of the great divergences in both vocabulary and phonological development, the name Tai is used only to designate a more restricted group that excludes such languages as Kam, Sui, and Bê. It has also been suggested, chiefly on the basis of some similar vocabulary items, that the Tai languages are related to the Kadai languages of southwestern China and Hainan Island and ultimately to the Austronesian (Malayo-Polynesian) languages, rather than to the Sino-Tibetan tongues.

Classification. Classifications have been made according to the geographical location of the Tai speakers, social, political, and cultural criteria, and literacy versus nonliteracy. The classification used for this article is based on linguistic relationships proposed in 1959–60; the criteria for it are lexical (involving similarities in vocabulary) and phonological (involving similarities in sounds and systems of sounds). According to these features the Tai languages are divided into the three groups mentioned above (see map). Languages of the Southwestern group are spoken in Thailand, Laos, North Vietnam, Burma, and Yunnan, China; they include Thai, or Siamese, Lao, Shan, Khün, Lü, White Tai, Black Tai, etc. The Southwestern division, geographically the most widespread group, consists of two-thirds of the Tai-speaking population and represents an expansion in comparatively recent periods. To the Central group belong the Tho dialects spoken in North Vietnam and the various dialects spoken in Kwangsi, such as Lungchow. The Chung-chia or Pu-yi dialects in Kweichow and the Chuang-chia dialects in Kwangsi belong to the Northern group. Some of the Northern dialects are also spoken in Yunnan and Vietnam, and one, called Saek, is spoken as far south as Laos and Thailand.

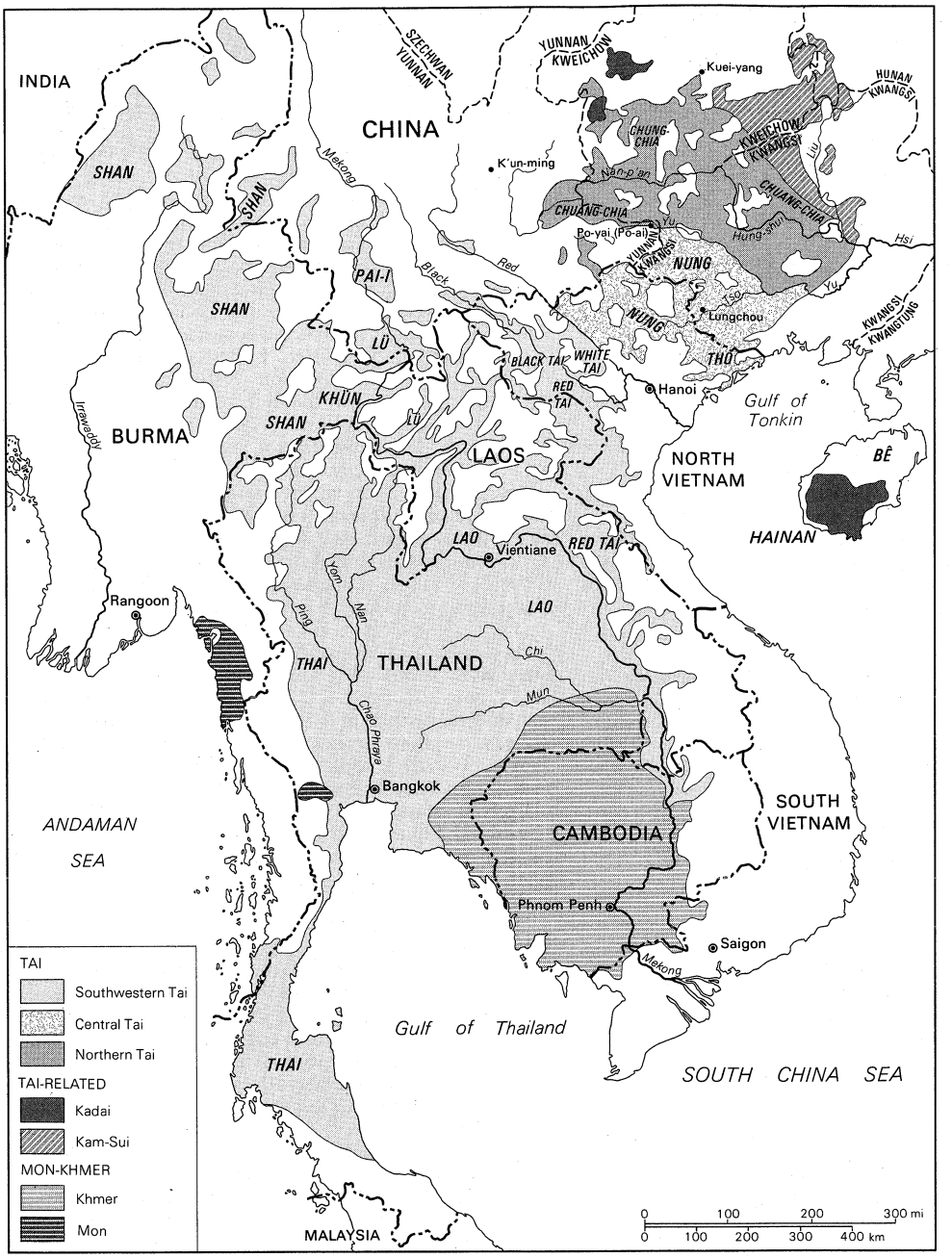
Main groups of Tai

The fairly large number of vocabulary items shared by languages of these three groups suggests their genetic relationship. There are also certain items that are shared by only two of the groups and not found in the other. For instance, the word for sky is shared by the Southwestern dialects (Siamese *fáa*) and the Central dialects (Lungchow *faa*), but another word is used in the Northern dialects (Po-ai *mïn*). Similarly, the word for beard is shared by the Central group (Lungchow *mum*) and the Northern group (Po-ai *mum*) but is replaced by another word in the Southwestern group (Siamese *nùat*); the term for knife is shared by the Southwestern (Siamese *mîit*) and the Northern groups (Po-ai *mit*) but not by the Central dialects. There are also vocabulary items that are found only in one group. In all, the evidence seems to indicate that there are three groups of dialects in the Tai family.

Dialect groups

Different phonological features may be attached to some words according to the dialect group. For instance, the Southwestern forms for the verb "to be" (Siamese *pen*) are derived from a protoform **pɛn* (with the vowel like *e* in "egg"), whereas the Central dialect forms (Lungchow *pin*) and the Northern forms (Po-ai *pan*) come from a protoform **bɛn*, as indicated by the tone. (A protoform is the presumed or reconstructed ancestral form of a word; an asterisk [*] marks those forms that are unattested and reconstructed.) Similarly, the Southwestern and the Central forms for the classifier for animals (Siamese *tua*, Lungchow *tuu*) are derived from a protoform **tua*, whereas the Northern forms (Po-ai *tuu*) are attributed to a protoform **dua*. (A classifier is a term that indicates the group to which a noun belongs, like "animate object," or designates countable objects or measurable quantities, like "yards" [of cloth] and "head" [of cattle]). Such words as the forms for "to be" and the classifier for animals are good indications of dialect boundaries.

In phonological development, the Northern dialects differ from the rest in not maintaining the distinction between aspirated and unaspirated voiceless stops. That is, the dialects have lost the feature of aspiration, which



Major divisions of the Tai languages and neighbouring languages.
Adapted from Frank M. LeBar, Gerald C. Hickey, and John K. Musgrave, *Ethnic Groups of Mainland Southeast Asia;* New Haven, Human Relations Area Files Press, 1964

sounds like a puff of breath accompanying a consonant. Aspiration may, however, be reintroduced in some dialects by later borrowing or secondary developments. The Central dialects differ from the other groups in the treatment of certain Proto-Tai consonant clusters, such as **tr-* and **thr-*. Although they have changed from the proto-forms, these are usually kept distinct in the other groups—*e.g.*, in Siamese as *taa* ("eye") and *haaŋ* ("tail"), in Po-ai as *taa* and *lïïŋ*. In the Central dialects, however, they have merged into a single sound—*e.g.*, Tho *thaa* and *thaaŋ*, Lungchow *haa* and *haaŋ*.

Phonological characteristics. The sound system of the Tai languages may best be described in terms of its syllabic structure. Each syllable consists of an initial consonant or consonant cluster followed by a vowel or vowel cluster (long vowel or diphthong), which may be further followed by a final consonant, usually a nasal sound or an unreleased stop. (An unreleased stop is a consonant in which there is complete stoppage of the airstream from the lungs, and in which the tongue or the lips maintain the position of the consonant without opening the stop-page.) In addition, each syllable has a tone. As an illustration of this type of structure, the system of the Thai, or standard Siamese, language as spoken around the Bangkok area may be given. Labials are sounds using the lips as articulators; alveolars are formed by placing the tip of the tongue at the gums behind the upper teeth; sibilants, or fricatives, involve local friction and a hissing sound resulting from incomplete closure in the mouth. Velars are sounds made with the back of the tongue touching the soft palate (velum); glottal sounds involve constriction of the vocal cords; semivowels are gliding sounds with some of the properties of both vowels and consonants; and liquids are frictionless consonants, produced with incomplete closure in the vocal tract, that sometimes function as vowels.

Five tones

There are five tones: level (*a*), low (*à*), falling (*â*), high (*á*), and rising (*ǎ*); for example, *maa* "to come," *màak* "areca nut," *mâak* "much," *máa* "horse," and *mǎa* "dog" are differentiated by the various tones.

Grammatical characteristics. The statements about morphology and syntax refer particularly to the standard language of Thailand, though they are applicable in gen-

Table 1: The Sounds of Thai (Siamese)

	initial consonants*					final consonants	
Labials	p	ph	b	m	f	p	m
Alveolars	t	th	d	n		t	n
Sibilants	c	ch			s		
Velars	k	kh		ŋ		k	ŋ
Glottals	ʔ				h	ʔ	
Semivowels		j		w			j w
Liquids		l		r			

	vowels			long vowels			diphthongs		
High	i	ï	u	ii	ïï	uu	ia	ïa	ua
Mid	e	ə	o	ee	əə	oo			
Low	ɛ	a	ɔ	ɛɛ	aa	ɔɔ			

*Certain consonant clusters are also permitted initially, such as *kl, kr, khl, khr, tr, pl, pr, phl, phr, kw,* and *khw.*

eral to all of the Tai languages. Words or morphemes (word elements) are, for the most part, monosyllabic, but there are also many polysyllabic words, mainly compounds and loanwords from Sanskrit and from Khmer, an Austroasiatic language spoken in Cambodia. There are no inflections in Tai comparable to *-ed* in English "walked" or *-s* in "dogs," but derivations are common. Derivation is the formation of new words by the addition of endings, prefixes, or other words, often resulting in a change of the part of speech or meaning of a term—*e.g.*, English "unionize" or "unionization" from "union." Compounding is the chief Tai method of derivation—*e.g.*, *nâa-taa* "countenance (face-eye)," *kèp-kìaw* "to harvest (gather-cut with a sickle)." Reduplication, the repetition of a word or part of it, is common—*e.g.*, *dii-dii* "very good" from *dii* "good." Partial reduplication is also found, such as *sanùk-sanăan* "to enjoy oneself" from *sanùk* "to have fun." Prefixes and infixes occur often in Indo-Aryan and Khmer loans—*e.g.*, *pra-thêet* "country" (from Sanskrit *pradeśa*) and *d-amn-əən* "to proceed" (from Khmer *dae* "to walk") and *d-amn-aə* "process." There also developed some native prefixes that are abbreviated forms of what once were full words. For example, *ma-*, a prefix used in many names of fruits, such as *ma-phráau* "coconut," *ma-mûaŋ* "mango," is derived from *màak* "areca nut" (originally "fruit"). Similarly, in the word *sa-dïï* "navel," *sa-* is the reduced form of *săaj* "line, string," which refers to the umbilical cord. Old processes of derivation involved using the alternation of consonants or of tone or both, such as *níi* "this" and *nîi* "here," *nɔ́ɔj* "small, little" and *nɔ́ɔj* "a little bit," *khiaw* "sickle" and *kìaw* "to cut with a sickle." Such processes are, however, no longer active.

A sentence is usually formed by a noun phrase (subject) followed by a verb phrase (predicate). There are, however, nominal predicates as well, in which the verb is lacking and a noun is used instead; an example is *wan-níi wan-săw* "today (is) Friday." A noun phrase consists of a noun, which may be followed by its modifiers (another noun, an adjective, or a verb phrase), which, in turn, are followed by a numeral with a classifier, and finally, by a demonstrative. For example, *tûu* "cabinet" and *náŋsïï* "book" become *tûu-náŋsïï* "bookcase"; *tûu-kèp náŋsïï* is "bookcase" ("cabinet-keep-book"); *tûu-yen* is "icebox" ("cabinet-cool"), and *tûu-náŋsïi sɔɔŋ-bai níi* is "these two bookcases" ("cabinet-book two-classifier these"). Classifiers must be used with numerals and vary according to the nouns they are to enumerate. It is not uncommon that a noun may serve as its own classifier.

Use of nominal predicates

A verb phrase may consist of an adjective or of a verb often followed by its object or its complements or both. A common type of complement denotes the direction of action by using such words as *paj* "to go," *maa* "to come," *khĭn* "to ascend," *loŋ* "to descend," and so on. Examples are *ʔaw náŋsïï paj* "Take the book away!"; *ʔaw náŋsïï maa* "Bring the book (here)!"; *aw náŋsïï khĭn maa* "Bring the book up!"; and *aw náŋsïï loŋ paj* "Take the book down!". A number of particles, particularly those occurring at the end of a sentence, are used to indicate a question or a command, to express emphasis or uncertainty, to show politeness and the sex of the speaker, etc. For example, *kháw maa lɛ́ɛw* "He has come"; *kháw maa máj*? "Is he coming?"; *khun maa máj khráp*? "Are you coming?" (man speaking, polite); *khun maa máj khá* "Are you coming?" (woman speaking, polite).

Vocabulary. The various Tai languages have been in contact with many different languages spoken in the same area of Southeast Asia, and it is inevitable that words from different sources have been adopted by them. There is a basic core of vocabulary that is shared by most of the Tai languages and may be considered as the native vocabulary of the protolanguage (earlier form of the Tai lan-

Table 2: The Thai Alphabet

consonants																vowels*	
middle			high			low										short	long
ก k			ข kh			ค kh	ฆ kh	ง ŋ								เ–ะ -a?	–า -aa
จ c			ฉ ch			ช ch	ฌ ch	ญ j								–ั/ -a-	
ฎ d	ฏ t		ฐ th			ฑ th	ฒ th	ณ n								–ิ -i(?)	–ี -ii
ด d	ต t		ถ th			ท th	ธ th	น n								–ึ -ï(?)	–ือ -ïï
บ b	ป p		ผ ph			พ ph	ภ ph	ม m								–ุ -u(?)	–ู -uu
			ฝ f			ฟ f										เ–ะ -e?	เ– -ee
			ศ s		ษ s	ย j	ร r	ล l		ว w						แ–ะ -ɛ?	แ– -ɛɛ
			ส s			ซ s										โ–ะ -o?	โ– -oo
อ ʔ			ห h			ฮ h	ฬ									– -o-	
																เ–าะ -ɔ?	–อ -ɔɔ
vowel clusters†																ไ– -aj	–า ย -aaj
–ัว -ua			–ว/ -ua-													ใ– -aj	
เ–ียะ -ia			เ–ีย -ïa													เ–า -aw	–า ว -aaw
เ–อ -əə			เ–ิ/ -əə-													ฤ rï(?)	ฤๅ rïï
เ–ย -əəj			–ำ -am													ฦ lï(?)	ฦๅ lïï

tonal markers (used with vowel symbols)

(1) – middle tone with middle and low consonants; rising tone with high consonants
(2) ˋ low tone with middle and high consonants; falling tone with low consonants
(3) ˇ falling tone with middle and high consonants; high tone with low consonants
(4) ˊ rising tone chiefly with middle consonants

*Pronunciation key (for symbols not used in the standard Roman alphabet): ɛ = *a* as in *bad;* ï = the *u*-sound in some pronunciations of *just* (*e.g.*, "just a minute"); ɔ = *o* as in *law;* ə = *a* as in *sofá;* ʔ = glottal stop (the catch sound for *tt* in Cockney or Brooklynese *bottle*). †For which the long-short distinction does not apply.

guages). Nevertheless, items in this list may be found to resemble forms in other languages, such as Chinese or Indonesian, and they have given rise to different interpretations of the relationship of the Tai languages to other linguistic families. Loanwords in later periods from the Khmer and Indo-Aryan languages (Sanskrit and Pāli) are particularly common in Siamese and Lao, and loans from Chinese are abundant in the Central and the Northern dialects.

Loan words

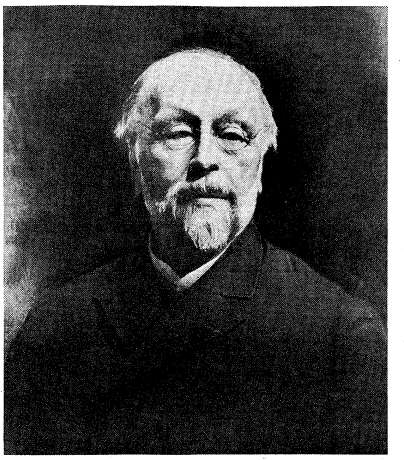
Writing. There are two kinds of writing used among the Tai languages. One, ultimately derived from Chinese, is used by the Central and the Northern dialects; the other comes from Indo-Aryan sources and is used in many languages of the Southwestern group. The Chinese-based system, employed chiefly to write songs, consists of both Chinese characters and modified Chinese characters, very much like the early writing in Vietnam. A specimen of this type of writing dates back perhaps to the 18th century, but it may have been in use much earlier. The script adapted from Indo-Aryan sources for the Tai languages dates perhaps from the 13th century. The earliest known example of such a writing system is the inscription of Ramkhamhaeng in northern Thailand from AD 1293. The Modern Thai alphabet (see Table 2) is a modified form of the original writing. It preserves the old distinction of voiced (low) and voiceless (middle or high) consonants, a distinction that is now lost, but leaves its effects on the tone. This system also provides an unambiguous method for indicating the vowels and tones. Similar types of writing are used in Lao, Lü, White Tai, Black Tai, etc. Another type of Indo-Aryan writing, used in Shan and Ahom, lacks the difference between high and low consonants, has an insufficient number of vowel signs, and provides no tone marks. Although the shapes of the letters in Shan differ greatly from those in Ahom, the basic principle is essentially the same.

BIBLIOGRAPHY. M.R. HAAS, *Thai-English Student's Dictionary* (1964), a concise dictionary arranged according to the order of the Thai alphabet with a brief description of the phonological and grammatical system, and with H.R. SUBHANKA, *Spoken Thai*, 2 vol. (1947), a textbook for learning spoken Thai step by step; F.M. LABAR, G.C. HICKEY, and J.K. MUSGRAVE, *Ethnic Groups of Mainland Southeast Asia*, pt. 3, pp. 187–244 (1964), a detailed description of different Tai ethnic groups, including their location, demography, and social organization; FANG KUEI LI, "Consonant Clusters in Tai," *Language*, 30:368–379 (1954), an article demonstrating the existence of a large number of initial consonant clusters in the protolanguage and their simplification in the modern dialects; "A Tentative Classification of Tai Dialects," in S. DIAMOND (ed.), *Culture in History*, pp. 951–959 (1960), a classification establishing three groups of Tai languages and dialects according to lexical distribution and phonological development, and "The Tai and the Kam-Sui Languages," *Lingua*, 14:148–179 (1965), a discussion of the relationship of these two language groups; R.B. NOSS, *Thai Reference Grammar* (1964), a detailed descriptive and structural analysis of the Thai language; P.K. BENEDICT, "Thai, Kadai and Indonesian: A New Alignment in Southeastern Asia," *American Anthropologist*, 44:576–601 (1942), a provocative attempt to show the relationship of the Tai languages to the Indonesian languages.

(F.K.L.)

Taine, Hippolyte

Hippolyte-Adolphe Taine achieved fame over a wide range of disciplines—as a leading French thinker, as a literary and art critic, and as a historian. His greatest influence upon his contemporaries, however, was as an intellectual leader, one of the most esteemed exponents of 19th-century French Positivism, the cult of science in its most devoted, high-minded, and rational form. His work, a reaction against spiritualist philosophy, was unified by his attempt to apply the scientific method to the study of literature and art, to psychology, to cultural history, and to the realms of metaphysics and ethics. He even aspired to base a new religion upon science, and his ideas helped to provide a theoretical basis for the literary movement of Naturalism; the novel, he argued, should be a "collection of experiments," a contribution to the scientific understanding of human nature, revealing, like the new scientific psychology he advocated, the physiological and psychological determinants of man's behaviour.



Taine, portrait by Léon Bonnat, 1889.
Hachette—J.P. Ziolo

Born on April 21, 1828, at Vouziers, in the Ardennes, Taine came from a professional middle class family; his father was a lawyer. He was educated privately at home until shortly after his father's death; thereafter, he went with his mother to live in Paris and became an outstanding pupil at the Collège Bourbon and then at the highly prestigious École Normale. He gained his *licence-ès-lettres* (preliminary degree) in 1848 and began to study for his *agrégation* (advanced degree) in philosophy, one of his dominant interests. He already held unorthodox intellectual views. He had apparently lost his Christian faith by the age of 15, and his youthful rationalist attitude led him to admire the ideas of the Idéologue philosophers who held that all knowledge must be based on sense experience, on observation and controlled experiment; this overriding conviction guided his later career. He was also already attracted by the metaphysical ideas of Hegel and Spinoza, which inspired in him a desire to find a total explanation of the causal forces of life and the universe. In contrast to these views, his new teachers of philosophy in Paris held the prevailing philosophical doctrine of eclecticism; consequently—and not without creating some scandal in academic circles—Taine's *agrégation* jury failed him in 1851. He then taught for brief periods at Nevers and Poitiers but in 1852 applied for leave of absence. Returning to Paris, he devoted himself to preparing his two dissertations for the doctorate in literature: *De Personis Platonicis* ("Concerning Plato's Characters") and, his first well-known work, a study of La Fontaine (1853; revised and published in 1861 as *La Fontaine et ses fables* ["La Fontaine and His Fables"]). He gained his doctoral degree in May 1853 and began an essay on Livy, *Essai sur Tite-Live* (1856), which, despite further criticism of his philosophical outlook, won a prize from the Académie Française. During this period he was also attending lectures in science and gathering the knowledge of physiology that he utilized later in his work on psychology. Reluctant to return to full-time teaching, he lived by private tutoring and as a man of letters. Even a holiday in 1854, necessitated by ill health, was turned to advantage: in 1855 he published a literary guide book based on his travels, *Voyage aux eaux des Pyrénées* ("Voyage to the Waters of the Pyrenees"). More important for his own development, he contributed frequent literary and historical articles to such leading journals as the *Revue des Deux Mondes*, the *Revue de l'Instruction Publique*, and the *Journal des Débats*, articles that provided the basis for three books further enhancing the reputation he had gained by his works on La Fontaine and Livy. These were *Les Philosophes français du XIXe siècle* (1857; "The French Philosophers of the 19th Century"), a critical polemic against the prevailing eclectic philosophy of Victor Cousin and his group, which also provides in its later chapters a lucid exposition of his own Positivist theory of knowledge; a

Attack on eclecticism

first collection of *Essais de critique et d'histoire* (1858; "Essays of Criticism and History"); and his notable *Histoire de la littérature anglaise* (4 vols., 1863–64; Eng. trans., 1871, *History of English Literature*), whose celebrated "Introduction" gives a succinct statement of his approach to literary and cultural history and a basic text for the understanding of his scientific attitude to literary criticism. The same great causal factors underlie any cultural artifact of a given age and society, he claims. By studying the literary documents the psychology of their author can be understood, and this, complemented by scrutiny of the facts of his life and personality, illuminates the "*faculté maîtresse*," the predominant characteristic that determines his work; this, in turn, can then be "explained" by reference to three great conditioning facts, "*la race*," "*le milieu*," and "*le moment*"; *i.e.*, the writer's inherited personality, his social, political, and geographical background, and the historical situation in which he writes. It is evident that Taine's interest here is less in literature itself than in historical causation and psychology, and his method may well be thought to have encouraged in his admirers an excessive preoccupation with biography and literary history at the expense of critical judgment, though Taine's own abilities as a critic were considerable.

Throughout the 1860s Taine continued indefatigably his researches and his writing. Even his travels (to England, Italy, Germany, and The Netherlands) were utilized to gather notes for future work—for example, his closely observed if simplifying *Notes sur l'Angleterre* (1872; Eng. trans., *Notes on England*, 1872); and even his life in Paris led to his *Notes sur Paris: Vie et opinions de M. Frédéric-Thomas Graindorge* (1867; Eng. trans., *Notes and Opinions of Mr. Frédérick-Graindorge*, 1875), perhaps the most personal and entertaining of his books. In 1864, by a happy decision of Napoleon III, he was appointed to succeed Viollet-le-Duc, the architect, as professor of aesthetics and of the history of art at the École des Beaux-Arts in Paris, where he lectured for 20 years. The lecture courses, which he eventually published, include *Philosophie de l'art* (1865; Eng. trans., *The Philosophy of Art*, 1865), *De l'idéal dans l'art* (1867; "On the Ideal in Art"), and those concerned with the philosophy of art in Italy (1866), The Netherlands (1868), and Greece (1869). This post also gave him a security that favoured his more protracted scientific studies and helped make the later 1860s a happy and fertile period in his life. He published, in addition to the works named, his second volume of essays, *Nouveaux essais de critique et d'histoire* (1865; "New Essays of Criticism and History"), including his perceptive articles on Racine, Balzac, and Stendhal (whose psychological acuity he was one of the first to admire). In 1868 he married Mlle Denuelle, the daughter of a well-known architect and artist, by whom he had a son and a daughter. In 1870 he published the two volumes of *De l'intelligence* (Eng. trans., 1871, *On Intelligence*), a major work in the discipline of psychology, which had interested him since his youth. His devotion to science is most fully illustrated here; he opposes the speculative and introspective approach of the eclectics and outlines a scientific methodology for the study of human personality that established him, alongside thinkers such as Théodule Ribot and Pierre Janet, as a founder of empirical psychology. Though much of the work is now outdated, in its day it helped to modify methods of research by its emphasis on experiment, the search for causes, the study of pathological cases, and the physiological basis of personality. It also intensified opposition to his ideas, and he was angrily accused of holding a strictly determinist and materialist view of man—not altogether unfairly, even though he claimed to reject materialism and argued that moral responsibility was compatible with determinism as he conceived it. The work also develops his long-standing attempt to fuse Positivism and Hegelian Idealism and to provide a method for a scientific metaphysics. Through such a metaphysics, he maintained, the final causes of life itself might be discovered; its insights inspired him to an exalted pantheistic trust in nature that is movingly expressed in essays on Marcus Aurelius (in *Nouveaux essais*) and Iphigeneia (in *Derniers essais*).

Publication of *De l'intelligence*

Germany's invasion and defeat of France in 1870–71 had a profound impact upon Taine (already prepared in his mind by a visit in 1869 that had disabused him of his earlier enthusiasm for German civilization). The French defeat, in his view, sprang from a deep national sickness, and he determined to devote his final years to examining its causes. A shift of interest toward politics is illustrated by a brochure of 1872 on the problems and effects of universal suffrage, but, above all, his approach was historical: to seek the sources of the political instability that he held responsible for his country's plight. This major reorientation of concern led to his great historical work, *Les Origines de la France contemporaine* ("The Origins of Contemporary France"), a monumental analysis, claiming scientific objectivity (although its factual and interpretative reliability have been challenged). It seeks to show that France's primary fault lay in excessive centralization, originating during the *ancien régime*, and intensified by the French Revolution, about which he shares and develops Burke's hostile view. A first volume, on *L'Ancien Régime* ("The Old Regime"), appeared in 1876, followed by three volumes on the Revolution (1878–85). In 1878 he was also elected to the Académie Française. To have more time for his self-appointed task he withdrew increasingly from Paris and after 1883 even resigned his professorship to live and work, surrounded by family and friends, in his house on the lake of Annecy at Menthon-Saint-Bernard, in the peaceful natural setting he had always loved. Only one volume of *Le Régime moderne* ("The Modern Regime"), however, was published in his lifetime (1891), the second volume coming out in November 1893. The entire work was reissued in 1899. There also appeared after his death his *Derniers essais de critique et d'histoire* (1894; "Last Essays of Criticisms and History") and an unfinished autobiographical and psychological novel, written about 1861, *Étienne Mayran* (1910). He died in Paris on March 5, 1893, and was buried at Menthon-Saint-Bernard, honoured as an outstanding scientific humanist of his time.

Historical theories

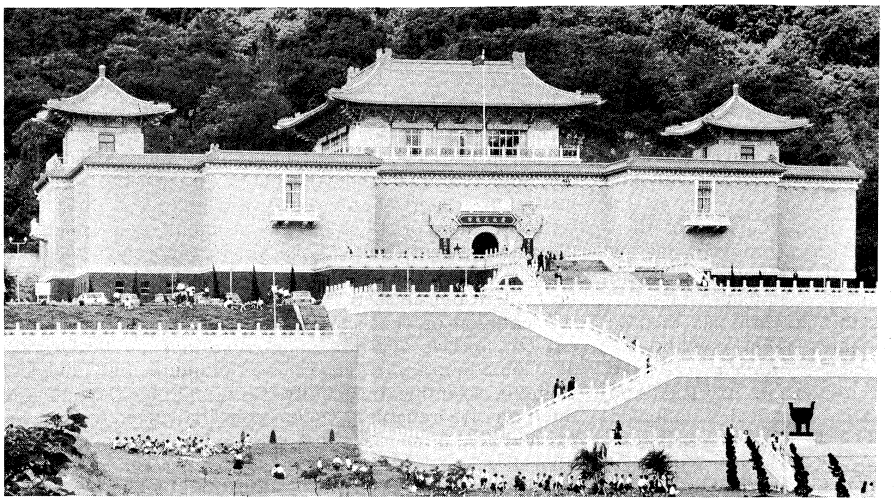
BIBLIOGRAPHY. An authoritative biography and a selection of Taine's correspondence was edited by his widow under the title *H. Taine, sa vie et sa correspondance*, 4 vol. (1902–07; Eng. trans., *Life and Letters of H. Taine*, 3 vol., 1902–08). Most of Taine's works have been translated into English. Useful monographs that include a survey of his life include V. GIRAUD, *Essai sur Taine, son oeuvre et son influence*, 6th ed. (1923); A. CHEVRILLON, *Taine: formation de sa pensée* (1932); and A. CRESSON, *Hippolyte Taine, sa vie, son oeuvre* (1951). In addition, studies of his philosophy include G. BARZELLOTTI, *La Philosophie de H. Taine* (1900); D.D. ROSCA, *L'Influence de Hegel sur Taine, théoricien de la connaissance et de l'art* (1928); and M. LEROY, *Taine* (1933), in French. For a critical study of his historical work, see A. AULARD, *Taine, historien de la Révolution française* (1907). His literary criticism is fully discussed by S.J. KAHN in *Science and Aesthetic Judgment: A Study in Taine's Critical Method* (1953); and his influence in England by F.C. ROE, *Taine et l'Angleterre* (1923). Shorter studies in English are found in D.G. CHARLTON, *Positivist Thought in France During the Second Empire, 1852–1870* (1959) and *Secular Religions in France, 1815–1870* (1963); and in G.P. GOOCH, *History and Historians in the Nineteenth Century*, 2nd ed. (1952).

(D.G.Ch.)

Taipei

Taipei (Tai-bei in Pin-yin romanization) is the provisional capital of the Republic of China and a principal city of the island of Taiwan, located in the Pacific, 121 miles from the southeast coast of China. The city became the capital in exile of the Nationalist government after the victories of the Communists on the mainland in 1949.

Taipei is situated in the north of Taiwan, surrounded on all sides by mountains. About 19 miles to the northeast is Chi-lung, one of the main ports of Taiwan. Taipei was also the capital of Taiwan Province until 1957, when the capital was moved to Chung-hsing-hsin-ts'un near the west coast. The population of the city in 1972 was about 1,865,000, of whom 40 percent had migrated from the mainland after the Communist victories.



The National Palace Museum, Taipei.
George Holton—Photo Researchers

Recent growth

In the 19th century Taipei became an important centre for overseas trade. Oceangoing junks could reach it by river from the ports of Chi-lung and Tan-shui. The Japanese acquired the island in 1895 after the Sino-Japanese War and made Taipei the capital. The island reverted to China in 1945, after Japan's defeat in World War II. With the arrival of the Nationalist government from the mainland in 1949, Taipei expanded greatly. In 1967 it was declared a special municipality and given the status of a province; the total area was increased from about 26 square miles (67 square kilometres) to about 105 square miles (272 square kilometres) with the absorption of the towns and villages of Pei-t'ou, Shih-lin, Nei-hu, Nan-chiang, Mu-cha, and Ching-mei. The metropolitan area may be considered to include the whole northern part of the island, including the ports of Chi-lung and Tan-shui.

Taipei and its environs are now a major industrial area. Most of the important textile factories are located there; other industries include electronic parts, electric machinery and appliances, wires and cables, canned goods, refrigeration, shipbuilding (including yachts and other pleasure craft), motorcycles, rubber goods, and various handicrafts.

Railways and bus lines connect Taipei with all parts of the island. Within the city there are both municipal and private bus lines. While the railroad station is still the hub of the city, much new construction has occurred in the area of the Presidential Building and the Kuomintang headquarters. Broad new boulevards radiate from there to all parts of the city. The Sung-shan International Airport on the outskirts is served by a number of international airlines.

The older buildings in Taipei are mostly low, unpainted, wooden structures. In the 1960s many of these were replaced with high-rise apartment houses and office buildings. Because of the population influx and the priority given to office and industrial construction, there has been an acute shortage of housing.

The climate is subtropical and moderate. Temperatures in the summer rarely go over 95° F (35° C), and there is usually a sea breeze in the evening. Winter temperatures are seldom below 59° F (15° C). Rainy seasons are from March to June and November to January or February. The mean annual rainfall is about 83 inches 2,112 millimetres). Because of its mild climate, Taipei is a green city throughout the year.

The city and the island have excellent public-health services and are therefore free of some of the diseases endemic to most tropical and subtropical countries. The annual death rate in Taipei was reduced from 12.5 per 1,000 in 1947 to 3.2 per 1,000 in 1970. The average life expectancy is about 65 years for males and 69 for females.

In 1968 a nine-year course of education became compulsory. Taipei had almost 100 primary schools in 1971, as well as about 40 high schools and a dozen trade schools. Taiwan had over 90 institutions of higher learning, of which about 20 were located in Taipei.

The National Palace Museum

While Taipei is not a place of antiquity, it is a good place to view antiquities. The National Palace Museum, named after its original in Peking, houses one of the world's largest collections of ancient Chinese artifacts, calligraphy, paintings, and porcelain, brought over from the mainland after 1949.

One of the most popular recreation areas is Yang-min Shan, only ten miles from the centre of the city. Both the mountain and the town of Pei-t'ou at its base are known for their hot springs. Pi-t'an (Green Lake), seven miles from the city, has boating and water sports. There are ocean beaches not far away, and Tan-shui has two excellent golf courses.

There is a remarkable number of Christian churches and denominations in Taipei, which came there after being driven from the mainland. They range from the American Southern Baptist Mission to the Watch Tower Bible and Tract Society. There are also a mosque and many Buddhist and Taoist temples and nunneries.

BIBLIOGRAPHY. Information on Taipei may be found in the *China Yearbook* and various statistical publications on Taiwan. See also DAN CHEN, "City in a Hurry," *Free China Review*, 17:21–27 (1967), a review of its political, social, and economic conditions; and TAIPEI, FU-MIN GEOGRAPHICAL INSTITUTE OF ECONOMIC DEVELOPMENT, *Research Report* (1948–).

(H.-p.T.)

Taira Family

One of the houses that most influenced Japanese history, the Taira family, also known by the Sinicized pronunciation as Heike ("the house of Taira"), were essentially samurai, warriors whose climb to power was accelerated by their function of providing troops for the Imperial guards or military and police supervision of outlying districts under Imperial rule.

Origins and first period of power. The clan had its origins in 825, at a time when government finances were at a low ebb and members of the Imperial line were numerous. In an attempt to eliminate some of the drain on the finances, collateral Imperial branches were given surnames (the Imperial family had none) and sent out into the provinces. The name of "Taira" was given to Prince Takamune, the son of Prince Kuzuhara and grandson of Kammu, the 50th emperor. His descendants were accordingly called Taira of Kammu. Takamochi, a nephew of Takamune, arrived in the Hitachi district (about 40 miles [60 kilometres] northwest of present-day Tokyo) as a local official and settled there. His descendants succeeded him in the post, and the family became powerful samurai in the district.

Taira Masakado

A great-grandson, Taira Masakado, acquired great power and soon governed the whole Kantō district. In 939, he established a government in the southern part of Kantō,

styling himself *shinnō* ("new emperor") in opposition to the Emperor in the capital at Kyōto, but was subdued in 940. In 1028, when Taira Tadatsune attempted to re-establish Taira domination over the Kantō, the court dispatched another warrior, Minamoto Yorinobu, to quell the rebellion, and three years later, Tadatsune surrendered. As a result the Taira family began to decline, and the Minamoto family, descendants of Seiwa, the 56th emperor, organized a big samurai group in Kantō, with the Tairas under them.

Second era of power. In later years the Fujiwara family, who, sharing power with the emperor, had monopolized the highest posts in the court from the mid-10th to the mid-11th century, began to decline. In the latter half of the 11th century, the emperor Shirakawa abdicated the throne in favour of his son and then introduced a new political system called *insei,* by which the former emperor, who was now freed from the ceremonial requirements of the Imperial office (but could count on the loyalty of his son the real emperor), was finally able to wrest the power of the throne away from the Fujiwaras. So as to retain absolute power, the former emperor Shirakawa summoned Taira Masamori, a descendant of the Taira of Kantō with considerable local power in the Ise (present-day Mie Prefecture) district, to suppress the Minamoto family, whose military strength had been helping to ensure the dominance of the Fujiwaras at court. Masamori's success was so absolute that he stood high in the former emperor Shirakawa's favour and won speedy promotion as a court official.

Masamori's son Tadamori continued his father's successes. By eliminating the pirates along the Inland Sea in western Japan, he curried Imperial favour.

Taira Kiyomori

Taira Kiyomori, the son of Tadamori and grandson of Masamori, continued to enlarge the family's holdings and to increase its influence at court, making a conflict between the Tairas and Minamotos inevitable. Finally in 1156 a dispute over control of the court between two brothers, the former emperor Sutoku and the reigning emperor Shirakawa II, resulted in the Hogen War between Kiyomori and the head of the Minamotos. Aided by the defection of a group of Minamoto warriors, Kiyomori emerged victorious. Three years later in the Heiji War of 1159, Kiyomori brutally eliminated those Minamotos who had sided with him in the Hogen War. After that Kiyomori became the most powerful figure in Japan. To complete his career he was in 1167 finally given the highest position of any court official, *dajo-daijin* ("prime minister"). Kiyomori also managed to marry his wife's younger sister to the former emperor Shirakawa II, whose son became the prince. Kiyomori thus became even more powerful than the former emperor.

The Taira family monopolized high positions as court officials, governing almost half of all the provinces and owning more than 500 manors. Kiyomori also made huge profits from trade with China. In 1179, the court nobles led by the former emperor Shirakawa II rebelled against him but were subdued, and Shirakawa II was imprisoned. As a result, Kiyomori's grip became positively dictatorial, the period being known as the "Rokuhara regime" since he lived at Rokuhara in Kyōto. In spite of his great powers, however, he failed to make any basic changes in the Imperial system. As a result, the Taira hold over the countryside weakened as the family became accustomed to the rich court life and lost touch with the provincial warrior groups.

Rise of the Minamoto family

In 1181, when Kiyomori died after an illness, movements were started against the tyrannical Taira clan all over the country. The strongest opposition was that of the Minamoto family, whose scion, Minamoto Yoritomo, living in Kantō, had been spared in the great conflicts of 1159 because of his extreme youth at the time. With the assistance of other samurai hostile to the Tairas, he rose in arms. When the Minamoto army advanced on the capital, the Tairas escaped from Kyōto and, taking the young emperor Antoku with them, attempted to establish themselves in their stronghold in western Japan. But they were defeated in two successive battles, one at Ichinotani, west of the city of Kōbe in Settsu Province, and the other at Yashima Island, along the Inland Sea in Sanuki Province (present Kagawa Prefecture). Forced to flee further west, the Taira family was finally completely destroyed in 1185 in the great sea battle of Dannoura, which occurred off the eastern end of the strait that separates Kyushu from Honshu. In this battle the emperor Antoku drowned, taking with him the great sword that was one of the Three Imperial Regalia, the symbols of divine authority that had supposedly been brought to Japan when the first emperor descended from heaven.

BIBLIOGRAPHY. For more information on the Taira family, see *The Heike Monogatari,* trans. by A.L. SADLER, *Transactions of the Asiatic Society of Japan,* vol. 46, pt. 2, and vol. 49, pt. 1 (1918–21); and CARL SPOHR, *Gempei: The Civil Wars of Old Japan* (1967).

(K.Ho.)

Taiwan

Taiwan (T'ai-wan in Pin-yin romanization) is an island off the southeast coast of the China mainland. Since 1949 it has been the seat of the government of the Republic of China (Nationalist China), which is located in the island's capital city of Taipei. In addition to the main island of Taiwan, Taiwan comprises 15 islands in the Taiwan group and 64 islands in the Pescadores Archipelago. The total area of Taiwan, including the islands, is 13,892 square miles (35,981 square kilometres). Of the total, the Taiwan island group accounts for 13,843 square miles, while the Pescadores Archipelago, located to the west of Taiwan, has an area of 49 square miles.

Taiwan is bounded to the north by the East China Sea, to the east by the Pacific Ocean, to the south by the Bashi Channel, which separates it from the Philippines, and to the west by Formosa (Taiwan) Strait, which separates it from the China mainland. Taiwan forms part of the great island system rimming the western Pacific Ocean. The origin of its Chinese name T'ai-wan (conventionally Taiwan) is not known; it is also often known as Formosa—meaning beautiful—a name given it by Portuguese explorers.

Since 1949 the Nationalist government on Taiwan has continued to claim jurisdiction over the China mainland, whereas the government of the People's Republic of China on the mainland has claimed jurisdiction over Taiwan; both governments are in agreement that the island is a part of China. (See the city article TAIPEI; for associated physical features, see CHINA SEA; PACIFIC ISLANDS; and PACIFIC OCEAN; for historical aspects, see CHINA, HISTORY OF.)

THE LANDSCAPE

The natural environment. *Relief features.* The island is formed by a great fault block trending north-northeast to south-southwest, which is tilted toward the west. Its eastern margin, facing the Pacific Ocean and often lined with cliffs, marks the edge of the continental shelf. Its western section is on the slope of the shelf as is the Formosa Strait, the shallower parts of which are only 320 feet below sea level. The coastline on the west is simple and straight, bordered with low sand dunes and lagoons. Deep-water ports are situated at Chi-lung (Kee-lung), at the northern tip of the island, and at Kao-hsiung, on the southwest coast.

The crest of the Chung-yang Shan-mo (Central Range) lies east of and parallels the island's axis. Scores of peaks rise above 10,000 feet, the highest being Hsin-kao Shan (Yushan) (13,113 feet [3,997 metres]). Around the mountainous area are numerous independent hills, with an average height of 5,000 feet.

The relief on the east is very steep and that on the west quite gradual. The terrace tablelands and alluvial coastal plains in western Taiwan form the principal area of dense population; most of the major cities are located there.

The rivers

Drainage and soils. The rivers, all of which rise in the Chung-yang Shan-mo, are short and subject to extreme variations in flow. They are mostly shallow or dried up in the dry season, while there are floods in the monsoon (rain-bearing wind) season. Lacking steady currents, most rivers are not navigable but are liable to cause

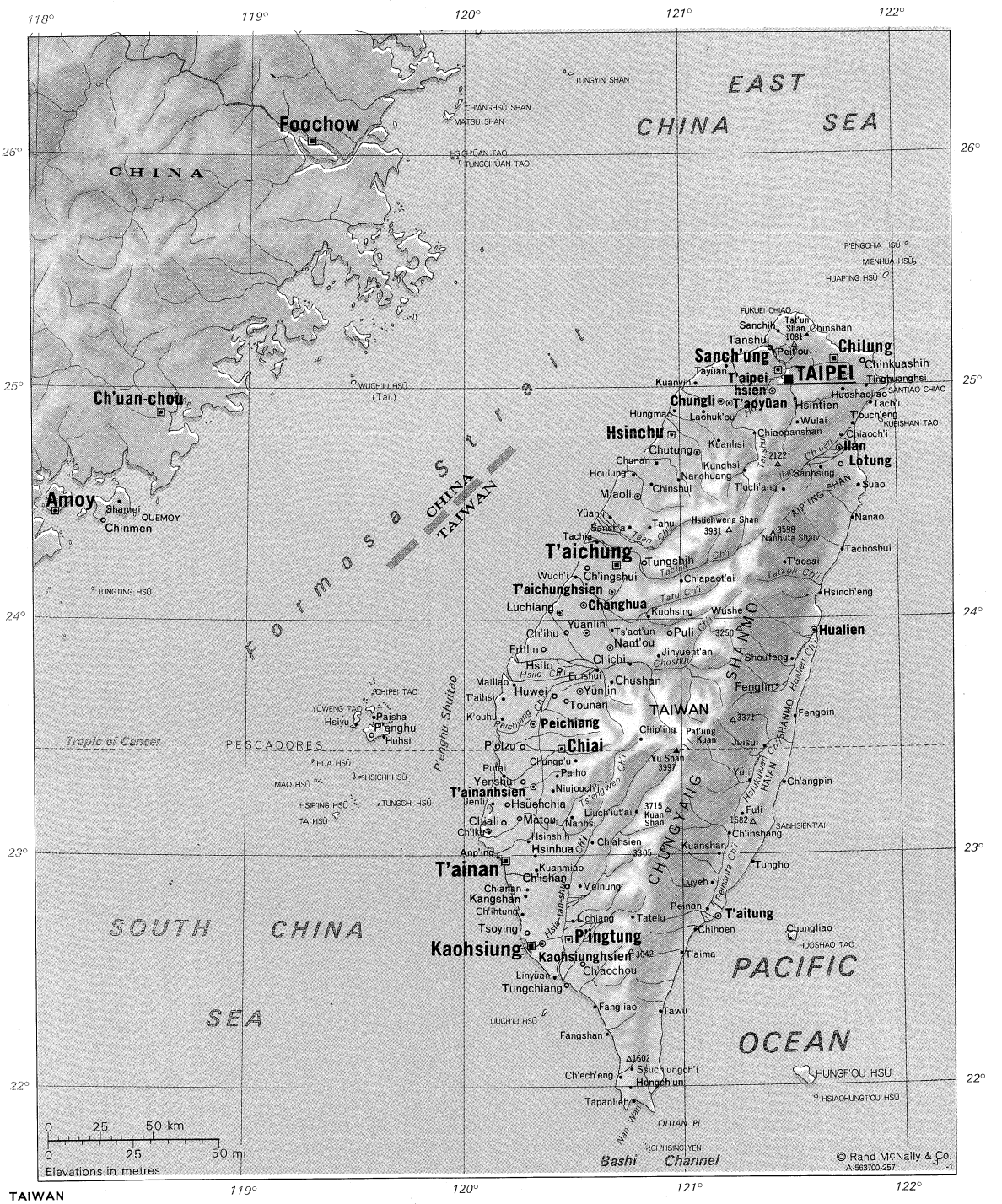
floods, especially in the plains, necessitating extensive drainage systems. The principal rivers are the Tan-shui Ch'i, which flows northward, and the Cho-shui Ch'i, which flows westward.

Soil of alluvial origin on the plains and in the valleys covers about one-fourth of the island and is its chief resource. The upland soils, subject to drastic erosion, are leached, acid, and infertile.

Climate. Taiwan straddles the tropical and subtropical zones and has warm summers and mild winters. The climate is moderated by the warm waters of the Kuroshio (Japan) Current. The summer is long, lasting from April until November (200 days or more). In cold months the mean monthly temperature is about 59° F (about 15° C). Beginning with April, the mean monthly temperature is above 68° F (20° C). The highest mean monthly temperature reaches 86° F (30° C) from June to September. Lowland Taiwan is frost free. The temperatures fall with altitude, and in winter the central mountains are covered with snow.

Mean annual precipitation is 102 inches, although in some years the precipitation in summer alone may exceed 200 inches. The upland area receives more rain than the lowlands, and the eastern part receives more than the western part. Rainfall is more even in the north. The precipitation in summer is about 80 percent of the annual total. There are more typhoons in July, August, and September than at other times; about 78 percent of the total occurs in these three months.

Vegetation and animal life. There are green plants on the plains all the year round. As the climate varies with altitude, so does the natural vegetation. Stands of mixed bamboo, palm, and tropical evergreen grow in the lowlands; subtropical evergreen forests including camphor



TAIWAN

Taiwan, Area and Population

	area		population	
	sq mi	sq km	1966 census	1972 estimate
Counties (*hsien*)				
Chang-hua	412	1,066	940,000	1,068,000
Chia-i	753	1,951	791,000	852,000
Hsin-chu	592	1,532	534,000	601,000
Hua-lien	1,787	4,629	315,000	340,000
I-lan	825	2,137	369,000	418,000
Kao-hsiung	1,094	2,833	755,000	859,000
Miao-li	703	1,820	467,000	532,000
Nan-t'ou	1,585	4,106	463,000	516,000
P'eng-hu	49	127	124,000	119,000
P'ing-tung	1,072	2,776	764,000	839,000
T'ai-chung	792	2,051	715,000	807,000
T'ai-nan	778	2,016	875,000	939,000
T'ai-pei	792	2,052	1,187,000	1,326,000
T'ai-tung	1,357	3,515	269,000	291,000
T'ao-yüan	471	1,221	634,000	761,000
Yün-lin	498	1,291	733,000	802,000
Municipalities (*shih*)				
Chi-lung	51	132	292,000	331,000
Kao-hsiung	44	114	675,000	889,000
T'ai-chung	63	163	394,000	477,000
T'ai-nan	68	176	431,000	490,000
T'ai-pei*	105	272	1,434,000	1,875,000
Total Taiwan	13,892†	35,981†	13,161,000†	15,130,000†

*Area and population figures include Yang-ming-shan special administrative district, which has been reported statistically with T'ai-pei municipality since 1967. †Figures do not add to total given because of rounding.
Source: Official government figures.

laurel are found from 2,000 to 6,000 feet; broadleaved evergreen forests of the temperate zone are represented by cedars, cypress, junipers, rhododendrons, maples, and cryptomeria (Japanese cedar) from 6,000 to 8,000 feet; and coniferous forests are found above 7,500 feet.

Animal life, similar to that of the southern Chinese mainland, includes deer, wild boars, bears, monkeys, goats, wildcats, and panthers. Birds include pheasant, geese, flycatchers, kingfishers, larks, and many other species. Fish abound in the coastal areas.

The landscape under human settlement. *Traditional regions.* Except for the aborigines, the inhabitants of Taiwan are the descendants of immigrants from the Chinese mainland. Settlement proceeded from the area around T'ai-nan, in the south, northward and from the coast inland. From the latter part of the Ming dynasty (early 17th century) until it was occupied by the Japanese in 1895, T'ai-nan was both the largest city and the capital of Taiwan. From T'ai-nan, development spread northward and southward. In the south, development spread from Kao-hsiung to P'ing-tung, 14 miles to the northeast, then to T'ai-tung, on the southeast coast, and Hua-lien, on the east central coast. In the north, it went up from Chang-hua, on the west central plain, to T'ao-yüan, on the northwest coast, Taipei, and I-lan, on the northeast coast. The principal regions are the Chianan plain, the earliest developed region; the Kao-ping plain, on the west coast; the Hengch'un peninsula, in the south; the Hua-tung valley, in the east; the Chang-hua plain; the Hsin-chu-T'ao-yüan terrace; the Taipei basin, the fertile northern portion of

Origin of the population

MAP INDEX

Cities and towns

An-p'ing 23·00n 120·09e
Chang-hua 24·05n 120·32e
Ch'ang-pin 23·20n 121·28e
Ch'ao-chou 22·33n 120·32e
Ch'e-ch'eng 22·05n 120·42e
Chia-hsien 23·05n 120·35e
Chia-i 23·29n 120·27e
Chia-li 23·10n 120·11e
Chia-nan 22·52n 120·16e
Chiao-ch'i 24·49n 121·46e
Chiao-pan-shan 24·49n 121·21e
Chia-pao-t'ai 24·11n 121·01e
Chih-pen 22·42n 121·03e
Ch'ih-shang 23·07n 121·12e
Ch'ih-tung 22·46n 120·16e
Ch'i-hu 23·58n 120·28e
Ch'i-ku 23·08n 120·07e
Chi-lung 25·08n 121·44e
Ch'ing-shui 24·15n 120·35e
Chin-kua-shih 25·08n 121·51e
Chin-men 24·27n 118·21e
Chin-shan 25·14n 121·36e
Chin-shui 24·36n 120·52e
Chi-p'ing 23·31n 120·49e
Ch'i-shan 22·53n 120·29e
Chu-nan 24·41n 120·52e
Chung-li 24·57n 121·13e
Chung-liao 22·40n 121·30e
Chung-p'u 23·25n 120·31e
Chu-shan 23·46n 120·41e
Chu-tung 24·44n 121·05e
Erh-lin 23·54n 120·22e
Fang-liao 22·22n 120·35e
Fang-shan 22·16n 120·39e
Feng-lin 23·44n 121·27e
Feng-pin 23·36n 121·31e
Fu-li 23·12n 121·16e
Heng-ch'un 22·00n 120·44e
Hsi-lo 23·49n 120·27e
Hsin-ch'eng 24·08n 121·39e
Hsin-chu 24·48n 120·58e
Hsin-hua 23·01n 120·20e
Hsin-shih 23·05n 120·17e
Hsin-tien 24·58n 121·33e
Hsi-yü 23·36n 119·30e
Hsüeh-chia 23·14n 120·11e
Hua-lien 23·58n 121·36e
Hu-hsi 23·32n 119·38e
Hung-mao 24·55n 120·58e
Huo-shao-liao 25·00n 121·45e
Hu-wei 23·42n 120·24e
I-lan 24·45n 121·44e
Jen-li 23·15n 120·08e
Jih-yüeh-t'an 23·52n 120·53e
Jui-sui 23·29n 121·23e
Kangshan 22·51n 120·17e
Kao-hsiung 22·38n 120·17e
Kao-hsiungh-sien 22·38n 120·21e
K'ou-hu 23·35n 120·10e
Kuan-hsi 24·48n 121·10e
Kuan-miao 22·58n 120·19e
Kuan-shan 23·03n 121·09e
Kuan-yin 25·02n 121·04e
Kung-hsi 24·39n 121·18e
Kuo-hsing 24·02n 120·51e
Lao-hu-k'ou 24·53n 121·03e
Li-chiang 22·44n 120·30e
Lin-yüan 22·30n 120·25e
Liu-ch'iu-t'ai 23·12n 120·48e
Lo-tung 24·41n 121·46e
Lu-chiang 24·04n 120·26e
Lu-yeh 22·55n 121·08e
Mai-liao 23·45n 120·15e
Ma-tou 23·11n 120·16e
Mei-nung 22·54n 120·32e
Miao-li 24·34n 120·48e
Nan-ao 24·28n 121·47e
Nan-hsi 23·11n 120·30e
Nan-t'ou 23·54n 120·41e
Pai-ho 23·21n 120·25e
Pai-sha 23·37n 119·35e
Pei-chiang 23·34n 120·18e
Pei-nan 22·48n 121·06e
Pei-t'ou 25·08n 121·29e
Peng-hu 23·34n 119·34e
P'ing-tung 22·40n 120·29e
P'o-tzu 23·29n 120·16e
Pu-li 23·58n 120·57e
Pu-tai 23·23n 120·09e
San-ch'a 24·25n 120·46e
San-chih 25·14n 121·37e
San-chung 25·04n 121·29e
San-hsing 24·40n 121·39e
Sha-mei 24·32n 118·25e
Shou-feng 23·52n 121·30e
Ssu-ch'ung-ch'i 22·06n 120·46e
Su-ao 24·36n 121·51e
Ta-ch'i 24·57n 121·53e
Ta-chia 24·21n 120·37e
Ta-cho-shui 24·19n 121·45e
Ta-hu 24·26n 120·52e
T'ai-chung 24·09n 120·41e
T'ai-chung-hsien 24·15n 120·43e
T'ai-hsi 23·42n 120·11e
T'ai-ma 22·37n 120·59e
T'ai-nan 23·00n 120·11e
T'ai-nan-hsien 23·18n 120·19e
Taipai 25·03n 121·30e
T'ai-pei-hsien 25·00n 121·27e
T'ai-tung 22·45n 121·09e
Tan-shui 25·10n 121·26e
T'ao-sai 24·16n 121·30e
T'ao-yüan 24·59n 121·19e
Ta-pan-lieh 21·59n 120·47e
Ta-te-lu 22·45n 120·47e
Ta-wu 22·22n 120·54e
Ta-yüan 25·06n 121·14e
Ting-huang-hsi 25·01n 121·52e
T'ou-ch'eng 24·51n 121·49e
Tou-nan 23·41n 120·28e
Ts'ao-t'un 23·59n 120·40e
Tso-ying 22·41n 120·17e
T'u-ch'ang 24·34n 121·29e
Tung-chiang 22·28n 120·26e
Tung-ho 22·58n 121·18e
Tung-shih 24·16n 120·50e
Wu-ch'i 24·13n 120·31e
Wu-lai 24·52n 121·33e
Wu-she 24·02n 121·08e
Yen-shui 23·20n 120·16e
Yüan-lin 23·57n 120·34e
Yü-li 23·20n 121·19e
Yün-lin 23·43n 120·33e

Physical features and points of interest

Bashi Channel 21·30n 120·45e
Ch'ang-hsü Shan, *island* 26·14n 119·58e
Ch'hsing Yen, *island* 21·44n 120·50e
Chi-pei Tao, *island* 23·42n 119·35e
Cho-shui Ch'i, *river* 24·03n 120·23e
Chung-yang Shan-mo, *mountain range* 23·30n 121·00e
East China Sea 26·15n 121·30e
Formosa Strait 24·00n 119·00e
Fu-kuei Chiao, *cape* 25·19n 121·34e
Hsiao-hung-t'ou Hsü, *island* 21·57n 121·35e
Hsia-tan-shui Ch'i, *river* 22·28n 120·25e
Hsi-chi Hsü *island* 23·15n 119·37e
Hsi-ch'üan Tao, *island* 25·59n 119·57e
Hsi-lo Ch'i, *river* 23·47n 120·15e
Hsi-p'ing Hsü, *island* 23·14n 119·36e
Hsiu-ku-luan Ch'i, *river* 23·28n 121·29e
Hsüeh-weng Shan, *mountain* 24·24n 121·12e
Hua Hsü, *island* 23·24n 119·20e
Hua-lien Ch'i, *river* 23·57n 121·36e
Hua-p'ing Hsü, *island* 25·24n 121·58e
Hung-f'ou Hsü, *island* 22·03n 121·33e
Huo-shao Tao, *island* 22·40n 121·30e
I-lan Ch'uan, *river* 24·42n 121·48e
Kuan Shan, *mountain* 23·14n 120·54e
Kuei-shan Tao, *island* 24·50n 121·56e
Liu-ch'iu Hsü, *island* 22·20n 120·22e
Mao Hsü, *island* 23·19n 119·19e
Ma-tsu Shan, *island* 26·09n 119·56e
Mien-hua Hsü, *island* 25·29n 122·06e
Nan-hu-ta Shan, *mountain* 24·22n 121·26e
Nan Wan, *bay* 21·53n 120·48e
O-luan Pi, *cape* 21·54n 120·51e
Pacific Ocean 22·30n 121·30e
Pa-t'ung Kuan, *pass* 23·30n 121·01e
Pei-chiang Ch'i, *river* 23·32n 120·09e
Pei-nan-ta Ch'i, *river* 22·46n 121·10e
P'eng-chia Hsü, *island* 25·38n 122·04e
P'eng-hu Shuitao, *strait* 23·30n 119·50e
Pescadores, *islands* 23·30n 119·30e
Quemoy, *island* 24·27n 118·23e
South China Sea 22·30n 119·00e
Ta-an Ch'i, *river* 24·22n 120·42e
Ta-chia Ch'i, *river* 24·19n 120·34e
Ta Hsü, *island* 23·10n 119·32e
T'ai'p'ing Shan, *mountain* 24·30n 121·38e
Tan-shui Ho, see Hsia-Tan-shui Ch'i
Ta-tu Ch'i, *river* 24·12n 120·29e
Ta-t'un Shan, *mountain* 25·10n 121·31e
Ta-tzu-li Ch'i, *river* 24·08n 121·40e
Ts'eng-wen Ch'i, *river* 23·04n 120·04e
Tung-chi Hsü, *island* 23·15n 119·40e
Tung-ch'üan Tao, *island* 25·58n 119·58e
Tung-ting Hsü, *island* 24·10n 118·14e
Tung-yin Shan, *island* 26·22n 120·30e
Wei-tou-wan, *bay* 24·34n 118·30e
Wu-ch'iu Hsü, *island* 25·00n 119·27e
Yu Shan, *mountain* 23·28n 120·57e
Yü-weng Tao, *island* 23·35n 119·30e

the west-coastal alluvial plain; and the I-lan plain, an isolated alluvial lowland on the northeastern coast.

Rural settlement. The patterns of rural settlement are determined by two major factors—water sources and the availability of arable land. In the northern part where water sources are abundant and the terrain is hilly, there are scattered villages. In the southern part, where the plains are flat and water is scarce, the villages are compact. Where rice paddies are located on the plain, the villages are compact and densely populated as a result of the concentration of cultivation. Upland villages are usually less compact and more widely dispersed. Villages on coastal land also are less compact and more widely scattered than those where rice is grown.

Urban settlements. Because of the straight, unbroken coastline, the only ports are Kao-hsiung, in the south; Chi-lung (the port of Taipei), in the north; and Hua-lien, in the east. In former days there were ports along the west coast at Anp'ing, Lu-chiang, and Pei-chiang, but, because of the extension of the seashore, their harbours have been filled in, and they are no longer on the sea. Because of the constant changes in water level, most rivers are not navigable, and there are few towns along the riverbanks.

Railways and highways are the principal means of inland transportation, and many towns are established along the transportation lines. In the north, the municipality of Taipei is the centre of a well-developed urban region, including Chi-lung and T'ao-yüan. In the south Kao-hsiung is the centre of an urban region that includes T'ai-nan and P'ing-tung. Other urban areas are centred on Hsin-chu and T'ai-chung.

PEOPLE AND POPULATION

The population, as already mentioned, is predominantly Chinese, most of whom are the descendants of immigrants from the southeastern Chinese provinces of Fukien and Kwantung; in addition, after the transfer of the Nationalist government to the island in 1949, there was a sizable immigration from other parts of China. The aborigines, of Indonesian origin, number about 150,000, or about 1 percent of the total population, and live in the foothills and highlands, where many practice shifting (temporary) cultivation. All have undergone partial assimilation with the Chinese.

Principal dialects

Linguistic and other groups. The post-1949 immigrants were predominantly Mandarin speaking and, because of their prominence in the Nationalist government and of the prevalence of education under the Nationalists, Mandarin has become a principal language. Several other Chinese dialects are still spoken, however, including the southern Fukien dialect and the Hakka dialect of Fukien Province. There are also several aboriginal dialects. The so-called Taiwanese dialect comprises the spoken languages of the southern Fukien areas, such as Lung-ch'i Ch'üan-chou, and Hsia-men. The aboriginal dialects include Atayal, Seepeg, Saisiat, Bunun, Tsou, Kanabu, Saroa, I-Rutkai, II-Rutkai, Paiwan, Puyuma, Ami, and Yami.

Ethnic groups. The Chinese population of Taiwan is almost entirely composed of Hans. The aborigines are divided into nine families: the Atayal, Saisiat, Paiwan, Rutkai, Bunun, Tsou, Yami, Ami, and Puyuma. Among them the Atayals predominate, with about 40 percent of the total aboriginal population; next are the Paiwans, with about 30 percent; and the Bununs, with about 20 percent.

Religious groups. In addition to the forms of worship of the aborigines, many kinds of religion have been introduced into Taiwan from many parts of the world over the past 300 years. The many Chinese immigrants brought their religions, principally Buddhism and Taoism, with them. In 1622 the Dutch introduced Protestant Christianity; in 1624 the Spanish brought Roman Catholicism to the island; and later Japan introduced Shintoism there. In addition, Confucianism has immensely influenced the Chinese people of Taiwan in ethics, morality, and academic thinking.

The principal religions in Taiwan today, in addition to the form of worship of the aborigines, are Taoism and Buddhism. There are about 2,000 Taoist temples. The Buddhist Association of the Republic of China is located in Taipei, and Buddhist associations are to be found throughout the island; there are more than 3,000 Buddhist temples and about 5,750,000 devotees. Protestant Christians number about 320,000; they have about 1,000 churches. The Roman Catholics, for their part, number about 310,000 and have more than 660 churches. There are about 40,000 Muslims, who have about 12 mosques.

Demography. *Birth and mortality rates.* The total population of Taiwan in 1972 was 15,130,000. The population of Taipei was 1,865,000—about 12 percent of the total—but it included many rural people living within the official municipality. Taiwan has had a rapid rate of population growth. Its population density in 1972—1,089 persons per square mile—was one of the highest in the world. Since the government adopted a new population policy, the birth rate has dropped from 38 per 1,000 in 1947 to 26 per 1,000 in 1971. Due to improvements in medical treatment and public sanitation, the death rate has dropped markedly, to 4.8 per 1,000 (1971).

Growth and density of population

Immigration. The Chinese mainlanders who migrated to the island after 1949 numbered over 1,800,000 by 1970. There has been a noticeable population movement from rural areas to the towns. The most rapidly growing urban areas are Taipei, Kao-hsiung, T'ai-nan, Tai-chung, Taipei County, and T'ao-yüan County.

Population distribution. The distribution of Taiwan's population is influenced by the island's terrain. The coastal plains and basins in the west are agricultural areas in which the population is dense because of the development of transportation and irrigation. The population density in the western plains reaches about 2,560 persons per square mile. In the craggy terrain of the central and eastern mountains, the density of population is fairly low.

Demographic trends. The population of Taiwan is comparatively young, with 38.7 percent under 15 years of age. The annual rate of population increase dropped from 28 per 1,000 in 1964 to 21 per 1,000 in 1971 after the government advocated birth control and family planning in 1964. Deaths from disease and calamity had also declined by the early 1970s. Rate of increase is expected to decline to 18 per 1,000 in the 1980s, when the population is thought to increase at the rate of 326,000 a year.

THE NATIONAL ECONOMY

Agriculture, long the leading element in the national economy, is developed mainly on the western plains, while industry is concentrated largely in the areas surrounding the harbours of Chi-lung and Kao-hsiung. Marked industrial development, especially in the 1960s, resulted by the 1970s in a notable change—*i.e.*, from a predominantly agricultural to a predominantly industrial economy. Export trade, which, because of Taiwan's rapid economic growth, has gained an important position, had gradually centred about industrial products; among these the proportion of textile goods had increased sharply, while the proportion represented by machinery and electrical appliances had also increased.

Resources. Although more than 50 kinds of minerals have been found in Taiwan, total mineral resources are modest. Coal is the most important, reserves occurring near Taipei and in several other areas of the north amounting to about 254,000,000 tons. Deposits of copper and gold are found in the Taipei area. Iron and sand reserves are small. Sulfur and sulfur-pyrite reserves amount to about 2,500,000 tons, with deposits in Pei-t'ou, Chu-tzu-hu, Ten-shan, and Ho-shan-chun all located north of Taipei. Marble and limestone are abundant in Hua-lien on the east coast, which has reserves of 330,000,000 tons. Dolomite (limestone or marble rich in magnesium carbonate) is mostly deposited near Hua-lien. Petroleum, phosphorus, and natural gas exist in small quantities and occurrences of manganese, asbestos, talc, glass sand, and other minerals have been found. Salt is produced by the evaporation of sea water.

Mineral resources

Forests are most abundant in the high mountains. One of the natural resources of Taiwan is camphor laurel, from which a great deal of camphor is produced.

The total output capacity of the power stations in the early 1970s was 15,171,000,000 kilowatt-hours, 20 percent of which was produced by hydroelectric stations and 80 percent by thermal stations. A nuclear power plant was under construction early in the 1970s.

Sources of income. *Agriculture.* One-quarter of the total area in Taiwan is arable and fully cultivated. Sloping areas and dry riverbeds are also under cultivation, as also are the tidal lands in the west.

Rice cultivation

The single most important crop is rice, with a total planted area of about 1,000,000 acres, about 50 percent of the total cultivated area and including most of the irrigated portion. More than 60 percent of the paddy fields are double cropped and nearly two-thirds of the rice produced is higher-grade *pon-lai* rice, introduced by the Japanese. With improved farming techniques, including increased and more effective use of fertilizer, the production of rice per acre had increased more than 50 percent between the early 1950s and the early 1970s.

Sugarcane is another important agricultural product, with a total planted area of 235,000 acres. Tea plantations are mostly in the north. Hemp and jute are grown in the middle and eastern part, while a different variety of hemp is grown in the south. Increased production and consumption of wheat reflects the presence of more northern Chinese than in earlier days. Pineapple is one of the principal export products. The total volume of pineapple production was about 359,000 tons in the early 1970s. Fruits and vegetables, such as bananas, litchis, longans, peaches, watermelons, and oranges abound; all these fruits are canned for export. Most vegetables are produced in the central and southern regions and are sufficient to meet the consumption needs of the island.

Forests and fisheries. With many mountains, Taiwan has abundant timber. The Taiwan Forest Bureau, in charge of the national forest land, is responsible for the planting, protection, and logging of forests. The annual production of forest products, including plywood, reaches an average of about 1,200,000 cubic yards, exclusive of firewood. Inaccessibility, low quality, and high costs limit production and have made import of lumber necessary. Overcutting and lack of reforestation have resulted in erosion and destructive floods.

Fishery production has increased to over 650,000 tons annually. About 12 percent of this is produced by fish farming. The balance is caught in rivers and in coastal waters. The warm currents in the Pacific Ocean off the east coast provide good deep-sea fishing grounds, especially for tuna. The Taiwan Fishery Bureau was established to help develop the fishing industry, and the number of fishing vessels is increasing in the early 1970s.

Mining and quarrying. Annual coal production averages between 4,000,000 and 4,500,000 tons. Copper production is about 3,900 tons. Quarrying of marble exceeds 1,100,000 tons and of dolomite 55,000 tons. The Chinese Petroleum Corporation not only prospects for and produces crude petroleum but also refines crude petroleum imported from abroad. Annual production of crude petroleum is about 88,000 tons and that of natural gas about 1,200,000 cubic yards. Salt production is located mainly on the southwestern coast. Production varies considerably, mainly because of weather factors, but averages about 440,000 tons.

Government-owned enterprises

Manufacturing. Industrialization was begun by the Japanese, and its growth has been accelerated under the Nationalist regime. Government-owned manufacturing enterprises include aluminum, cement, coal, fertilizer, machinery, paper, and sugar. Textile manufacture, the leading industry, is largely privately owned. Food processing, chemicals, and chemical products, electrical appliances and equipment, and tobacco processing are also important, as are the production of glass, cement, leather goods, and rubber goods. Printing and publishing, especially the reprinting, without authorization, of trade books and textbooks and sets of reference works copyrighted in other countries, has flourished, although in the early 1970s the signing by the government of international copyright agreements was being mooted, and some publishers were adhering to them.

Foreign trade. Trade balances after World War II were generally unfavourable until the mid-1960s. As before the war, exports were principally agricultural products—for example, sugar and rice—accounting for 78 percent of total value in 1952. With the growth of industry, the relative importance of agricultural exports had declined. By the 1970s, although agricultural exports—especially fruits and vegetables, including bananas—were still important, manufactured goods accounted for more than 50 percent of foreign trade, led by textile yarns and fabrics, clothing, telecommunications equipment, and plywood. Imports were highly diversified, consisting of a variety of consumer goods and raw materials, including bauxite, lumber, rubber, chemicals, cotton, wool, and synthetic fibres and resins. The government has encouraged multilateral trade, and trade patterns are diverse, especially when compared with the prewar period when they were virtually limited to Japan. In the 1970s trade has been conducted with Japan, the United States, Hong Kong, South Vietnam, West Germany, and other countries.

Management of the economy. *The private sector.* The government has established an advisory committee that is responsible for overseeing production; its main purpose is to promote foreign sales. There is only one organ of cooperative financing—the Cooperative Bank of Taiwan, jointly owned by the government and cooperatives. In addition there are four private commercial banks. Most of the public enterprises are being gradually opened to private operation, except for certain government monopoly businesses.

The public sector. Since 1953 the government has promulgated successive four-year economic development plans to spur industrial expansion and economic growth. Ten enterprises, in the fields of sugar, electricity, petroleum and natural gas, fertilizer, caustic soda, aluminum, copper, shipbuilding and repairs, and railways, are government monopolies under the Ministry of Economic Affairs.

In order to promote economic development, the government controls and helps private enterprises to employ new techniques and locate foreign investment. Most manufacturing industries receive technical cooperation from foreign companies.

Taxation. Taxes in Taiwan include income, legacy, commodity, stamp, stock, farm, land, increment on land value, and business taxes. Taxes are levied according to a progressive rate; people with small incomes pay little tax.

Trade unions. The Chinese Federation of Labor (CFL) is a nationwide organization of industrial and craft unions. Five other national labour organizations include those for seamen, railway workers, postal workers, miners, and salt workers. There are local unions in all factories, transportation and public utility units, and occupational and vocational groups.

Problems and prospects. Because Taiwan is an island and thus limited in resources and in domestic market, it must exchange goods with other countries and exploit overseas markets in order to support mass production and cut down its production costs. Although labour costs are low, shortages of labour in the future could result from success in achieving rapid economic growth. If Taiwanese society remains stable and production quality is improved, prospects for economic growth are promising.

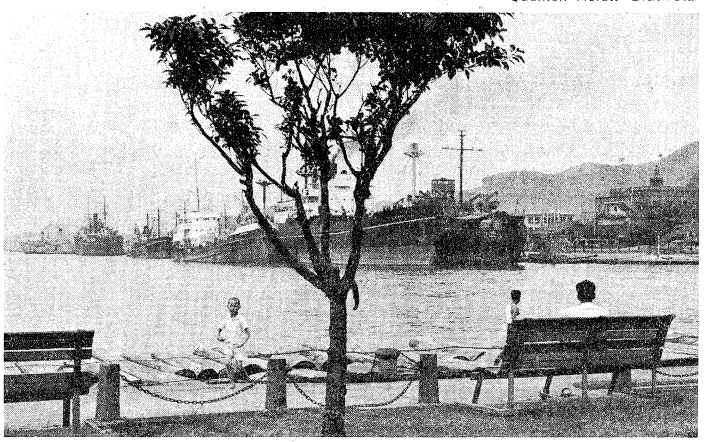
Transportation. External transport links are by air and sea. Railways and highways for internal communications are well developed.

Road networks

The well-built road networks have a total length of about 10,000 miles. The principal roads consist of a highway running around the island and three east–west highways crossing the island in the northern, middle, and southern regions of Taiwan. A north–south expressway was under construction in the 1970s. Passenger-bus transportation connecting large cities and small towns throughout the island is provided by the Taiwan Highway Bureau and private transportation companies provide both passenger and freight service. In large cities there are public buses as well as private buses and taxis. Few people own cars, but many have motorcycles.

Railways, operated by the Taiwan Railway Administration, have a total track length of about 1,000 miles, including 140 miles of double track. The main lines include those from Chi-lung, in the north, to Kao-hsiung, in the southwest; from Chi-lung to Su-ao, in the northeast; from Kao-hsiung to Fang-liao, in the south; and from Hua-lien, in the east, to T'ai-tung, in the southeast. Private railroads, including those belonging to sugar, forestry, and salt enterprises, total about 2,000 miles; they also carry a small number of passengers.

Quentell Violett—Black Star



Chi-lung harbour, the port of Taipei, with facilities to accommodate ships up to 20,000 gross tons.

The international seaports are Chi-lung, in the north; Kao-hsiung, in the south; and Hua-lien, in the east. Chi-lung is the port for Taipei city, while Kao-hsiung is an industrial port. Both cities have good facilities for anchoring large ships. Hua-lien, with facilities that are considerably more limited, has been open as an international port since 1963.

Civilian airports include Taipei International Airport, in the north; Kao-hsiung Airport, in the south; and Hua-lien Airport, in the east. Civilian air transports to T'ai-nan, T'ai-chung, T'ai-tung, and P'eng-hu are permitted to use the military air bases in those areas. From Taipei International Airport service is available to the United States and many Asian cities. Two Airlines provide several round-trip flights daily between Taipei and Hua-lien Kao-hsiung, P'eng-hu, Tai-nan, and Tai-chung.

ADMINISTRATION AND SOCIAL CONDITIONS

Structure of government. *Constitutional framework.* The central government of the Republic of China combines the Cabinet and presidential systems of government. The president is the supreme leader of the nation. Under him there are five independent *yüans* (councils)—executive, legislative, judicial, examination (fulfilling civil service commission functions), and control (an investigative agency).

Regional, state, and local government. According to the constitution of the Republic of China, all levels of local government below the provincial level have self-government—Taiwan having the status of a province of the republic. The provincial government is headed by a governor and has a secretariat, with a secretary general, consisting of departments, bureaus, and offices in charge of various governmental functions. The provincial government council consists of 21 members selected by the national executive *yüan* and appointed by order of the president.

The provincial government

Under the control of the provincial government, there are 16 counties (*hsien*) and five municipalities (*shih*). Eight municipalities are under the control of county governments. County governments are headed by magistrates and cities by mayors, both elected for four-year terms, as are the county and city councils.

Taipei (*q.v.*) became a special municipality under control of the executive *yüan* in 1967; it has a position equal to that of the Taiwan provincial government.

The political process. The president is elected by the National Assembly, which is composed of delegates elected from each county (or the equivalent) throughout the country. The president of the executive *yüan* is nominated by the national president with the consent of the legislative *yüan*, which is composed of the members elected by the people. The presidents of the judicial *yüan* and the examination *yüan* are also nominated by the national president with the consent of the control *yüan*. The president of the control *yüan* is elected by the provincial and municipal councils. According to the constitution, suffrage in all local elections is universal, equal, direct, and secret.

There are three parties: the Kuomingtang (Nationalist), the Young China Party, and the Democratic Socialist Party. The Kuomingtang, the oldest among the three parties, has been the ruling party since the establishment of the Republic of China; it was founded in 1894. According to the constitution of the Republic of China, citizens have the powers of election, recall, initiative, and referendum. The rights of initiative and recall have, however, been suspended, and the National Assembly exercises these rights on behalf of the people.

Justice. The judicial *yüan* is the highest organization of the nation. Under its control are the three principal judicial organizations—the Supreme Court, the Administrative Court, and the Committee on Discipline of Public Functionaries. In addition, there are district courts in the districts and autonomous regions and high courts in provinces, or special areas.

The armed forces. The armed forces include the air force, army, navy, combined service force, military police, and garrison force. The army has a strength of 390,000 men, equipped and trained by the United States; the navy has a strength of 35,000 naval personnel and 35,000 marines; and the air force has a personnel of 80,000. All of the military units are said to be well trained and fully equipped. The armed forces are under the administration of the Ministry of National Defense. The president is the supreme commander of the national armed forces.

The social milieu. *Education.* The educational system is divided into three stages: national education, middle education, and higher education. There is also preschool education (kindergarten) and social education, including adult education and special education.

National education is compulsory for nine years (six years in primary school and three years in junior middle school); middle education includes middle schools, normal schools, and vocational schools; higher education includes graduate schools, universities, colleges, and junior colleges. There are over 90 institutions of higher education, more than half of them private. The two largest are the National Taiwan University (founded 1928) at Taipei and National Chung Hsing University (founded 1961) at T'ai-chung.

Health and welfare. In the early 1970s there were more than 100 private hospitals, as well as a considerable number of public hospitals, including four general hospitals and five specialist hospitals in Taipei. There are also about 350 county sanitariums. Government sources state that all the hospitals and sanitariums are well equipped.

Housing. To meet the needs of population growth and urban development, the government has built apartments that are sold to the public by means of long-term, low-interest loans. The government provides houses for the poor without charge.

Police. Under the control of the provincial police department, every municipality or county has its own police unit for traffic regulation, maintaining order, dealing with emergencies, such as fires, handling criminal cases, preventing smuggling and larceny, and protecting alien residents.

Living conditions. Taiwan has a stable, prosperous society. Individual income has increased every year. Wages in Taiwan are comparatively low, however, and food expense accounts for a large percentage of the cost of living. National savings have increased rapidly.

Cooperatives

The national constitution provides that "cooperative enterprises shall receive encouragement and assistance from the State," and there are agencies at all government levels to assist such enterprises. Regional cooperatives are organized to serve rural districts, villages, or townships. Specialized cooperatives include those for agricultural and industrial production, marketing, rental service (*e.g.*, agricultural machinery), labour, transportation, public utilities, procurement and consumption, credit, and insurance. In addition, there are over 200 cooperative farms, and the number of cooperatives grew from fewer than 500 in 1947 to almost 3,000 early in the 1970s.

After years of effort, Taiwan has greatly improved public health conditions. As of 1970, the average life span was 69 years for females and 65 years for males. There have been no international quarantines of outbreaks of contagious diseases in Taiwan in recent years. Malaria was eliminated by 1965.

Cultural life and institutions. Chinese culture is predominant. Fine arts, surviving national folk traditions, and popular culture include Chinese painting, and Chinese music. Popular musical instruments include the fiddle, lute, flute, and zither. Dances include court dances, aboriginal dances, and Chinese folk dances. Theatre includes traditional Chinese opera, Taiwanese operas, which are mostly copied from the Chinese opera and use the Fukien dialect instead of Mandarin, and modern drama presented on the stage as well as over radio and television. Puppet shows are popular. Taiwan rivals Hong Kong in the making of Chinese motion pictures. Taiwanese handicrafts are prized for their beauty.

The cultural bureau of the Ministry of Education is in charge of the administration of cultural affairs. The National Palace Museum collection includes ancient Chinese paintings and books, pottery, porcelain, curios, and sculptures. The Art House in Taiwan holds art exhibitions and has a stage for drama and dance performances.

The media

There are about 30 daily newspapers, two of which are English language publications; 16 daily papers are published in Taipai and the rest in other major cities. There are about 1,000 periodicals, most of them house organs of various organizations. There are about 30 radio broadcasting stations; the three television stations include the Taiwan Television Enterprise, the China Television Company, and the Chunghua Television Service. Many of their programs are in colour. The Chunghua Television Service also features educational programs.

Prospects for the future. Taiwan is one of the most strategically important islands in the western Pacific; its future political role is certain to be of significance.

The government of Taiwan has proclaimed the recovery of the Peoples' Republic of China to be a matter of national policy. It thus devotes a large part of the national budget to military expenditures. Nevertheless, the growth rate of the economy in the early 1970s was keeping pace with the growth of the population. Grain crops were sufficient to supply the needs of the population, and, in addition, the government is continuing to take effective measures to improve fertilization and irrigation techniques and to develop new strains of seeds. Industry has continued to develop, both with domestic capital and foreign investment. In consequence, Taiwan has become a self-supporting country that is no longer dependent on foreign aid and that is an exporter of technical skills, particularly to developing countries in Asia and Africa.

BIBLIOGRAPHY. CHENG-SIANG CHEN, *A Geography of Taiwan*, 2 vol. (1960), fully illustrated, is a valuable reference. CHIAO-MIN HSIEH, *Taiwan-Ilha Formosa* (1964), deals with the land and people of the island. The *National Atlas of China*, vol. 1, *Taiwan*, ed. by CHI-YUN CHANG, (1963), has detailed statistical maps. The *China Yearbook* (annual), ed. by the INFORMATION BUREAU OF CHINA, treats Taiwan from various aspects, especially administration. The *Taiwan Statistical Data Book* (annual), published by the TAIWAN PROVINCIAL GOVERNMENT, is an authoritative official source. Of special interest are MEI-LING HSU, *Taiwan Population Distribution, 1965* (1969); and *Industry of Free China* (monthly), which reports economic developments.

(Y.-c.K.)

Takla Makan Desert

A desert of Central Asia and one of the largest sandy deserts in the world, the Takla Makan (Ta-ko-la-ma-gan in Chinese Pin-yin romanization) occupies the central part of the Tarim Basin in China. The desert area extends about 600 miles from west to east, and it has a maximum width of 260 miles and a total area of about 105,000 square miles (270,000 square kilometres). The desert reaches altitudes of 3,900 to 4,900 feet in the west and south and from 2,600 to 3,300 feet in the east and north.

For associated physical features, see KUNLUN MOUNTAINS; PAMIR MOUNTAIN AREA; TIEN SHAN (MOUNTAINS); and TARIM RIVER.

Physical characteristics. The Takla Makan is flanked by high mountain ranges: the Kunlun in the south, the Tien Shan in the north, and the Pamirs in the west. There is a gradual transition to the marshy Lop Nor (Lop Lake) Basin in the east, and, in the south and west, between the sandy desert and the mountains, lies a band of sloping desert lowland composed of pebble-detritus deposits.

Several small mountain ranges and chains, composed of sandstones and clays from the Paleogene and Neogene periods (10,000 to 65,000,000 years ago), rise in the western part of the desert. The arc-shaped Mazar Tagh (Mazar Mountains), located between the Ho-t'ien Ho (Ho-t'ien River) or Khotan Darya (Khotan River) and the Yarkand River valleys, arch toward the southwest. Ninety miles long and from two to three miles wide, with a maximum height of 5,363 feet, they rise an average of only 1,000 to 1,150 feet above the surface of the sandy plain. The Chiao-lo Shan (Chöl Tagh) is also an insular range, surrounded on all sides by massifs of moving sands. Rosstagh, also known as Tokhtakaz Mountain, its highest peak, is 5,117 feet (1,560 metres), and the range rises from 600 to 800 feet above the plain. Both ranges are covered by a shallow mantle of eluvium and rock debris and have sparse, desert-type vegetation. In the north, the sands of the Takla Makan form a clear boundary with the vegetated Tarim River Valley.

The general slope of the plain is from south to north, and thus the rivers running off from the Kunlun Mountains flow in this direction. The Ho-t'ien and K'o-li-ya (Keriya) river valleys have survived up to the present, but a majority of the shallower rivers have been lost in the sands, after which their empty valleys fill up with windborne sand.

Wind-formed surface features

The surface of the Takla Makan is composed of friable alluvial deposits several hundred feet thick. This alluvial stratum has been affected by the wind, and its wind-borne sand cover is as much as 1,000 feet thick. There is a variety of eolian (wind-formed) topographic features, and sand dunes of various shapes and sizes are encountered. These eolian sand dunes were formed through the weathering of the alluvial and colluvial deposits of the basin and of the foothill plains of the Kunlun and eastern Tien Shan. The size of the larger sand-dune chains is considerable: they range from 100 to 500 feet high and 800 to 1,650 feet wide, with a distance between the chains of from one-half to three miles. The highest eolian topographic forms are the pyramidal dunes, rising 650 to 1,000 feet. In the eastern and central parts of the desert, networks of hollow dunes and large, complex sand-dune chains predominate. They are also common in the western portion of the desert (east of the Ho-t'ien Ho Valley), where transverse and longitudinal (with respect to the wind) topographic forms coexist. On the edge of the desert, semipermanent, clustered sand dunes with tamarisk and niter-bushes, as well as clayey regions with disconnected sand dunes, predominate. Such a diversity in wind-formed features is a result of the complex wind conditions of the basin.

Climate. The Takla Makan's climate is moderately warm and markedly continental, with a maximum annual temperature range of 70° F (21° C). Precipitation is very low, ranging from 1.5 inches in the west to 0.4 inch in the east. The air temperature in the summer is high, rising as high as 100° F (38° C) on the eastern edge of the



The Takla Makan Desert.

desert. In July the average air temperature reaches 77° F (25° C) in the eastern regions. Winters are cold: in January the average air temperature is 16° to 14° F (−9 to −10° C), and the low in winter is generally below −4° F (−20° C). Northerly and northwesterly winds predominate in the summer in the western region. These two air currents, on meeting near the desert's centre at the end of the K'o-li-ya Ho, create a complex circulation system that is clearly reflected in the topography of the sand dunes. In the spring, when the surface sand becomes warm, ascending currents develop, and northeasterly winds become particularly strong. During this period hurricane-force dust storms, filling the atmosphere with dust to altitudes of from 11,500 to 13,000 feet, often occur. Winds from other directions also raise clouds of dust into the air, covering the Takla Makan with a shroud for almost the entire year.

Hydrology. Since the Tarim depression is an internal-drainage basin, the entire runoff from the surrounding mountains collects in the basin itself, feeding the rivers and the groundwater strata. In all probability, the groundwater table under the sands has an unbroken course, flowing west to east to Lop Nor. The importance of rainfall in moistening the sands and feeding the groundwaters is slight, due to its small quantity and high rate of evaporation. The rivers draining the Kunlun Mountains penetrate from 60 to 120 miles into the desert, gradually drying up in the sands. Only the Ho-t'ien Ho crosses the desert and, in July and August, occasionally carries its waters to the Tarim River.

The groundwater table

Vegetation and animal life. Vegetation is very sparse in the Takla Makan—almost the entire territory is devoid of plant cover. In depressions among the sand dunes, where the groundwaters lie no deeper than 10 to 16 feet from the surface, thin thickets of tamarisk, niter-bushes, and reeds may be found. The thick strata of moving sands, however, prevent the wider spread of this vegetation. Along the edges of the desert, in the area where the sand dunes meet the river valleys and deltas and where the groundwaters lie comparatively close to the surface, the vegetation is richer. Here, in addition to the plants mentioned above, a number of species characteristic of river valleys are found: Turanga poplar, oleaster, camel thorn, Zygophyllaceae (caltrop family), and saltworts. Sand dunes in hummocks frequently form around the scrub.

The animal life of the Takla Makan is also very sparse. Only in peripheral regions of the desert, in ancient and modern river valleys and deltas where water and vegetation appear, are the fauna more diverse. Herds of gazelles are found in open spaces, and there are usually wild boars inhabiting thickets of bulrushes and reeds in the river valleys. Wolves and foxes are among the carnivores. As late as the beginning of the 20th century, tigers could be found, but they have since been exterminated. Rare animals include the Siberian deer, which inhabits the Tarim River Valley, and the wild camel, which at the end of the 19th century roamed over a considerable portion of the Takla Makan as far as the Ho-t'ien Ho but now appears only occasionally, in the eastern desert region.

There are a large number of rabbits, sand eels, field mice, and jerboas in the sand dunes, and, among insectivores, the long-eared hedgehog and bats are common. Small, tufted larks and the Tarim jay are the commonest birds.

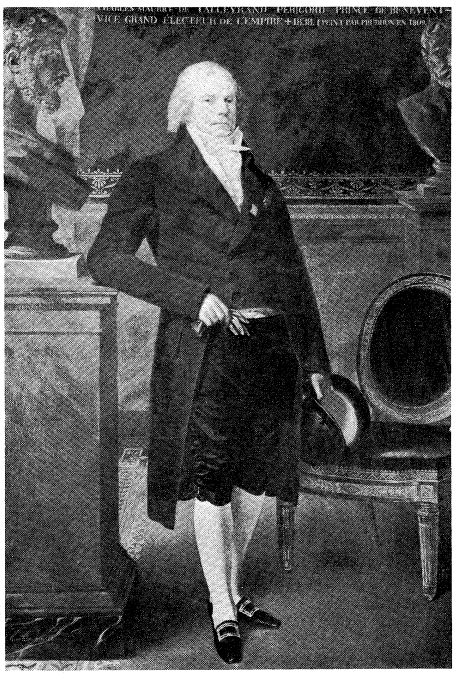
There is no fixed population in the Takla Makan. Hunters make periodic visits, but the area territory is not used by stock breeders because of the almost complete absence of vegetation.

Indications of the presence of petroleum have recently been found in the two regions—Kashgar–Khargalik at the western part of the Tarim Basin and Aksu Kucha at the northern edge of the basin. (M.P.Pe.)

Talleyrand

Charles-Maurice de Talleyrand-Périgord, a French statesman and diplomat notable for his capacity for political survival, held high office during the French Revolution, under Napoleon, at the restoration of the Bourbon monarchy, and under Louis-Philippe. As Louis XVIII's delegate to the Congress of Vienna, his skill in dividing the four allies, who had defeated Napoleon, won for France an effective voice in the settlement of Europe.

Giraudon



Talleyrand, portrait by P.P. Prud'hon, 1809. In a private collection, Paris.

EARLY LIFE

Education. Talleyrand was born in Paris on February 2, 1754, the son of Charles-Daniel, comte de Talleyrand-Périgord, and of Alexandrine de Damas d'Antigny. His parents came from old, aristocratic families but were not rich. Charles-Maurice was sent to be nursed in a Paris suburb, where, when he was four years old, he is said to have fallen off a chest of drawers, dislocating his foot. It is possible, however, that his clubfoot was congenital, linked, perhaps, to a brain disorder. At any rate, Talleyrand's clubfoot was of cardinal importance in his choice of career. As he could not follow the family tradition by going into the army, his parents intended him for the church. From the age of eight he was a pupil at the Collège d'Harcourt in Paris, and at 15 he became an assistant to his uncle Alexandre, then coadjutor to the archbishop of Reims, in the hope that the luxurious life led by the princes of the church would awaken in him a taste for an ecclesiastical career. He liked what he saw, and in 1770 he entered the seminary of Saint-Sulpice in Paris. There, he certainly learned theology; but he also read, in the seminary's large library, the works of the Philosophes, the contemporary progressive thinkers. Thus he began his political education and acquired a skepticism regarding men and affairs that was never to leave him. In addition, while still a seminarian he took his first mistress.

Clerical career. Expelled for his conduct (1775), he nevertheless received minor orders in April of that year and was, six months later, nominated by the King as abbot of Saint-Denis, at Reims. In March 1778 Talleyrand received his degree in theology from the Sorbonne, and in December 1779 he was ordained. The following day his uncle Alexandre, now archbishop of Reims, appointed him his vicar general.

At that time Talleyrand appeared to be a typical court cleric, spending more time with the most fashionable wits and beauties of the day than with churchmen. Yet he did not devote all his time to pleasure; he believed in his future and desired above all to become a bishop. The surest way to his goal was appointment to the well-paid post of agent general of the clergy, who represented the French church in its dealings with the French government in the interval between meetings of the Assembly of the Clergy, which were held regularly every five years. Talleyrand was appointed agent general in 1780. There were in fact, two agents general; but his colleague's reputation had been undermined by a notorious scandal, and Talleyrand was in practice the sole representative of the French church between 1780 and 1785. He seems to have taken his role very seriously: in any case, he showed great energy in defending the most archaic as well as the most controversial privileges of the church. He fought vigorously for the right of the church to retain all its property and did his best to smother discussion of the question (which had arisen in the Habsburg Empire) of whether the church should sell part of its property in order to relieve the state's financial distress. He also tenaciously defended ecclesiastical jurisdiction against the encroachments of royal justice, argued for the continued exemption of the clergy from ordinary taxes, and played a part in raising the standard of living of the lower clergy. His activities brought him into regular contact with various ministers of the crown. His participation in the meetings of the Assembly of the Clergy gave him an exceptional opportunity to acquire parliamentary experience. Finally, his activities won him the coveted bishopric: in November 1788 he was appointed bishop of Autun. When he took possession of his see on March 15, 1789, the Revolution was an the verge of breaking out.

Appointment as agent general of the clergy

CAREER AS STATESMAN AND DIPLOMAT

During the Revolution. His first task was to prepare the elections to the States General, the National Assembly in which the Estates (nobility, clergy, and Third Estate) were separately represented. The assembly had not met in France since 1614 and was now summoned for May 5, 1789. Talleyrand, elected by his clergy as their deputy, like all delegates had prepared a list of grievances. His list contained demands for reforms in the status of the clergy as well as for a constitution that would provide a genuine representative government guaranteeing equality for all citizens, especially fiscal equality, thus abolishing those financial privileges that he had defended four years earlier. This was the first of a long series of reversals that were to be characteristic of his career.

At the solemn opening of the States General at Versailles, Talleyrand was described by a witness as having "an angelic face animated by a diabolical spirit." The parliamentary experience he had gained at the meetings of the Assembly of the Clergy stood him in good stead: he immediately attracted attention and soon acquired great influence. During the earlier deliberations, the clergy and the nobility wanted to sit in separate chambers, as they had in 1614; Talleyrand, however, favoured uniting the three orders into one National Assembly, and this proposal was accepted on June 27.

To the great astonishment of his fellow bishops who remembered the zeal with which he had defended the church's rights in 1780 and 1785, Talleyrand, during the famous night of August 4, when the abolition of feudalism was voted, urged the repeal of the tithe. The speech created a sensation, foreshadowing his proposal, on October 10, 1789, to nationalize French church property (amounting to about 15 percent of the territory of France). The land thus appropriated was to be used to pay off the state's debts. When nationalization was voted on November 2, Talleyrand emerged as one of the most revolutionary deputies. Many other signs confirmed the revolutionary position adopted by the "limping devil." In addition to proposing the introduction of a new, uniform system of measures, the decimal system, he also drafted a plan that abolished the old system of education. When he celebrated mass at the Festival of the Federation on July 14, 1790, in commemoration of the taking of the Bastille

Advocacy of abolition of ecclesiastical privileges

the year before, he seemed the veritable "bishop of the Revolution." The same month Talleyrand succeeded in having the Civil Constitution of the Clergy adopted, which, without papal approval, completely reorganized the French church on a democratic basis. The first bishop to take the oath of loyalty to this constitution, he also consecrated the first bishops elected according to the new procedure. The resulting excommunication by the pope did not distress Talleyrand, who was already planning to leave the church. It no longer offered sufficient scope for his ambition, and, since, under the law expropriating the church he had been deprived of his property, a bishopric could no longer satisfy his monetary needs. Elected administrator of the *département* of Paris in January 1791, he resigned as bishop of Autun. On the death some weeks later of the comte de Mirabeau, the most influential deputy in the National Assembly, Talleyrand, the only member with a comparable reputation, had hoped to replace Mirabeau as secret counselor to Louis XVI. He was, however, deprived of the opportunity by the royal family's unsuccessful flight in June and the dissolution of the National Assembly in September.

Civil servant and diplomat

As deputy to the National Assembly, Talleyrand was not eligible for re-election to the assembly replacing it, the Legislative Assembly. Yet his skill as a clever negotiator had been noted, and when at the close of 1791 the French government wanted to prevent England and Prussia from joining Austria in a coalition against France, the foreign minister decided to send Talleyrand to London to persuade England to remain neutral.

Arriving in London in January 1792, Talleyrand proposed to William Pitt, the prime minister, that both countries should guarantee each other's territorial integrity: France specifically recognizing British sovereignty in Ireland and India. Returning to Paris in March without a definite answer from the British, Talleyrand persuaded Charles-François Dumouriez, the new foreign minister, to appoint the young Marquis de Chauvelin as ambassador to London, and returned there as his assistant. The two men arrived in London on April 29, just after France had declared war on Austria, with whom Prussia immediately allied itself. Although Talleyrand secured a declaration of neutrality from the British government on May 25, the storming of the Tuileries palace by the Paris mob on June 20 made his position difficult, and he left London on July 5. The overthrow of the monarchy on August 10 and the massacre of royalist prisoners in September alienated what sympathy the London government had for France and, at the same time, made it advisable for Talleyrand—suspected as a lukewarm revolutionary by the extremists—to leave Paris. After drafting—as a pledge of loyalty to the Provisional Executive Council—a circular to all European governments attributing responsibility for the events of August 10 to Louis XVI, he obtained a passport to go to London in a private capacity. Arriving on September 18, he made every effort to avert war with Great Britain, but the invasion of Belgium by the French, opening the Scheldt Estuary to international shipping (November 1792), followed by the execution of Louis XVI in January 1793, made war inevitable. Talleyrand, who had been denounced in the National Convention (the assembly elected after the overthrow of the monarchy) for conducting a secret correspondence with Louis XVI, also became undesirable in England, where the most counterrevolutionary of the French émigrés were demanding his expulsion. Expelled in January 1794, he embarked for the United States in March. He remained there for two years, engaging in profitable financial speculations that enabled him to rebuild his fortune.

Foreign residence

During the Directory. After the fall, in July 1794, of Maximilien Robespierre, the chief instigator of the Reign of Terror, Talleyrand petitioned the National Convention to remove his name from the list of émigrés, as he had left France on an official passport. His request was granted in September 1795, but he did not reach Paris until September 1796. As soon as he arrived he took his seat in the Institut National (a creation of the National Convention re-establishing, in a new form, the 18th-century academies, among them the Académie Française), to which he had been elected during his absence. The paper that he read there in July 1797, in which he concluded that France would be unable to reconquer its American colonies and should therefore attempt to establish colonies in Africa, particularly in Egypt, showed that he again hoped to enter politics. A few days later, his paper, which raised him in public esteem, and his connections with a member of the Directory governing France at the time gained him the post of foreign minister, although the four other members of the Directory did not favour him. Talleyrand's position was confirmed a few weeks later by the coup d'etat of September 4, 1797, as a result of which the two directors most hostile to him were dismissed.

As foreign minister Talleyrand confirmed Napoleon's conclusion of the Treaty of Campo Formio (October 1797) following his great victories against Austria. Talleyrand negotiated the agreements annexed to the treaty, which are said to have brought him more than a million francs in bribes. Jointly with Napoleon he urged on the Directory his idea of a military expedition to Egypt (May 1798). Led by Napoleon, the expedition, after brilliant initial successes, ended in failure. Talleyrand alone, however, was responsible for a breach between France and the United States, following the indignant withdrawal of three American envoys from whom Talleyrand had demanded enormous bribes. They had come to Paris to negotiate a treaty to replace the one of 1778, but the breakdown in negotiations led to an undeclared war between the United States and France. Moreover, the failure of the Egyptian expedition had resulted in the formation of a new coalition against France composed of England, Austria, Russia, and Turkey. Acknowledging the failure of his policies, Talleyrand resigned on June 20, 1799. After two years as foreign minister he had, by various means, amassed, as he put it, an "immense fortune" of more than three million francs, which he deposited abroad.

Partnership with Bonaparte

During the Consulate and the empire. Five months after Talleyrand's resignation, Napoleon returned from Egypt, and, following his coup d'etat of November 9–10, 1799, established the Consulate, consisting of him as the actual ruler and two other consuls. Talleyrand supported him and returned to the foreign ministry on November 22. He was to stay there until 1807. Talleyrand's principal aim was the pacification of Europe, and he immediately began negotiations with the belligerent nations. Operating from a position of strength after Napoleon's victories in Italy and in Germany, his negotiations with Austria were the first to result in a peace treaty, signed at Lunéville on February 9, 1801. Negotiations with England, in which Talleyrand played a less direct part, led to the Treaty of Amiens (March 27, 1802). For the first time in six years, Europe was at peace. Talleyrand contributed to the realization of Napoleon's ambitious plans to remodel Europe by helping him to establish French supremacy in Italy, Germany, and Switzerland. At the meeting of leading members of the Cisalpine Republic at Lyons early in 1802, he engineered various changes in the republic's new constitution (including changing its name to the more promising one of Italian Republic) and the election of Napoleon as its president. Talleyrand was also the chief author of the treaty of 1803, which extensively altered Germany's political geography. Much to his own profit, he supervised the allocation of numerous secularized church lands as compensation to those princes who had lost to France territory on the left bank of the Rhine.

Return to foreign ministry

At home Talleyrand used his influence to facilitate the return of émigrés and priests who had refused the oath of loyalty to the various constitutions since 1790 and urged the signing of the Concordat between Napoleon and Pope Pius VII (July 1801), which re-established religious peace. Then, taking advantage of the Concordat's provisions, he married his mistress, Madame Catherine Grand, the divorced French wife of an English employee of the British East India Company.

Talleyrand's policy would have been completely successful had he been able to prevent the renewal of war between France and England in May 1803. This time,

however, he did not resign. He had helped Napoleon establish himself as "consul for life" in 1802, and he continued to support him when, after the resumption of the war, Napoleon wanted to show, by means of some striking action, that he would never come to terms with the Bourbons. And so Talleyrand participated in one of the most dreadful crimes with which Napoleon has ever been charged. When Talleyrand and Joseph Fouché, the minister of police, learned that a Bourbon prince, whom they believed to be Louis-Antoine-Henri de Bourbon-Condé, duc d'Enghien, was planning the assassination of the First Consul, they advised his abduction. Although the duke was living in neutral territory, in the duchy of Baden, Talleyrand promised that he would smooth over any protests resulting from the violation of international law. And so the Duc d'Enghien was kidnapped, arrested, and transferred to Paris, where he was tried, condemned to death, and executed (March 21, 1804). Later, Talleyrand tried to remove from the archives documents proving his involvement in the death of the Duc d'Enghien. It was this crime that consolidated Napoleon's power, and when on May 18, 1804, he was proclaimed emperor, he appointed Talleyrand grand chamberlain of the empire, with an annual income of 500,000 francs. Later, Talleyrand was also appointed to the sinecure of vice elector; finally, on June 5, 1806, he became prince and duke of Benevento (a city in southern Italy that had been seized from the papacy).

Murder of the Duc d'Enghien

Nevertheless, after 1805 Talleyrand's influence diminished. He failed to persuade Napoleon to conciliate Austria after his victory at Austerlitz or to preserve the Holy Roman Empire, which was replaced by the Confederation of the Rhine and the Austrian Empire. In 1806, the negotiations he had opened with a view to restoring peace with England broke down, and a Franco-Russian peace treaty signed in Paris by Talleyrand and by the Russian ambassador was repudiated by the Tsar. When France's war with Russia and Prussia was finally ended by the Treaties of Tilsit (July 7 and 9, 1807), Talleyrand had no part in the negotiations. Alarmed by Napoleon's insatiable ambition, which, as he clearly saw, could lead only to disaster, he resigned his office in August 1807. It was not without pleasure that Napoleon accepted it. He had received innumerable complaints about Talleyrand's venality and, even was beginning to doubt his loyalty.

Between empire and restoration. Though no longer a minister, Talleyrand was still consulted by Napoleon, and it seems likely that he did not advise against the French invasion of Spain. In September 1808 Talleyrand accompanied Napoleon to a congress of European sovereigns at Erfurt. There Talleyrand had secret talks with Tsar Alexander I, urging him to oppose Napoleon, and thereafter conducted a clandestine correspondence with both Russia and Austria. This treasonable activity did not, in fact, involve Talleyrand in any great risk, since it was approved by Fouché, the minister of police, who shared Talleyrand's opposition to Napoleon's policies.

After Napoleon had divorced the empress Joséphine (1810), Talleyrand played a part in arranging the Emperor's marriage with Marie-Louise of Austria, daughter of Emperor Francis I, in the hope that this union would modify Napoleon's ambition. But nothing, apparently, could accomplish that, and when Talleyrand learned that the imperial army had invaded Russia, he declared that it was "the beginning of the end." After the disastrous retreat from Russia, Napoleon, in January 1814, asked him to return to the foreign ministry in order to negotiate with the allies, but Talleyrand, who was already planning to restore the Bourbons, refused, unmoved by the Emperor's anger. When the allies entered Paris, on March 31, 1814, the Tsar took up residence at Talleyrand's mansion and was eventually convinced by him that only the restoration of the Bourbons could guarantee peace in Europe. Talleyrand persuaded the Senate (only 63 of whose 140 members were present) to establish a provisional government of five members, including himself, and to declare Napoleon deposed. The new government immediately recalled Louis XVIII, who on May 13, 1814, appointed Talleyrand his foreign minister.

During the Restoration and the July Monarchy. As France's representative at the Congress of Vienna (1814–15), which was convened to rebuild Europe after 23 years of upheaval, Talleyrand exhibited his diplomatic skill to the full, but it is doubtful whether it benefitted France. He managed to divide the Allies, urging Austria and England to conclude secret agreements with France to prevent Russia from annexing the whole of Poland, and Prussia from annexing the whole of Saxony. This new triple alliance succeeded in reducing the territorial claims of the other great powers, and led to the agreement by which France retained the frontiers of 1792 (which were pushed back to those of 1790 after the Hundred Days, the period during which Napoleon reigned in Paris after his escape from Elba). But, in accepting the cession to Prussia of the greater part of the left bank of the Rhine, Talleyrand created a serious danger for France that became especially clear in 1870, 1914, and 1939.

Congress of Vienna

Talleyrand remained in Vienna during the Hundred Days. On Louis XVIII's return to Paris, he was appointed president of the council, while retaining the office of foreign minister. But the ultraroyalists now in power were violently opposed to a ministry dominated by two former revolutionaries, Fouché and Talleyrand. Talleyrand, moreover, had lost the Tsar's support by his tactics at Vienna, and on September 24, 1815, he was forced to resign. He then lived in retirement, writing his memoirs, until 1829, when his scheming political sense led him to ally himself with the Liberals in order to remove Charles X, the brother and successor of Louis XVIII. He established contact with Louis-Philippe, duc d'Orléans, and helped to make him king during the July Monarchy of 1830. As Louis-Philippe's ambassador in London, from 1830 to 1834, he played a vital part in the negotiations between France and Britain that resulted in the creation of a neutral kingdom of Belgium. His diplomatic career was crowned by the signing of an alliance between France, Great Britain, Spain, and Portugal in April 1834.

Talleyrand died in Paris on May 17, 1838. He received the last sacraments, having signed, a few hours before his death, a document in which he declared himself reconciled with the church and expressed regret for his former conduct. He was buried at his château of Valençay. He had separated from his wife in 1815 and left no legitimate descendants.

Assessment

Talleyrand was indisputably a man of exceptional intelligence and a skillful diplomat. During the 19th century many liberals persisted in acclaiming him as the great statesman whose political realism restored peace and order to Europe at the Congress of Vienna; and during the many international crises of the 20th century, some historians have regretted the absence of a man of his extraordinary diplomatic gifts. But his total lack of principle, his treachery and venality, leave a sense of distaste that his talents, even though almost always used in the service of peace and liberalism, cannot totally dispel.

BIBLIOGRAPHY

Sources: Talleyrand's *Mémoires* were edited by the DUC DE BROGLIE, 5 vol. (1891–92; Eng. trans. of vol. 3–5, 1891–92). They are unreliable and should be used with discretion. A number of Talleyrand's letters and dispatches have been published in the following works: G. PALLAIN, *La Mission de Talleyrand à Londres en 1792* (1889); *Le Ministère de Talleyrand sous le Directoire* (1891); *Correspondance inédite du prince de Talleyrand et du roi Louis XVIII pendant le Congrès de Vienne* (1881); and *L'ambassade de Talleyrand à Londres, 1830–1834*, 2 vol. (1891); P. BERTRAND, *Lettres inédites de Talleyrand à Napoléon, 1800–1809* (1889); H. HUTH and W.J. PUGH, *Talleyrand in America As a Financial Promoter, 1794–1796*, 3 vol. (1942).

Biographies and monographs: The basic biography of Talleyrand is G. LACOUR-GAYET, *Talleyrand, 1754–1838*, 4 vol. (1928–34), which has not been translated into English. The author has consulted the main sources, both printed and in manuscript. This work has been superseded in certain areas but has not been surpassed. The study by E.L. DARD, *Napoléon et Talleyrand* (1935; Eng. trans., 1937), which is based on extensive work in the Austrian archives, should also be noted. The following biographies contain subjective judgments of Talleyrand: A. DUFF COOPER, *Talleyrand* (1932); the COMTE

DE SAINT-AULAIRE, *Talleyrand* (1935); CRANE BRINTON, *The Lives of Talleyrand* (1936); G. FERRERO, *Talleyrand à Vienne, 1814–1815* (1940); I. MADELIN, *Talleyrand* (1942); J. BERTAUT, *Talleyrand* (1945); and J. ORIEUX, *Talleyrand, ou la fuite au désert* (1970), all favourable accounts; R. LAFORGUE, *Talleyrand, l'homme de la France* (1947), a psychoanalytical study; M. MISSOFFE, *Le Coeur secret de Talleyrand* (1956), on the romantic life of Talleyrand; E. TARLE, *Talleyrand* (French trans. from the Russian, 1958), hostile; and L.S. GREENBAUM, *Talleyrand, Statesman-Priest* (1970), an original study of Talleyrand's ecclesiastical career before 1789.

Film: *Le Diable Boiteux* by SACHA GUITRY, the famous French actor, is a pleasant and amusing film that takes considerable liberties with historical truth.

(J.Go.)

Talmud and Midrash

Talmud and Midrash are commentative and interpretative writings that hold a place in the Jewish religious tradition second only to the Bible (Old Testament).

DEFINITION OF TERMS

The Hebrew term Talmud ("study" or "learning") commonly refers to a compilation of ancient teachings regarded as sacred and normative by Jews from the time when it was compiled until modern times and still so regarded by traditional religious Jews. In this broad sense, it is a set of books consisting of the Mishna ("repeated study"), the Gemara ("completion"), and certain auxiliary materials. The Mishna is a collection of originally oral laws supplementing scriptural laws. The Gemara is a collection of commentaries on and elaborations of the Mishna, which in "the Talmud" is reproduced in juxtaposition to the Gemara (see illustration). For present-day scholarship, however, Talmud in the precise sense refers only to the materials customarily called Gemara—an Aramaic term prevalent in medieval rabbinic literature that was used by the church censor to replace the term Talmud within the Talmudic discourse in the Basel edition of the Talmud, published 1578–81. This practice continued in all later editions.

Mishna and Gemara

The term Midrash ("exposition" or "investigation"; plural, Midrashim) is also used in two senses. On the one hand, it refers to a mode of biblical interpretation prominent in the Talmudic literature; on the other, it refers to a separate body of commentaries on Scripture using this interpretative mode.

NATURE AND FUNCTIONS

Role in Jewish religion and society. Torah in Judaism refers to the divine guidance, teaching, or law for the Jewish people (and, ultimately, for mankind). In its written form, it was considered to be especially present in the first five books of the Bible (the Pentateuch), which therefore came to be called Torah. In addition to this written Torah, or "Law," there were also unwritten laws or customs and interpretations of them, carried down in an oral tradition over many generations, which acquired the status of oral Torah.

The Talmud is the literary culmination of this oral tradition, which, according to the rabbis who created the Talmud, originated at Mt. Sinai as part of the divine revelation vouchsafed to Moses, along with the material recorded in the Pentateuch. The oral tradition interpreted the written Torah, adapted its precepts to ever-changing political and social circumstances, and supplemented it with new legislation. Thus the oral tradition added a dynamic dimension to the written code, making it a self-regenerating, endless source of guidance, a perpetual process rather than a closed system. The vitality of this tradition is fully demonstrated in the way the ancient laws were adapted after the destruction of the Temple in 70 CE (Common Era = AD) and by the role the Talmud played in the survival of the Jewish people in exile. By the 11th century, Diaspora Jews lived within a Talmudic culture that unified them and that superseded geographical boundaries and language differences. Jewish communities governed themselves according to Talmudic law, and individuals regulated the smallest details of their lives by it.

But despite the central place of the Talmud in traditional Jewish life and thought, significant Jewish groups and individuals have opposed it vigorously. The Karaite sect in Babylonia, beginning in the 8th century, refuted the oral tradition and denounced the Talmud as a rabbinic fabrication. Medieval Jewish mystics declared the Talmud a mere shell covering the concealed meaning of the written Torah, and heretical messianic sects in the 17th and 18th centuries totally rejected it. The decisive blow to Talmudic authority came in the 18th and 19th centuries when the Haskala (the Jewish Enlightenment movement) and its aftermath, Reform Judaism, secularized Jewish life and, in doing so, shattered the Talmudic wall that had surrounded the Jews. Thereafter, modernized Jews usually rejected the Talmud as a medieval anachronism, denouncing it as legalistic, casuistic, devitalized, and unspiritual.

Anti-Talmudic positions

There is also a long-standing anti-Talmudic tradition among Christians. The Talmud was frequently attacked by the church, particularly during the Middle Ages, and accused of falsifying biblical meaning, thus preventing Jews from becoming Christians. The church held that the Talmud contained blasphemous remarks against Jesus and Christianity and that it preached moral and social bias toward non-Jews. On numerous occasions the Talmud was publicly burned, and permanent Talmudic censorship was established.

On the other hand, since the Renaissance there has been a positive response and great interest in rabbinic literature by eminent non-Jewish scholars, writers, and thinkers in the West. As a result, rabbinic ideas, images, and lore, embodied in the Talmud, have permeated Western thought and culture.

Content, style, and form. The Talmud is first and foremost a legal compilation. At the same time it contains materials that encompass virtually the entire scope of subject matter explored in antiquity. Included are topics as diverse as agriculture, architecture, astrology, astronomy, dream interpretation, ethics, fables, folklore, geography, history, legend, magic, mathematics, medicine, metaphysics, natural sciences, proverbs, theology, and theosophy.

This encyclopaedic array is presented in a unique dialectic style that faithfully reflects the spirit of free give-and-take prevalent in the Talmudic academies, where study was focussed upon a Talmudic text. All present participated in an effort to exhaust the meaning and ramifications of the text, debating and arguing together. The mention of a name, situation, or idea often led to the introduction of a story or legend that lightened the mood of a complex argument and carried discussion further.

This text-centred approach profoundly affected the thinking and literary style of the rabbis. Study became synonymous with active interpretation rather than with passive absorption. Thinking was stimulated by textual examination. Even original ideas were expressed in the form of textual interpretations.

The subject matter of the oral Torah is classified according to its content into Halakha and Haggada and according to its literary form into Midrash and Mishna. Halakha ("law") deals with the legal, ritual, and doctrinal parts of Scripture, showing how the laws of the written Torah should be applied in life. Haggada ("narrative") expounds on the nonlegal parts of Scripture, illustrating biblical narrative, supplementing its stories, and exploring its ideas. The term Midrash denotes the exegetical method by which the oral tradition interprets and elaborates scriptural text. It refers also to the large collections of Halakhic and Haggadic materials that take the form of a running commentary on the Bible and that were deduced from Scripture by this exegetical method. In short, it also refers to a body of writings. Mishna is the comprehensive compendium that presents the legal content of the oral tradition independently of scriptural text.

Halakha and Haggada; Midrash and Mishna

Modes of interpretation and thought. Midrash was initially a philological method of interpreting the literal meaning of biblical texts. In time it developed into a sophisticated interpretive system that reconciled apparent

biblical contradictions, established the scriptural basis of new laws, and enriched biblical content with new meaning. Midrashic creativity reached its peak in the schools of Rabbi Ishmael and Akiba, where two different hermeneutic methods were applied. The first was primarily logically oriented, making inferences based upon similarity of content and analogy. The second rested largely upon textual scrutiny, assuming that words and letters that seem superfluous teach something not openly stated in the text.

The Talmud (*i.e.*, the Gemara) quotes abundantly from all Midrashic collections and concurrently uses all rules employed by both the logical and textual schools; moreover, the Talmud's interpretation of Mishna is itself an adaptation of the Midrashic method. The Talmud treats the Mishna in the same way that Midrash treats Scripture. Contradictions are explained through reinterpretation. New problems are solved logically by analogy or textually by careful scrutiny of verbal superfluity.

Midrashic–Talmudic thinking

The strong involvement with hermeneutic exegesis—interpretation according to systematic rules or principles—helped develop the analytic skill and inductive reasoning of the rabbis but inhibited the growth of independent abstract thinking. Bound to a text, they never attempted to formulate their ideas into the type of unified system characteristic of Greek philosophy. Unlike the philosophers, they approached the abstract only by way of the concrete. Events or texts stimulated them to form concepts. These concepts were not defined but, once brought to life, continued to grow and change meaning with usage and in different contexts. This process of conceptual development has been described by some as "organic thinking." Others use this term in a wider sense, pointing out that, although rabbinic concepts are not hierarchically ordered, they have a pattern-like organic coherence. The meaning of each concept is dependent upon the total pattern of concepts, for the idea content of each grows richer as it interweaves with the others.

ORIGINS AND DEVELOPMENT

Ezra the scribe who, according to the Book of Ezra, reestablished and reformed the Jewish religion in the 5th century BCE (Before the Common Era = BC), began the "search in the Law . . . to teach in Israel statutes and ordinances."

His work was continued by *soferim* (scribes), who preserved, taught, and interpreted the Bible. They linked the oral tradition to Scripture, transmitting it as a running commentary on the Bible. For almost 300 years they applied the Torah to changing circumstances, making it a living law. They also introduced numerous laws that were designated "words of the *soferim*" by Talmudic sources. By the end of this period, rabbinic Judaism—the religious system constructed by the scribes and rabbis—was strong enough to withstand pressure from without and mature enough to permit internal diversity of opinion.

At the beginning of the 2nd century BCE, a judicial body headed by the *zugot*—pairs of scholars—assumed Halakhic authority. There were five pairs in all, between *c.* 150 and 30 BCE. The first of the *zugot* also introduced the Mishnaic style of transmitting the oral tradition.

The *tannaim*

The making of the Mishna: 2nd–3rd centuries. Hillel and Shammai, the last of the *zugot*, ushered in the period of the *tannaim*—"teachers" of the Mishna—at the end of the 1st century BCE. This era, distinguished by a continuous attempt to consolidate the fragmentary Midrashic and Mishnaic material, culminated in the compilation of the Mishna at the beginning of the 3rd century CE. The work was carried out in the academies of Hillel and Shammai and in others founded later. Most scholars believe that Halakhic collections existed prior to the fall of Jerusalem, in CE 70. Other compilations were made at Yavne, a Palestinian town near the Mediterranean, as part of the effort to revitalize Judaism after the disaster of CE 70 (see JOHANAN BEN ZAKKAI). By the beginning of the 2nd century there were many such collections. Tradition has it that Rabbi Akiba organized much of this material into separate collections of Midrash, Mishna, and Haggada and introduced the formal divisions in tannaitic literature. His students and other scholars organized new compilations that were studied in the different academies.

After the rebellion of the Jews against Roman rule led by Simeon bar Kokhba in 132–135, when the Sanhedrin (the Jewish supreme court and highest academy) was revived, the Mishnaic compilation adopted by the Sanhedrin president became the official Mishna. The Sanhedrin reached its highest stature under the leadership of Judah ha-Nasi (Judah the Prince, or President); he was also called Rabbi, as the pre-eminent teacher.

It seems certain that the official Mishna studied during his presidency was the Mishna we know and that he was its editor. Judah aimed to include the entire content of the oral tradition. He drew heavily from the collections of Akiba's pupils but also incorporated material from other compilations, including early ones. Nevertheless, the accumulation was such that selection was necessary. Thus almost no Midrash or Haggada was included. Colleagues and pupils of Judah not only made minor additions to the Mishna but tried to preserve the excluded material, the Baraitot ("Exclusions"), in separate collections. One of these was the Tosefta ("Addition"). Midrashic material was gathered in separate compilations, and later revisions of some of these are still extant. The language of all of the tannaitic literature is the new Hebrew developed during the period of the Second Temple (*c.* 6th century BCE–1st century CE).

The *amoraim*

The making of the Talmuds: 3rd–6th centuries. The expounders of the Mishna were the *amoraim* ("interpreter"), and the two Talmuds—the Palestinian (or Jerusalem) and the Babylonian—consist of their explanations, discussions, and decisions. Both take the form of a running commentary on the Mishna.

The foundations for these two monumental works were begun by three disciples of Judah ha-Nasi: Johanan bar Nappaḥa, Rav (Abba Arika), and Samuel bar Abba, in their academies at Tiberias, in Palestine, and at Sura and Nehardea in Babylonia, respectively. Centres of learning where the Mishna was expounded existed also at Sephoris, Caesarea, and Lydda in Palestine. In time new academies were established in Babylonia, the best known being those at Pumbedita, Mahoza and Naresh, founded by Judah bar Ezekiel, Rava, and Rav Pappa, respectively. The enrollment of these centres often numbered in the thousands, and students spent many years there. Those who no longer lived on the academy grounds returned twice annually for the *kalla*, a month of study in the spring and fall.

Academies differed in their methods of study. Pumbedita, for example, stressed casuistry, while Sura emphasized breadth of knowledge. Students often moved from one academy to another and even from Palestine to Babylonia or from Babylonia to Palestine. This kept open the channels of communication between the various academies and resulted in the inclusion of much Babylonian material in the Palestinian Talmud, and vice versa.

Despite the overwhelming similarity of the two Talmuds, however, they do differ in some ways. The Palestinian Talmud is written in the Western Aramaic dialect, the Babylonian in the Eastern. The former is invariably shorter, and, not having been subject to final redaction, its discussions are often incomplete. Its explanations tend to remain closer to the literal meaning of the Mishna, preferring textual emendation to casuistic interpretation. Finally, some of the legal concepts in the Babylonian Talmud reflect the influence of Persian law, for Babylonia was under Persian rule at the time.

The main endeavour of the *amoraim* was to thoroughly explain and exhaust the meaning of the Mishna and the Baraitot. Apparent contradictions were reconciled by such means as explaining that conflicting statements referred to different situations or by asserting that they stemmed from the Mishnayot (Mishnas) of different *tannaim*. The same techniques were used when amoraic statements contradicted the Mishna. These discussions took place for hundreds of years, and their content was passed on from generation to generation, until the compilation of the Talmud.

The compilation of the Palestinian and Babylonian Talmuds

The portion of the Palestinian Talmud dealing with the three Bavot ("gates")—*i.e.*, the first three tractates of the fourth order of the Mishna (for orders and tractates, see *Talmudic and Midrashic literature*, below)—was compiled in Caesarea in the middle of the 4th century and is distinguished from the rest by its brevity and terminology. The remainder was completed in Tiberias some 50 years later. It seems likely that its compilation was a rescue operation designed to preserve as much of the Halakhic material collected in Palestinian academies as possible, for by that time the deterioration of the political situation had forced most Palestinian scholars to emigrate to Babylonia.

The Babylonian Talmud was compiled up to the 6th century. Some scholars suggest that the organization of the Talmud began early and that successive generations of *amoraim* added layer upon layer to previously arranged material. Others suggest that at the beginning a stratum called Gemara, consisting only of Halakhic decisions or short comments, was set forth. Still others theorize that no overall arrangement of Talmudic material was made until the end of the 4th century.

The statement in the tractate *Bava metzi'a* that "Rabina and Rav Ashi were the end of instruction" is most often understood as referring to the final redaction of the Talmud. Since at least two generations of scholars following Rav Ashi (died 427) are mentioned in the Talmud, most scholars suggest that "Rabina" refers to Rabina bar Huna (died 499) and that the redaction was a slow process lasting about 75 years to the end of the 5th century.

According to the tradition of the *geonim*—the heads of the academies at Sura and Pumbedita from the 6th to the 11th centuries—the Babylonian Talmud was completed by the 6th-century *savoraim* ("expositors"). But the extent of their contribution is not precisely known. Some attribute to them only short additions. Others credit them with creating the terminology linking the phases of Talmudic discussions. According to another view, they added comments and often decided between conflicting opinions. The proponents of the so-called Gemara theory noted above ascribe to them the entire dialectic portion of Talmudic discourse.

TALMUDIC AND MIDRASHIC LITERATURE

Mishna. The Mishna is divided into six orders (*sedarim*), each order into tractates (*massekhtot*), and each tractate into chapters (*peraqim*). The six orders are *Zera'im, Mo'ed, Nashim, Neziqin, Qodashim*, and *Ṭohorot*.

Zera'im, Mo'ed, and *Nashim*

1. *Zera'im* ("Seeds") consists of 11 tractates: *Berakhot, Pea, Demai, Kilayim, Shevi'it, Terumot, Ma'aserot, Ma'aser sheni, Ḥalla, 'Orla*, and *Bikkurim*. Except for *Berakhot* ("Blessings"), which treats of daily prayers and grace, this order deals with laws related to agriculture in Palestine. It includes prohibitions against mixtures in plants (hybridization), legislation relating to the sabbatical year (when land lies fallow and debts are remitted), and regulations concerning the portions of harvest given to the poor, the Levites, and the priests.

2. *Mo'ed* ("Season" or "Festival") consists of 12 tractates: *Shabbat, 'Eruvin, Pesaḥim, Sheqalim, Yoma, Sukka, Betza, Rosh Hashana, Ta'anit, Megilla, Mo'ed qaṭan*, and *Ḥagiga*. This order deals with ceremonies, rituals, observances, and prohibitions relating to special days of the year, including the Sabbath, holidays, and fast days. Since the half-shekel Temple contribution was collected on specified days, tractate *Sheqalim*, regarding this practice, is included.

3. *Nashim* ("Women") consists of seven tractates: *Yevamot, Ketubot, Nedarim, Nazir, Soṭa, Giṭṭin*, and *Qiddushin*. This order deals with laws concerning betrothal, marriage, sexual and financial relations between husband and wife, adultery, and divorce. Since Nazirite (ascetic) and other vows may affect marital relations, *Nedarim* ("Vows") and *Nazir* ("Nazirite") are included here.

4. *Neziqin* ("Damages") consists of 10 tractates, the first three of which were originally considered one (the *Bavot*): *Bava qamma, Bava metzi'a, Bava batra, Sanhedrin, Makkot, Shevu'ot, 'Eduyyot, 'Avoda zara, Avot*, and *Horayot*. This order deals with civil and criminal law concerning damages, theft, labour relations, usury, real estate, partnerships, tenant relations, inheritance, court composition, jurisdiction and testimony, erroneous decisions of the Sanhedrin, and capital and other physical punishments. Since idolatry, in the literal sense of worship or veneration of material images, is punishable by death, *'Avoda zara* ("Idolatry") is included. *Avot* ("Fathers"), commonly called "Ethics of the Fathers" in English, seems to have been included to teach a moral way of life that precludes the transgression of law.

Neziqin, Qodashim, and *Ṭohorot*

5. *Qodashim* ("Sacred Things") consists of 11 tractates: *Zevaḥim, Menaḥot, Ḥullin, Bekhorot, 'Arakhin, Temura, Keretot, Me'ila, Tamid, Middot*, and *Qinnim*. This order incorporates some of the oldest Mishnaic portions. It treats of the Temple and includes regulations concerning sacrifices, offerings, and donations. It also contains a detailed description of the Temple complex.

6. *Ṭohorot* ("Purifications") consists of 12 tractates: *Kelim, Ohalot, Nega'im, Para, Ṭohorot, Miqwaot, Niddа, Makhshirin, Zavim, Ṭevul yom, Yadayim*, and *'Uqtzin*. This order deals with laws governing the ritual impurity of vessels, dwellings, foods, and persons, and with purification processes.

Tosefta. The Tosefta ("Addition") closely resembles the Mishna in content and order. In its present form it at times supplements the Mishna, at other times comments on it, and often also opposes it. There is no Tosefta on the tractates *Avot, Tamid, Middot*, and *Qinnim*. The Talmud quotes from many other collections of Mishnaiot and Baraitot: some are attributed to *tannaim*, and predate the established Mishna; and others, to *amoraim*. The original material is lost.

Talmud (Gemara). Although the entire Mishna was studied at the Palestinian and Babylonian academies, the Palestinian Talmud (Gemara) covers only the first four orders (except chapters 21–24 of *Shabbat* and chapter 3 of *Makkot*) and the first three chapters of *Nidda* in the sixth order. Most scholars agree that the Palestinian Talmud was never completed to the fifth and sixth orders of the Mishna and that the missing parts of the other orders were lost. A manuscript of chapter 3 of *Makkot* was, in fact, found and was published in 1946.

Incompleteness of the Talmuds

The Babylonian Talmud does not cover orders *Zera'im* (except *Berakhot*) and *Ṭohorot* (except *Nidda*) and tractates *Tamid* (except chapters 1,2,4), *Sheqalim, Middot, Qinnim, Avot*, and *'Eduyyot*. Scholars concur that the Talmud for these parts was never completed, possibly because their content was not relevant in Babylonia.

Midrashim. *Halakhic.* Halakhic Midrashim are exegetic commentaries on the legal content of Exodus, Leviticus, Numbers, and Deuteronomy. The five extant collections are *Mekhilta*, on Exodus; *Mekhilta deRabbi Shim'on ben Yoḥai*, on Exodus; *Sifra*, on Leviticus; *Sifre*, on Numbers and Deuteronomy; *Sifre zuṭa*, on Numbers. (*Mekhilta* means "measure," a norm or rule; *Sifra*, plural *Sifre*, means "writing" or "book.") Critical analysis reveals that *Mekhilta* and *Sifre* on Numbers differ from the others in terminology and method. Most scholars agree that these two originated in the school of Ishmael and the others in that of Akiba. In their present form they also include later additions. Mention should also be made of *Midrash tannaim* on Deuteronomy, consisting of fragments recovered from the Yemenite anthology *Midrash ha-gadol*.

Haggadic. Haggadic Midrashim originated with the weekly synagogue readings and their accompanying explanations. Although Haggadic collections existed in tannaitic times, extant collections were compiled between the 4th and 11th centuries. Midrashic compilations were not authoritatively edited and tend to be coincidental and fragmentary.

Biblical and nonbiblical Haggadot

Most notable among biblical collections is *Midrash rabba* ("Great Midrash"), a composite of commentaries on the Pentateuch and five Megillot (Song of Songs, Ruth, Ecclesiastes, Esther, Lamentations) differing in nature and age. Its oldest portion, the 5th-century *Genesis*

rabba, is largely a verse-by-verse commentary, while the 6th-century *Leviticus rabba* consists of homilies and *Lamentations rabba* (end of 6th century) is mainly narrative. The remaining portions of *Midrash rabba* were compiled at later dates.

The *Tanḥuma* (after the late-4th-century Palestinian *amora* Tanḥuma bar Abba), of which two versions are extant, is another important Pentateuchal Midrash. Additional Midrashic compilations include those to the books of Samuel, Psalms, and Proverbs. Mention should also be made of *Pesiqta* ("Section" or "Cycles") *deRab Kahana* (after a Babylonian *amora*) and *Pesiqta rabbati* ("The Great Cycle"), consisting of homilies on the Torah (Pentateuch) readings that occur on festivals and special sabbaths.

Haggadic compilations independent of biblical text include *Avot deRabbi Natan*, *Tanna deve Eliyyahu*, *Pirqe* ("Chapters") *deRabbi Eliezer*, and tractates *Derekh eretz* ("Correct Conduct"). These primarily deal with ethics, moral teachings, and biblical narrative.

Among the medieval anthologies are the *Yalquṭ* ("Compilation") *Shimoni* (13th century), *Yalquṭ ha-makhiri* (14th century), and *ʿEn Yaʿaqov* ("Eye of Jacob," 16th century). The two most important modern Haggadic anthologies are those of Wilhelm Bacher and Louis Ginzberg.

Codes. The Talmud's dialectic style and organization are not those of a code of laws. Accordingly, codification efforts began shortly after the Talmud's completion. The first known attempt was *Halakhot pesuqot* ("Decided Laws"), ascribed to Yehudai Gaon (8th century). *Halakhot gedolot* ("Great Laws"), by Simeon Kiyyara, followed 100 years later. Both summarize Talmudic Halakhic material, omitting dialectics but preserving Talmudic order and language. The later *geonim* concentrated on particular subjects, such as divorce or vows, introducing the monographic style of codification.

Codification literature gained impetus by the beginning of the 11th century. During the next centuries many compilations appeared in Europe and North Africa. The most notable, following Talmudic order, were the *Hilkhot Harif*, by Isaac Alfasi (11th century), and *Hilkhot Harosh*, by Asher ben Jehiel (13th–14th centuries). Though modelled after *Halakhot gedolot*, the *Hilkhot Harif* encompasses only laws applicable after the destruction of the Temple but includes more particulars. The *Hilkhot Harosh* closely follows Alfasi's code but often also includes the reasoning underlying decisions.

The Mishne Torah, Sefer ha-ṭurim, and Shulḥan ʿarukh

The most important of the topically arranged codifications were: the *Mishne Torah, Sefer ha-ṭurim*, and *Shulḥan ʿarukh*. (1) The *Mishne Torah* ("The Torah Reviewed") by Maimonides (12th century), is a monumental work, original in plan, language, and order; it encompasses all religious subject matter under 14 headings and includes theosophy, theology, and religion. (2) The *Sefer ha-ṭurim* ("Book of Rows or Parts"), by Jacob ben Asher (14th century), the son of Asher ben Jehiel, introduced new groupings, dividing subject matter into four major categories (*ṭurim*) reminiscent of the Mishnaic orders; it includes only laws applicable after the destruction of the Temple. (3) The *Shulḥan ʿarukh* ("The Prepared Table") by Joseph Karo (16th century), the last of the great codifiers, is structured after the *Sefer ha-ṭurim*, but presents the Sefardic (Near Eastern and North African) rather than the Ashkenazic (Franco-German and eastern European) tradition, with decisions largely following those of Alfasi, Maimonides, and Rabbi Asher. When the 16th-century Ashkenazic codifier Moses Isserles added his notes, this became the standard Halakhic code for all Jewry.

Commentaries. The interpretive literature on the Talmud began with the rise of academies in Europe and North Africa. The earliest known European commentary, though ascribed to Gershom ben Judah (10th–11th centuries), is actually an eclectic compilation of notes recorded by students of the Mayence (Mainz) Academy. Compilations of this kind, known as *quntresim* ("notebooks"), also developed in other academies. Their content was masterfully reshaped and reformulated in the renowned 11th-century commentary of Rashi (acronym of *Ra*bbi *Sh*lomo *Yitz*ḥaqi), in which difficulties likely to be encountered by students are anticipated and detail after detail is clarified until a synthesized, comprehensible whole emerges.

The commentaries of Ḥananel ben Ḥushiel and Nissim ben Jacob ben Nissim, the first to appear in North Africa (11th century), are introductory in nature. They summarize the content of Talmudic discussions, assuming that details will be understood once the general idea becomes comprehensible. This style was later followed by the Spanish school, including Joseph ibn Migash and Maimonides. However, as Rashi's work became known, it displaced all other commentaries. (Note its predominant role in the sample page of Talmud.)

Tosafot, novellae, and responsa

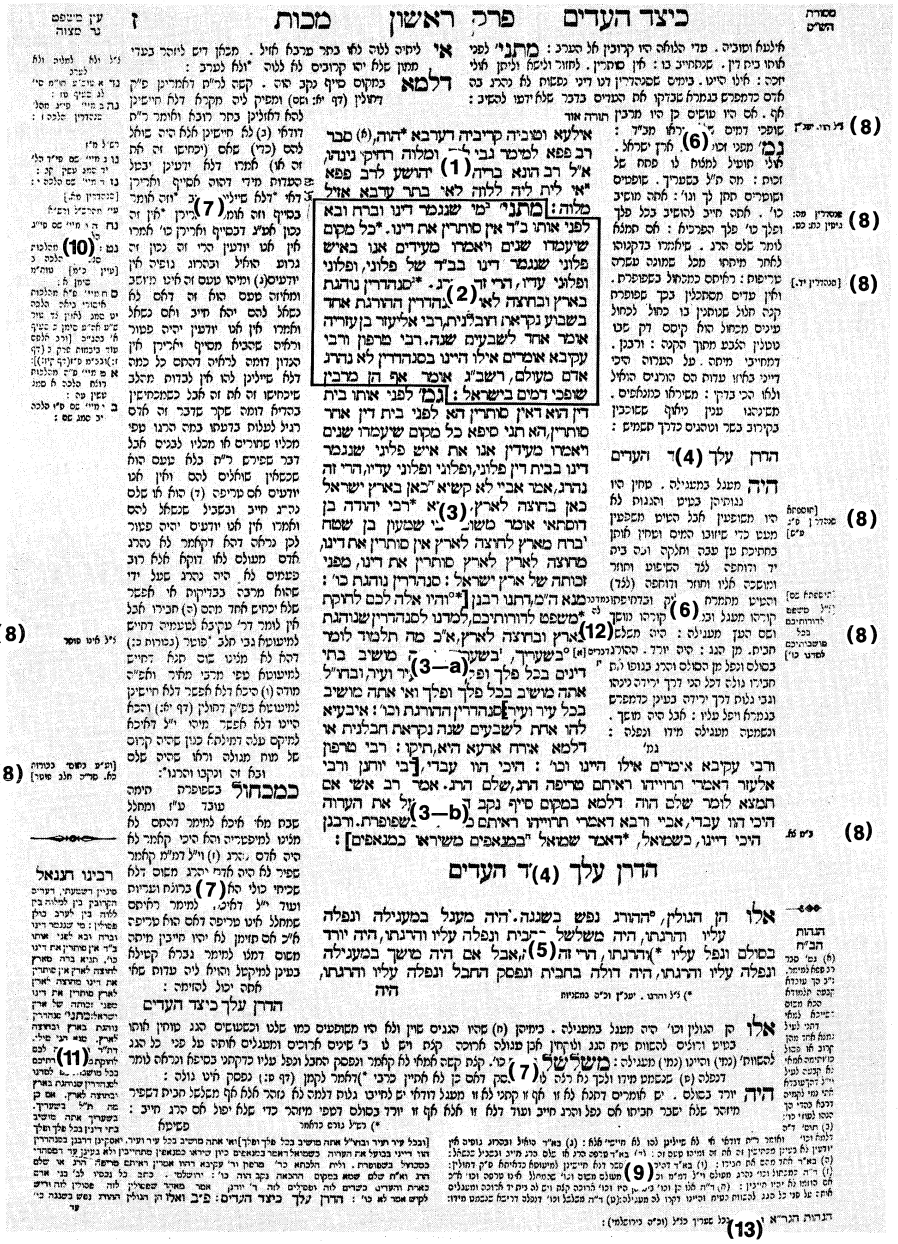
A new phase in Talmudic literature was initiated by Rashi's grandchildren, Rabbis Isaac, Samuel, and Jacob, the sons of Meir, who established the school of *tosafot*. (These medieval "additions" are not to be confused with the tannaitic Tosefta discussed above.) Reviving Talmudic dialectic, they treated the Talmud in the same way that it, in turn, had treated the Mishna. They linked apparently unrelated statements from different Talmudic discourses and pointed out the fine distinctions between seemingly interdependent statements. This dialectic style was soon adopted in all European academies. Even the writings of Ravad (Abraham ben David), Zerahiah ha-Levi, and Yeshaya deTrani, three of the most original Talmudists (12th century), reflect the strong impact of Tosafist dialectic.

The works of Meir Abulafia and Menaḥem Meiri, although of the North African genre, include a strong dialectic element. In Spain such dialectic works were known as *ḥiddushim* or *novellae* (since they sought "new insights"), the most famous being those written by four generations (13th–14th centuries) of teacher and pupil: Ramban (Naḥmanides, or Moses ben Naḥman), Rashba (Solomon ben Adret), Ritba (Yomtov ben Abraham), and Ran (Nissim ben Reuben Gerondi).

A major role in establishing Talmudic authority was also played by the *responsa* literature, which gave replies (*responsa*) to legal and religious questions. Beginning in the 7th century, when the Babylonian *geonim* responded in writing to questions concerning the Talmud, it developed into a branch of Talmudic literature that continued in virtually the same form to the present day. Then, as now, Talmudic authorities were approached for explanations and decisions. Among the *geonim* the best known respondents were Sherira (10th century) and his son Hai. In the Middle Ages the most important were Alfasi, Ibn Migash (Joseph ibn Migash), Maimonides, Ravad (Abraham ben David of Posquières), Ramban, Rashba, Rosh (Asher ben Jehiel), Ran, and Ribash (Isaac ben Sheshet Perfet).

Writing and printing of the Talmuds. Study in the academies was always oral; hence the question of when the Mishna and Talmud were first committed to writing has been the subject of much discussion. According to some scholars, the process of writing began with Judah ha-Nasi. Others attribute it to the *savoraim*.

The Palestinian Talmud was first printed in Venice (1523–24). All later editions followed this one. Printing of the Babylonian Talmud was begun in Spain about 1482, and there have been more than 100 different editions since. The oldest extant full edition appeared in Venice (1520–23). This became the prototype for later printings, setting the type of page and pagination (a total of close to 5,500 folios). The standard edition was printed in Vilna beginning in 1886. It carries many commentaries and commentaries upon commentaries. In the sample page reproduced here, the Mishna and the Gemara are placed in the centre column of the page and are printed in heavy type. The commentary of Rashi is always located in the inner column of the page and the *tosafot* in the outer column. Other commentaries and references to legal codes and to scriptural verses surround the major commentaries, in smaller type. Talmudic citations are made by tractate name, folio number, and side of the folio (a or b).



Sample page (7a) of the tractate *Makkot* (of the fourth order, *Neziqin*) of the Vilna edition of the Babylonian Talmud, first printed in 1880–86. It discusses the fate of a man who was convicted and escaped and how he is to be judged. Code numbers, a box surrounding the Mishna, and brackets indicating the extent of comments (3-a) and (3-b) have been superimposed onto the original page. (1) End of the Gemara to the previous Mishna. (2) Mishna. (3) Gemara. (3-a) Halakhic Midrash supporting the Mishna. (3-b) Three short comments from Palestine, Sura, and Pumbedita. (4) Mark indicating the end of the chapter. (5) Mishna of the next chapter. (6) Commentary of Rashi (1040–1105). (7) *Tosafot*, discussing special points in the Gemara and Rashi. (8) Cross-reference notes to other Talmudic and Rabbinic sources and textual variants. (9) Notes by Joel Sirkes (1561–1640). (10) References to the codes of Maimonides, Moses of Concy, the *Tur*, and the *Shulhan 'arukh*. (11) Commentary of Hananel of North Africa, early 11th century. (12) References to scripture. (13) Notes by Elijah Gaon of Vilna (1720–97).

By courtesy of the Library of the Jewish Theological Seminary of America, New York; Frank J. Darmstaedter

NONLEGAL SUBJECT MATTER

God, Torah, and Israel

Main religious doctrines. While the Talmudic rabbis never formally systematized their beliefs, their underlying religious concepts are clearly reflected in their decisions, ideas, and attitudes. Pre-eminent in rabbinic thinking were the concepts of God, Torah, and Israel.

God. The rabbinic God was primarily the biblical God who acted in history, the creator and source of life who was experienced through the senses rather than intellect. In reaction to sectarian teachings (*i.e.*, Gnosticism and early Christianity), however, the rabbis stressed God's universality, absolute unity, and direct involvement with the world. At one and the same time his immanence and transcendence (being present in and beyond the universe) were emphasized, and biblical anthropomorphisms (ascribing human attributes to God) were explained metaphorically. The rabbis also stressed an intimacy into the relationship between God and man. God became the father to whom each individual could turn in direct prayer for his needs. To the names YHWH and Elohim, which traditionally were identified with God's mercy and judgment, respectively, the rabbis added new terms reflecting his other attributes—*e.g.*, Shekhina ("Presence"), representing his omnipresence, or immanence; and Maqom ("Place"), his transcendence.

Torah. Torah, in the Talmudic sense, refers to all religious and ethical teachings handed down by tradition. According to the rabbis, God created the Torah long

before the world. It served as his blueprint for creation and contained the eternal divine formula for the world's future workings and thus the answers to all problems for all times and all people. God himself is depicted as studying the Torah, for even he cannot make decisions concerning the world that contradict it.

Israel. The people Israel, according to the rabbis, were chosen by God to be the guardian of his Torah, and, just as God chose Israel, Israel chose God. Thus, the concept of Israel as a nation bound together by an irrevocable commitment to bring the Torah to the world, and bearing corporate responsibility for this mission, was formed. No Jew can free himself from this commitment, but anyone accepting it, regardless of race, becomes a full-fledged Jew with obligations binding him and his descendants.

With this in mind, the rabbis repeatedly emphasized the importance of studying Torah. They pointed out that the Torah is not a declaration of religious beliefs. Rather it is a statement of a discipline regulating each detail of life. Any transgression of this discipline hampers the divine plan of establishing God's way of life in this world.

Sanctification of everyday life

Worship. The intensive rabbinic religious involvement led to the growth of a new concept of worship. While in the Bible worship was usually centred in the sanctuary of the Temple in Jerusalem, the rabbis, particularly after the destruction of the Second Temple (CE 70), attempted to sanctify all of life. Thus, they said that one must bless God upon arising in the morning, before dressing, before and after meals, and in all ordinary daily actions or routines. Each move in life should be an act of worship glorifying God's name.

Messianic kingdom. In rabbinic thinking the establishment of God's kingdom was tied to the Messiah, who was to be a descendant of King David, wise, just, a great scholar, a moral leader, and courageous king. He would redeem the Jews from exile and re-establish their independence in the land of Israel. With this the world would be ushered into a new era of righteousness and universal peace. The rabbis referrred to this era as "the world to come," portraying it as an immense academy in which the righteous would study Torah without interruption. They refrained, however, from describing it further, saying that human language and fantasy are inadequate to depict its wonders.

The nature of the Messiah and the time of his arrival raised much speculation. Following the defeat of Bar Kokhba, leader of the revolt against Roman rule (CE 135), the Messiah's coming, in rabbinic thought, faded into the mysterious and distant future, and descriptions concerning his personality assumed supernatural overtones.

For a fuller discussion of major religious doctrines, see JUDAISM.

Doctrine of man. The fate of man, his achievements and failures, his being and nothingness, occupy an important place in Talmudic literature. The rabbis' concept of man was a universal one. While they assumed that Jews are bound by greater religious duties than others, they considered all men equal, all created in the image of God. "Therefore, but a single man was created . . . That none should say to his fellow, 'My father was greater than thy father' " (tractate *Sanhedrin*).

The world, according to the Talmud, was created for the sake of man, and it is incumbent upon him to keep it in order. His responsibility begins at home. Man must care for his health, marry, build a family, provide for and educate children, honour parents, friends, and elders. He also carries social responsibilities and has to be part of the community. He must learn a trade and work so that he does not become a burden to the community.

The uniqueness of man in this world, likened by the Talmud to the uniqueness of God in the universe, lies in his freedom of choice. Nature follows its laws and angels their missions, but man is his own master. In contrast to St. Paul's doctrine that the original sin of Adam made sin an integral part of human nature, the rabbis considered man a wondrous and harmonious being. The duality of his nature was explained by the existence of a good and bad impulse, personified by two angels, *yetzer ha-ṭov* (the good inclination) and *yetzer ha-raʿ* (the evil inclination), which enter each man after birth. It is the duty of man to overcome his evil inclination, and it is for this that he is rewarded. Moreover, since there is corporate responsibility, not only is the sinner punished but the community at large also suffers. Here again, however, man is his own master. He can reverse the course of sin and punishment by repentance. Although repentance may be accompanied by formal and ceremonial acts, such as fasting, its basic principle is the renunciation of the sin and the wholehearted decision not to repeat it. When a man transgresses against God, his sin is forgiven by repentance alone, but, when he transgresses against his fellow man, he must make good his wrongdoing as well as repent.

The *yetzer ha-ṭov* and *yetzer ha-raʿ*

Medicine and science. The Talmud devoted considerable attention to the maintenance of good health, regarding it a religious duty. A keen understanding of the importance of hygiene in preventing illness was reflected in an emphasis upon bodily cleanliness. The rabbis also stressed the necessity for moderation in eating and drinking and the importance of a proper diet. The Talmud prescribed remedies for illnesses and mentioned surgical techniques, such as cesarean section.

Religious concerns surrounding the calendar, prohibitions against planting seeds of different kinds together, dietary laws, and sabbath-walking limits resulted in an intense rabbinical interest in astronomy, zoology, mathematics, and geometry.

Legend and folklore. Side by side with the Midrashic Haggada, which was the outgrowth of Bible exegesis and developed in the academies, the Talmuds and Midrashic collections contain a large quantity of Haggadic material with mythological rudiments, allusions to pagan beliefs and customs, and folkloristic elements of a world strange to the rabbis. Folktales and legends, animal lore, and adventure narratives, containing pagan ideas and beliefs, that were told by their Gentile neighbours were no doubt a major attraction to the common Jews, especially those in the countryside (the *ʿam ha-aretz*, or "people of the land"). The rabbis realized the great danger involved in this situation and developed their own folk material. They adopted the dramatic and artistic parts of these stories but rejected the unwanted elements, replacing them with their own ideas. Thus the animals and birds in fables quote the Bible and discuss it in the same manner that the rabbis do.

The challenge of pagan myths, legend, and folklore

Ancient mythology seems to have been well known and liked by the Jewish masses. Again, in order to fight its influence, the rabbis reworked its content in their own spirit. They retained the mythological suspense—the sea tries to drown the earth—but there is no mythological struggle between equal powers; angels try to prevent the creation of man, but they do not possess titanic power. All are subdued by the command of God. Thus, the rabbis transformed the ancient myths into dramatic evidence against polytheism. (See also JEWISH MYTH AND LEGEND.)

Astrology, magic, and divination. Astrology was a recognized science in the ancient world. The rabbis could not reject it entirely, and some concluded that the power of the stars is confined to Gentiles. Others made it part of God's order, saying that stars influence this world in the same way that climate influences plants. The rabbis strenuously objected to omens and other forms of divination because they considered them magic. Dreams were considered by some rabbis as meaningless, while others saw in them an element of prophecy.

The rabbis believed in the efficacy of magic but strenuously objected to its practice. They permitted only magic that had been proved effective in healing. They also permitted the use of incantations for the purpose of counteracting the hold of magic. Because of their supposedly protective nature, the use of amulets was also countenanced.

Demons (evil spirits) and the evil eye

The existence of a demonic kingdom was accepted by the rabbis without question. Evil spirits are invisible and fill the nether world. They avoid sunlight and concentrate in waters and deserted places. They also mingle with people, trouble them, and help them. They have passions and are born and die like people. However, they also have some of the traits and powers of angels. The evil eye was considered as dangerous as evil spirits. It was thought that

for mysterious reason some people have the power to injure others by looking at them and that it is generally jealousy that triggers this effect. The rabbis, however, repeatedly emphasized that all of these strange powers are under the divine government and, moreover, that they cannot hurt the pious.

TALMUDIC LAW AND JURISPRUDENCE

Sacral basis of Talmudic law

Unlike the Romans, who considered ritual law (*fas*) God-given and social law (*lex*) man-made, the rabbis believed all Jewish law to be of divine origin. Thus, for example, unfairness in labour relations was considered a religious sin and caring for the sick a religious obligation. Though familiar with the concept of natural law (ethical principles inherent in the nature of things and apprehensible through human reason), the rabbis objected to making nature the basis of law. Even rabbinic ordinances were regarded as having validity only because the authority of the rabbis is sanctioned by the Torah.

Methods of arriving at legal principle and decisions. Ancient Halakha knew no controversy. The earliest controversy dates to the pre-tannaitic *zugot*. Hillel and Shammai differed on significant issues, and, with the rise of their schools, Halakhic uniformity began to crumble. Halakha became a scholastic discipline that developed in academic rather than judicial settings, more and more issues remaining unresolved. Over 300 controversies between the schools of Hillel and Shammai (called the House of Hillel and the House of Shammai, respectively) are reported in Talmudic sources. As time passed, disputes proliferated even more and were considered legitimate provided they conformed to the rule of Halakhic discipline.

No attempt was made to restore Halakhic uniformity until the beginning of the 2nd century CE. Controversies were sometimes resolved by citing old traditions, by establishing precedents, or, when the sages could convene, by vote taking.

At Yavne, Gamaliel II, the president of the revived Sanhedrin (*c.* CE 80–*c.* 115), attempted to suppress diversity of opinion, but failed. The right to differ was already established. Moreover, in the Halakhic collection compiled at Yavne (tractate *'Eduyyot*), the views of individual scholars were preserved. The sages at Yavne, however, did take a major step toward restoring Halakhic consistency by upholding the generally more lenient views of the House of Hillel over those of the House of Shammai, thus establishing the Hillelite tradition as the main trend of rabbinic Judaism.

The principle that differing opinions should be recorded was followed by Judah ha-Nasi in his Mishna. Modern scholars differ as to whether he meant to compile a code of law or merely a Halakhic collection. The *amoraim*, however, accepted his Mishna as the definitive code and introduced a set of guidelines according to which disputes were decided. Thus, for example, collective (" . . . the sages said") and individual opinions stated anonymously were taken as law; Akiba's decisions were upheld over those of his colleagues. Similar guidelines developed also with regard to amoraic controversies.

Guidelines for settling Halakhic controversies

With the completion of the Talmud, a new phase in Halakhic development began. Not only were there two different Talmuds and a large Haggadic literature but even within each of the Talmuds diversified opinions were reported. The *geonim* laid down rules governing the use of this enormous literature for lawmaking. They designated the Babylonian Talmud the highest authority, taking the Palestinian Talmud into consideration only when it did not disagree with the Babylonian or when the latter expressed no opinion on a subject. They also deprived the Haggadic literature of Halakhic authority and set guidelines for the precedence of opinion among *amoraim*. These geonic rules served as the basis of all future codifications.

After the geonic period two methods of decision making were applied. The first of these relied primarily upon the authoritative codes. The Mediterranean rabbis, for example, made the code of Maimonides the source of all of their lawmaking. The second method relied on the original Talmudic sources for decision making. This method was applied by the Tosafists and their followers, who, though they consulted the older codes, did not accept them as the final authorities. The *responsa* literature represents a synthesis of these two methods. Although it makes use of codes as the main source of law, its decisions are always accompanied by a discussion and analysis of earlier relevant literature. This approach has been used by rabbis to the present day.

In addition to the above, in particular instances throughout the ages rabbinic authorities promulgated ordinances (*taqqanot*) and edicts (*gezerot*). These were made in response to pressing needs of time and circumstance, and this form of lawmaking was most frequently used by rabbinic synods in the Middle Ages.

Administration of justice. *Courts.* A comprehensive judicial system is described in Talmudic sources. The highest court was the Great Sanhedrin. It consisted of 71 members and convened daily in one of the Temple halls. It was the highest legal and religious authority in the country and had exclusive jurisdiction over matters of a national and public nature. It also functioned as the court of appeals, dealing with cases that were not resolved by the lower courts.

The Great and Lesser Sanhedrins

Next in line of judicial authority was the Lesser Sanhedrin. Each town with a population of 120 or more had a court of this kind. These courts each consisted of 23 members and dealt with cases involving capital punishment.

The members of the Sanhedrins had to be ordained, pious, mature in age, sound in mind and body, of wide knowledge, and of pure Jewish descent. Persons who were too old or who had never had children were ineligible, for it was thought that they might not be merciful.

The lower courts dealt with all remaining cases. Each consisted of three members and convened on Mondays and Thursdays. In cases involving a penalty the three judges had to be ordained, but in those involving ordinary monetary ligitation ordination was not required. In the latter type of case, concerned parties were allowed the alternative of setting up ad hoc arbitration bodies.

Rules of evidence. Jewish law was extremely strict regarding evidence acceptable in court. In cases entailing physical punishment, no circumstantial evidence, confession, or self-incrimination was recognized. The testimony of two eyewitnesses who confronted the defendant was required. In monetary cases documentary evidence and, at times, oaths were acceptable. Any mental or moral defects or self-interest in the case disqualified witnesses. Relatives could not serve as judges or witnesses.

Trial procedure. Jewish law knows of no lawyers. After the facts were presented, the court investigated, deliberated, and made its decision by voting. Both sides had to be treated equally, even to the point of seeing to it that neither should be dressed more richly than the other. Each side could be heard only in the presence of the other.

In the trial procedure of capital cases, there was a clear tendency toward bias in favour of the defendant. Thus, only the judges could argue for conviction, but all present could argue for acquittal. The most junior judges voted first so that they would not be unduly influenced by their seniors. A majority of one was sufficient for acquittal, but a majority of two was necessary for conviction. A verdict of acquittal could be reached on the same day but one of conviction only on the following day. When the court erred, only its convictions, and not its acquittals, were reversed.

Criminal law. In Jewish law, ritual and nonritual transgressions were crimes punishable by court. Each of the 36 most severe transgressions (*e.g.*, adultery, sodomy, idolatry, sorcery, or murder) carried one of four types of death penalty (stoning, burning, beheading, and strangling). Rabbinic law, however, tended to minimize the practice of capital punishment. The rigorous cross-examination of witnesses and the warning of impending punishment that the transgressor had to receive immediately before committing his crime made it almost impossible to reach a death verdict.

Death penalties

If despite all of this a death verdict was reached, every legal effort was made to allow for a last-minute reversal. Execution was expedited and carried out in the most humane manner possible, the accused being given an opiate before dying. To show their compassion the judges fasted on the day of execution. According to tradition the death penalty was abolished 40 years before the destruction of the Temple, when the Great Sanhedrin was exiled from the Temple complex.

The punishment for 207 other transgressions (*e.g.*, perjury, some forms of incest, the eating of forbidden foods) was flagellation. Here, too, the rabbis tended to be lenient. As in capital cases, a rigorous cross-examination and a warning were required. The maximum number of stripes administered was 39. Prior to flagellation the transgressor was examined medically to determine the number of stripes he could withstand.

Side by side with the above penalties, the courts also inflicted *makkat mardut* (disciplinary stripes) and excommunication in cases where regular flagellation could not legally be applied. These two punishments were generally used in Babylonia, where ordained courts did not exist. It should be mentioned also that the Mishna includes a few obscure references to a form of imprisonment used instead of capital punishment.

Civil and social law. Although the rabbis considered both ritual and nonritual law sacred, they demonstrated great independence in supplementing the relatively brief relevant scriptural comments and regulations with a comprehensive system of civil and social law. In response to variations in social and economic circumstances, certain differences in Palestinian and Babylonian Talmudic law emerged. The Babylonian rabbis, for example, recognized the law of the state as binding in monetary matters, while the Palestinian rabbis did not. In general, however, Jewish civil law developed relatively autonomously. In instances where the rabbis did adopt alien legal concepts, they elaborated upon them until they could be fully integrated into the spirit and structure of Jewish law.

Social welfare, torts, family law, financial law

The following are some of the areas covered: (1) Social welfare: a comprehensive social welfare system was worked out, including obligations to provide for children, educate them, and train them for a profession. Regulations of charity, medical assistance, and burial of the dead were established. (2) Torts: included were all damages caused by a person directly or indirectly via his property. The main aim was to compensate for damages. Consequently, no torts were classified as criminal. Even "an eye for an eye" was interpreted to mean financial compensation. (3) Family law: included were regulations concerning marriage and divorce procedures and the innovation of the *ketubba* (marriage contract), which spells out the mutual obligations of husband and wife in the areas of finance, medical care, clothing, housework, sexual relations, and child care. According to biblical law, the right to inherit belongs to sons first. To protect the rights of wives and daughters, rabbinic law obligated the sons to maintain the widows and unmarried daughters. (4) Financial law: except for Gen. 23:9 ff., Jer. 32:10, and Ruth 4:8, Scripture makes no reference to transaction procedures. The growth of finances, industry, and land estates led the rabbis to develop laws concerning contracts, partnerships, and legal arrangements to circumvent the biblical prohibition against usury. A series of modes of transaction effecting the transfer and acquisition of property evolved. Labour relations, rents, and leases were also carefully regulated.

THE TALMUD TODAY

Role in the Jewish community. With the rebirth of a Jewish national state (since 1948) and the concomitant revival of Jewish culture, the Talmud has achieved renewed importance. Orthodox Jewry has always focussed upon its study and has believed it to be the absolute Halakhic authority. This belief has now become even further intensified. While rabbinic courts in Israel have jurisdiction only in the area of family life, it has become one of the aims of religious (Orthodox) Jewry there to establish Talmudic law as the general law of the state.

It should also be noted that, aside from the special case of Israel, the legal system described above has continued to function down to the present day in Jewish communities all over the world. The jurisdiction of rabbinic courts is voluntarily accepted by Orthodox Jews. These courts continue to exert authority, especially in the areas of family and dietary law, the synagogue, and the organization of charity and social activity.

Conservative Jewry, too, has always been committed to rabbinic tradition. It has, however, conceptualized this tradition as an evolutionary process in which Halakha changes to meet the challenge of new conditions. Professional scholarship was considered crucial for understanding the furthering of this process. More recently, however, as a result of revived nationalism, new emphasis has been put upon lay education. Thus, a network of day schools and higher institutions of learning in which rabbinic tradition occupies a major role in the curriculum has been established. Scores of young Conservative Jews now search in the Talmud for answers to crucial problems, such as abortion and civil violence.

Classical (19th-century) Reform Judaism not only disassociated itself from the Talmud but negated it. More recently, however, Reform leaders have been inclined to re-establish some measure of ritual practice and rabbinic climate. Thus, it is now not unusual to find them stating their decisions in the form of *responsa* and using the rabbinic style of argument and even the casuistic type of Talmudic dialectic (*pilpul*) to justify their religious practices.

Talmudic scholarship. Although Talmudic scholarship continues to be advanced by individuals in a number of countries, its two main centres are in Israel and the United States. The Israeli centre has tended to focus upon research of a critical nature. Like Bible criticism, this work is divided between source criticism (*i.e.*, discovering the different sources, their dates, and the methods by which Talmudic literature was formed) and textual criticism (*i.e.*, establishing the correct text and reading). Research is also being done on Haggadic concepts and thinking, Talmudic law, and Halakhic development.

Talmudic scholarship in the United States has tended to be more philosophically and historically oriented. There has been great interest in the development of Halakha and also in folklore and custom. Essential work has also been done and continues to be done in the areas of source criticism. A work unique in scope and method is S. Lieberman's commentary on the Tosefta.

BIBLIOGRAPHY. B. GERHARDSSON, *Memory and Manuscript* (1961), contains a description of the methods and techniques by which the oral tradition was transmitted. H.L. STRACK, *Introduction to the Talmud and Midrash* (1931); and M. MIELZINER, *Introduction to the Talmud*, 4th ed. (1968), are still the best introductions for the general reader. The latter is particularly helpful in explaining Talmudic dialectic terminology and debate. J. BOWKER, *The Targums and Rabbinic Literature* (1969); and E. DEUTSCH, *The Talmud* (1895), are both descriptions of the Talmud, the former concentrating upon Talmudic literary compilations and the latter upon Talmudic content. The introduction of J. GOLDIN, *The Living Talmud* (1957), contains a vivid description of Talmudic debate. C. ALBECK, *Introduction to the Talmud*, in Hebrew (1969); and Z.H. CHAJES, *The Student's Guide Through the Talmud* (Eng. trans. 1952), are more advanced introductions, the former analytical and scientific and the latter representing the traditional view. J. NEUSNER (ed.), *The Formation of the Babylonian Talmud* (1970), contains summaries of some research by modern scholars on the question of how the Talmud was formed. L. GINZBERG, "Introduction to the Talmud," in *A Commentary on the Palestinian Talmud*, vol. 1 (1941), is the only introduction to the Palestinian Talmud available in the English language. L. FINKELSTEIN, *Akiba* (1962), is a historical and sociological approach to the development of Halakha. S. LIEBERMAN, *Greek in Jewish Palestine* (1942) and *Hellenism in Jewish Palestine* (1950), deal with the interrelationship between the ancient rabbinic world and its Gentile environment. B. COHEN, *Jewish and Roman Law*, 2 vol. (1966); and I. HERZOG, *The Main Institutions of Jewish Law*, 2 vol. (1966–67), are the best English descriptions of Jewish law. J. Z. LAUTERBACH, *Rabbinic Essays* (1951); G.F. MOORE, *Judaism*, 3 vol. (1927–30); S. SCHECHTER, *Aspects of Rabbinic Theology* (1961); and E.E. URBACH,

The Sages: Their Concepts and Beliefs, in Hebrew (1969), cover the major aspects of rabbinic theology. I. HEINEMAN, *Paths of the Aggadah*, in Hebrew (1970); M. KADUSHIN, *The Rabbinic Mind*, 2nd ed. (1965); and D. BEN AMOS, *Narrative Forms of the Haggadah: Structural Analysis* (1969), discuss Haggadic methods, forms, concepts, and thinking. L. ZUNZ, *Die Gottesdienstlichen Vorträge der Juden* (1892; updated Hebrew translation, 1950), is a thorough historical survey of Haggadic literature. D. NOY, *Motif-Index of the Talmudic-Midrashic Literature* (1954); J.J. SLOTKI, *Index Volume to the Soncino Talmud* (1952); M. GASTER, *The Exempla of the Rabbis* (1924, rev. ed. 1968); and C.G. MONTEFIORE and H. LOEWE, *A Rabbinic Anthology* (1938), are very helpful as reference guides; while W.O.E. OESTERLEY, H. LOEWE, and E.I.J. ROSENTHAL, *Judaism and Christianity*, rev. ed. (1969); and C. MERCHAVIA, *The Church Versus Talmudic and Midrashic Literature*, in Hebrew (1970), describe the relationship between the church and rabbinic Judaism. L. FINKELSTEIN (ed.), *The Jews*, 3rd ed., 2 vol. (1960); and E.R. BEVAN and C. SINGER (eds.), *The Legacy of Israel* (1927), deal with the influence of Judaism on world culture. Extensive bibliographies may be found in the works of Gerhardsson, Mielziner, Bowker, and Ben Amos.

(H.Z.D.)

Tamil Nadu

One of the southern states of India, Tamil Nadu (Land of Tamil) is located in the extreme southeast of the Indian peninsula. Representing the Tamil-speaking area of what was formerly Madras Presidency, it has an area of 50,180 square miles (129,966 square kilometres) and a population of about 41,103,000. It is bounded by the Indian Ocean to the east and south, by the states of Kerala to the west, Mysore to the northwest, and Andhra Pradesh to the north. Its capital is Madras.

While it has an industrial core in Madras, the state is essentially agricultural. Proud of their Dravidian language and culture, the Tamils have resisted attempts to make Hindi the national language. (For information on the state capital, see the city article MADRAS.)

The three Tamil kingdoms

History. *Early history.* The history of Tamil Nadu begins with the establishment of a trinity of Tamil powers in the region—the Cēra, Cōḷa, and Pāṇḍya kingdoms. By about AD 200 the influence of northern Aryan powers had progressed, and the Aryan saint Agastya had established himself as a cultural hero. The use of Roman gold and lamps and the consumption of Italian wine attest to the extensive foreign trade of the period. From the mid-6th century until the 9th century, the Cālukyas of Bādāmi, the Pallavas of Kāñcī, and the Pāṇḍyas of Madurai fought a long series of wars in the region. The period, nonetheless, was marked by a revival of Hinduism and the advance of the fine arts. From about AD 850, Tamil Nadu was dominated by the Cōḷas, of whom Rājendra I (reigned 1014–44) was the most distinguished ruler. In the mid-14th century the Hindu kingdom of Vijayanagar, including Tamil Nadu, came into prominence. During the 300 years of Vijayanagar administration, Telugu-speaking governors and officials were established.

The British period and independence. In 1639 the British East Indian Company opened a trading post at the fishing village of Madraspatnam (now Madras) with the permission of the declining Vijayanagar Empire. The history of Tamil Nadu from the mid-17th century to 1946 is the story of the Madras Presidency in relationship to the rise and fall of British power in India. After 1946 the Madras Presidency was able to make steady progress, as it had a stable government. In 1953 the Telugu-speaking state of Andhra Pradesh was formed, and in 1956 the presidency was further divided into the states of Kerala, Mysore, and Tamil Nadu.

The landscape. *Relief, drainage, and soils.* The state is divided naturally between the flat country along the eastern coast and the hilly regions in the north and west. The broadest part of the eastern plains is the fertile Cauvery Delta; farther south are the arid plains of Rāmanāthapuram and Madurai. The high Western Ghāts range runs all along the state's western border. The lower hills of the Eastern Ghāts and its outliers—locally called the Javādis, Kalrāyans, and Shevaroys—run through the centre of the region.

The rivers

The important rivers include the Cauvery, the Ponnaiyār, the Pālār, the Vaigai, and the Tambraparni, all of which flow eastward from the inland hills. The Cauvery and its tributaries are Tamil Nadu's most important sources of water and power.

Apart from the rich alluvial soil of the river deltas, the predominant soils are clays, loams, sands, and red laterites (soils with a high content of iron oxides and aluminum hydroxide). Black cotton soil is found in parts of Salem and Coimbatore in the west, Rāmanāthapuram and Tirunelveli in the south, and Tiruchchirāppalli in the central region of Tamil Nadu.

Climate. The climate is essentially tropical. The temperature in summer seldom exceeds 109° F (43° C) and in winter seldom falls below 64° F (18° C). The lowest temperatures are recorded during December and January, and the highest in April to June. The average annual rainfall, falling mainly between October and December, depends on the southwest and northeast monsoons and ranges between 25 and 75 inches a year. The most precipitation falls in the Nīlgiris and other hill areas, the lowest in the Rāmanāthapuram and Tirunelveli districts.

Vegetation and animal life. Forests cover about 15 percent of the state. At the highest altitudes of the Western Ghāts—the Nīlgiri, Anamalai, and Palni Hills—the mountains support subalpine vegetation. Along the eastern side of the Western Ghāts and in the hills of the northern and central districts, the plant life is a mixture of evergreen and deciduous plants, some of which are markedly adapted to arid conditions. The state's aquatic birds are represented in the bird sanctuary at Vedantangal; other forms of wildlife may be seen in the game sanctuary at Mudumalai.

Population. *Ethnic groups.* The present population has changed little over the centuries. Population characteristics include short to medium stature, dark to brown skin colour, and wavy hair. Dark pigmentation is more general in the southern districts and among the lower castes and tribes. Most of the hill tribes are of Veddoid origin. Caste is still strong, though untouchability has been banned by law. Former untouchables, now officially termed "scheduled castes," number about 6,000,000; and hill tribes number about 250,000.

Linguistic patterns. Tamil, the official state language, is spoken by over 83 percent of the people. For a considerable number of the population long resident in the state, Tamil has almost become a mother tongue. Telugu is spoken by almost 10 percent of the population, Kannada by almost 3 percent, Urdu by less than 2 percent, and Malayalam by more than 1 percent. In the Nīlgiris district in the west, Kannada (and its dialect Badaga) and Malayalam are stronger. English is spoken as a subsidiary language.

Religious affiliations. The main religions are Hinduism, Christianity, Islām, and Jainism. Followers of the first three are found in all districts, but Jains are confined to North and South Arcot and Madras city. Hindus form almost 90 percent of the population, Christians 5 percent. The largest concentration of Christians is in the Tirunelveli and Kanyakumari districts. The growth of atheism is a recent development; in 1961 about 34,000 persons declared themselves atheists, possibly as a protest against Brahmin ritualism.

Patterns of settlement. Though Tamil Nadu is one of the two most urbanized states of India (27 percent of the population is urban), it is still a rural land. Almost 25,000,000 people live in more than 14,000 nucleated villages. The poorest low-caste villagers live in segregated areas called *cēri*. The urban population lives in almost 443 towns, 17 of which have more than 100,000 inhabitants. The Madras metropolitan conurbation, covering the industrial areas, townships, and villages, has a population of about 2,500,000. There are 16 other conurbations, of which those around Madurai, Coimbatore, and Tiruchchirāppalli are the most important.

Administration. *State government.* The governor, the Legislative Assembly (235 members), Legislative Council (63 members), and the chief minister and his cabinet together constitute the legislative and executive wings of

the state government. The ministries are housed in Ft. St. George in Madras, though the offices of several heads of departments are in multistoried buildings outside the fort.

Local government. The state has 14 administrative districts, each administered by a district collector. Lower administrative units are divisions called *taluka, firka,* and villages. All of these units are responsible to the Revenue Department and the Board of Revenue. After independence, new units—*pañcāyat*s (village councils)—were established for purposes of local self-government and rural development. Above the *pañcāyat* there are *pañcāyat* unions and development councils. Villagers get their first lessons in democracy and administration in the *pañcāyat*s.

Justice. The judiciary is controlled by the High Court at Madras; there are district judges and magistrates at the lower levels.

Social conditions. *Health and the cost of living.* In the early 1970s, the annual per capita income in Tamil Nadu was Rs. 564, or about U.S. $79. Distribution of wealth is uneven. Though the death rate of the same period, eight per 1,000 persons, was low, and the birth rate of 24 per 1,000 moderate, infant mortality was as high as 60 per 1,000 live births. Smallpox seemed to be under control, but cholera, malaria, and filariasis (disease caused by infestation of the blood and tissues by parasitic worms) were not. Urban sanitation and drainage were substandard. There were more than 1,500 public and private hospitals, dispensaries, and primary health centres in the state.

Education. The literacy rate was almost 40 percent in 1971. About 92 percent of the children between the ages of six and 14 were at school in 1968–69. There were over 31,500 primary and middle schools, more than 2,500 high schools, and almost 150 arts and science colleges, as well as medical colleges, engineering colleges, polytechnic institutes, and industrial training institutes. The three state universities are at Madras, Anamalai, and Madurai. A vigorous effort is being made to make Tamil instead of English the medium of instruction at the university level.

Welfare. In the early 1970s the lowest castes, called Harijan (Children of God), continued to be the most miserable section of the Hindu population. The Department of Harijan Welfare was in charge of programs to improve their housing, education, and economy. Government positions were reserved for them, and there were also reserved constituencies for them for elections to the state legislature and Parliament. The State Social Welfare Board and Department of Women's Welfare are concerned with the problems of the handicapped, widows, and children. The state government provides free lunches for poor children at the elementary schools.

The economy. *Agriculture.* Agriculture is the mainstay of life for about three-quarters of the rural population. From very early times Tamil farmers have skillfully conserved scarce rainwater in small and large irrigation reservoirs, or "tanks." Government canals, tube wells, and ordinary wells also form part of the irrigation system. There are 17 river valley irrigation schemes, with 21 reservoirs and about 14,000 miles of distribution canals. As several of the river valley projects depend for water on the erratic northeast monsoon, the government is trying to tap subsoil water sources.

The irrigation system

Agricultural practices have undergone radical improvement since 1950. Multiple cropping, the use of new and better strains of rice, cotton, sugar, and millet, and the use of chemical fertilizers have been widely adopted. By 1967 the state was self-sufficient in the production of food grains. Tamil Nadu accounts for about 11 percent of India's rice production and 15 percent of its peanuts. Cotton and sugarcane are other important cash crops.

Industry. Improved port facilities and the effective use of electric power resources have helped industrial development. The state is the third most industrialized of the Indian states. In 1971 there were four thermal and 15 hydroelectric power stations with a capacity of almost 2,000,000 kilowatts. These were to be further augmented upon completion of an atomic power plant at Kalpakkam (scheduled for 1974). Cotton ginning, spinning, and weaving continued to be the major industries, followed by the manufacturing of cars, motorcycles, electric motors, and transformers. Other major undertakings include refineries and fertilizer plants and tire, aluminum, cement, chemical, and paper factories.

Transport and communications. The transport system of the southern Indian states converges on Madras. About 235 miles of broad-gauge or metre-gauge railways run through Tamil Nadu. The artificial harbour of Madras handles seaborne traffic. Apart from the international airport at Meenambakkam near Madras, there are three other airports at Tiruchchirāppalli, Madurai, and Coimbatore. There are 2,209 miles of motorable roads. Passenger bus transport is being nationalized; express buses take passengers to all important towns and places of interest. Post and telegraph offices are widely distributed and telephone services are available. The two radiobroadcasting stations are at Madras and Tiruchchirāppalli.

Cultural life. *The cultural milieu.* Despite attempts to break the religious and poltical power of the priestly Brahmin caste, Hinduism remains at the core of the culture. The more than 9,300 larger temples come under the administrative control of the Hindu Religious and Charitable Endowments Department. In most towns—and particularly in Chidambaram, Kānchipuram, Thanjāvūr (Tanjore), and Madurai—the *gopuram*s, or gateway towers, of the temples are dominant. The cycle of temple festivals attracts large congregations of devotees. Noteworthy are the car festivals during which large chariots decorated with icons are dragged around the temple. Hindu families owe allegiance to a number of sectarian monastic institutions, or *mutt*s, of which the most important is the Sankara Mutt at Kumbakonam.

The cultural role of Hinduism

The arts. Classical *bhārata-nātya* dancing and Carnatic music are widely practiced. Painting and sculpture, however, are less developed although there are schools for stone and bronze sculptors. Tamil literature rapidly adapted to the Western literary forms of the novel and the short story. The poet Subrahmanya Bharati (who died in 1922) was one of the first to modify traditional Tamil poetry. Since the 1940s, the motion picture has become the most popular form of mass entertainment. There are both touring and permanent cinema theatres; and sentimental and spectacular films, often featuring light music and dancing, are produced by the state's film studios.

The press. In 1968 there were over 400 periodicals published in Tamil, of which 41 were daily newspapers. The *Dinathanthi* is the leading paper. Among English newspapers, *The Hindu* of Madras maintains excellent journalistic standards.

BIBLIOGRAPHY. K.A. NILAKANTA SASTRI, *The Culture and History of the Tamils* (1964) and *A History of South India from Prehistoric Times to the Fall of Vijayanagar,* 3rd ed. (1966), are eminently readable and factual texts. Important data and an excellent atlas of the state may be found in the *Census of India 1961,* vol. 9, *Madras,* pt. 1–8 (1965–67). The proceedings of the 21st International Geographical Congress, Madras, *This Is Tamil Nadu* (1968), contains a collection of papers on a wide range of topics.

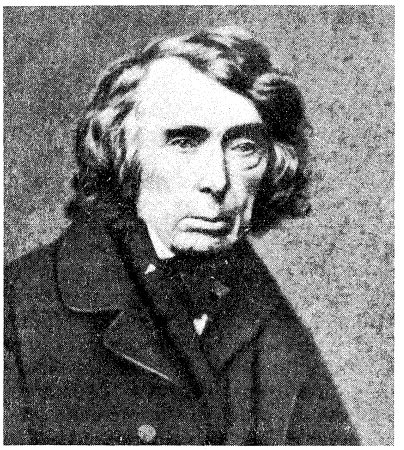
(A.Ai.)

Taney, Roger Brooke

Roger Brooke Taney, fifth chief justice of the Supreme Court of the United States, is principally remembered for the decision in the Dred Scott case, with which the court's majority opinion set back the cause of the antislavery forces by many years. Yet Taney was more than merely a rigid supporter of states' rights; in many important cases he advanced, in strict accordance with the Constitution, the power of the federal government.

He was born March 17, 1777, the son of Michael and Monica (Brooke) Taney. Of English ancestry, Michael Taney had been educated in France and was a prosperous tobacco grower in Calvert County, Maryland.

After graduation from Dickinson College in Pennsylvania, in 1795, Taney studied law with Judge Jeremiah Chase, of the Maryland General Court. He was admitted



Taney, photograph by Mathew Brady (c. 1823–96)
By courtesy of the Library of Congress, Washington, D.C.

to the bar in 1799 at Annapolis and served one year in the Maryland House of Delegates before settling down in Frederick, Maryland, to practice law. In 1806 he married Anne Key, whose brother, Francis Scott Key, later wrote "The Star-Spangled Banner."

Taney was a member of the conservative, property-conscious Federalist Party until 1812, when the party opposed the war against England. He returned to the Maryland House of Delegates in 1816, when, as a political maverick, he was elected to the state senate. Two years after his term expired in 1821, he moved his family to Baltimore, where he was soon recognized as an excellent lawyer. Juries were impressed with his sense of fair play and his courtesy toward opposing attorneys. In 1827 he was appointed attorney general of Maryland. By this time he had aligned himself with Andrew Jackson, the leader of the Democratic Party, and when Jackson, elected president in 1828, reorganized his cabinet in 1831, he appointed Taney attorney general of the United States.

Fight against the Bank of the United States

Throughout his tenure in Washington, Taney had been an outspoken leader in the Democrats' fight against the central bank, the Bank of the United States, which was widely regarded as a tool of Eastern financial interests. Taney believed it had abused its powers, and he strongly advised the president to veto the congressional bill that would renew the bank's charter and wrote much of the veto message; he also recommended that government funds be withdrawn from the bank and be deposited in a number of state banks.

As a result of his role in the fight over the Bank of the United States, Taney had become a national figure and in 1833 President Jackson appointed him secretary of the treasury. But opposition to Taney and his financial program was so strong that the Senate rejected him in June 1834, marking the first time that Congress had refused to confirm a presidential nominee for a cabinet post.

Taney returned to Baltimore to rebuild his law practice. A year later, Jackson nominated him to the Supreme Court of the United States as an associate justice. Taney's enemies stalled the nomination indefinitely. Then, on July 6, 1835, Chief Justice John Marshall died, and Taney was nominated to fill his place on the bench.

Despite powerful resistance, led by such prominent politicians as Henry Clay, John C. Calhoun, and Daniel Webster, Taney was sworn in as chief justice in March 1836. Although he had inherited the conservative tradition of the Southern aristocracy and had supported states' rights, the Taney court did not discard John Marshall's ideas of federal supremacy. Taney believed firmly in divided sovereignty, but he also believed it was the Supreme Court's role to decide which powers should be shared. Eventually, many of those who had opposed Taney's appointment came to respect him.

One of the most important decisions for which the Taney court is noted concerned rights granted by charters. The majority opinion in *Charles River Bridge* v. *Warren Bridge* (1837) declared that rights not specifically conferred could not be inferred from the language of a document. In this decision Taney rejected the claim of a bridge company that the subsequent grant by the state legislature of a charter to another bridge company impaired the legislature's charter to the first company.

The Dred Scott case

But the majority opinion that Taney delivered 20 years later, on March 6, 1857, in *Dred Scott* v. *Sanford* is the one for which he became famous. In essence, the decision argued that Scott was a slave and as such was not a citizen and could not sue in a federal court. Taney's further opinion that Congress had no power to exclude slavery from the territories and that Negroes could not become citizens was bitterly attacked in the Northern press. The Dred Scott decision probably created more disagreement than any other legal opinion in U.S. history; it became a violently divisive issue in national politics and dangerously undermined the prestige of the Supreme Court.

Whenever state authorities threatened or interfered with the execution of federal power, however, Taney upheld federal supremacy. His opinion in *Ableman* v. *Booth* (1858), denying state power (in this case the courts of the state of Wisconsin) to obstruct the processes of the federal courts, remains a magnificent statement of constitutional federalism. Under Taney's leadership federal judicial power was expanded over corporations, the federal government was held to have paramount and exclusive authority over foreign relations, and congressional authority over U.S. property and territory was vigorously upheld. His conflict with President Lincoln over the president's suspension of a citizen's petition for a writ of habeas corpus in time of war made him an object of bitter criticism, although in time many jurists came to agree with Taney's defense of an individual's constitutional rights.

Taney, a deeply religious Roman Catholic, considered slavery an evil. He had freed the slaves he had inherited before he came to the Supreme Court. It was his belief, however, that slavery was a problem to be resolved gradually and chiefly by the states in which it existed. As an adherent of the South, he could do nothing but watch the defeat of his cause. He died in Washington on October 12, 1864. Even though his thinking ran counter to the dominant historical trends of his time, he had an enduring influence on the substance and evolution of American constitutional law.

BIBLIOGRAPHY. C.B. SWISHER, *Roger B. Taney* (1935), a definitive biography; FELIX FRANKFURTER, *The Commerce Clause Under Marshall, Taney and Waite* (1937); A.J. SCHUMACHER, *Thunder on Capitol Hill: The Life of Chief Justice Roger B. Taney* (1964), a juvenile biography with emphasis on Taney as a political figure; SAMUEL TYLER, *Memoir of Roger Brooke Taney, LL.D.*, 2nd rev. ed. (1876), contains Taney's autobiography and an appendix of legal opinions, including that of the Dred Scott case; A. DUNHAM and P.B. KURLAND (eds.), *Mr. Justice*, rev. ed. (1964), and K.B. UMBREIT, *Our Eleven Chief Justices: A History of the Supreme Court in Terms of Their Personalities*, vol. 1 (1938, reprinted 1969), collections of biographical essays with analyses of the contributions of chief justices to the development of the Supreme Court.

(A.J.Sc.)

T'ang Hsüan Tsung

Under Hsüan Tsung (in Pin-yin romanization Tang Xüan Zung), the sixth emperor of the T'ang dynasty, the T'ang reached the zenith of their power and the peak of their culture. Hsüan Tsung ascended the throne in 712 under the nominal regency of his father, Jui Tsung; he achieved full power in 713 and abdicated in 756.

Early life and career

Hsüan Tsung, whose personal name was Li Lung-chi, was the third son of Jui Tsung by his empress *née* Tou. He was born in Lo-yang in 685, during a period when actual power was entirely in the hands of Empress Wu, although his father was nominal emperor. Hsüan Tsung survived purges in the 680s and early 690s, living with his father in strict seclusion, while many of the T'ang Imperial family were killed. Enfeoffed as prince of Ch'u in 687, he was re-enfeoffed as prince of Lin-tz'u in 693 after Empress Wu's usurpation of the throne. Toward the end

of her reign he was appointed an officer in the Imperial Guard of the Right and to an office in the Imperial Household department. On the restoration to the throne in 705 of his uncle Chung Tsung, he became vice president of the Court of Imperial Insignia, which controlled the arsenals of the capital. Although these posts were probably ceremonial, they gave him influence over the Imperial guards and palace armies.

In 710 Chung Tsung was poisoned by his domineering empress Wei, who appointed a 15-year-old puppet to the throne, intending to rule herself. This plan was thwarted by a coup led by the princess T'ai-p'ing, daughter of Empress Wu and sister of Jui Tsung. The empress Wei was killed, the puppet emperor deposed, and Jui Tsung restored to the throne. As a result of Hsüan Tsung's key role in this coup, he was appointed heir apparent.

Gradually, the court divided into factions around the princess T'ai-p'ing, who wished to continue a system of corrupt patronage, and Hsüan Tsung, who had instituted various administrative reforms. In 712 Jui Tsung abdicated in favour of Hsüan Tsung but, at T'ai-p'ing's urging, remained "Supreme Emperor," a sort of regent with control over appointments to high offices, which were filled with the Princess' supporters. The struggle between Hsüan Tsung and his aunt was intensified. In 713 she planned yet another coup, but this was forestalled by Hsüan Tsung, and she committed suicide. Hsüan Tsung then assumed full authority as emperor, and his father retired into seclusion.

Assumption of full imperial authority

Hsüan Tsung's reign began well. He carried out a sweeping reform of the bureaucracy, which had become vastly inflated by great numbers of nominal and supernumerary officials, many of whom had been appointed by patronage and many more by the open purchase of their posts. Some stability was regained, and the normal channels of authority, which had fallen into disorder under Empress Wu, were restored. The purchase of office was restricted. The numbers of Buddhist clergy and monasteries were reduced, and irregular ordinations, carried out to evade tax and civil responsibility, were severely restricted. Some financial reforms were effected, and the canal system, upon which the capital at Ch'ang-an relied and which had fallen into decay while Empress Wu resided in Lo-yang, was restored to action. Successful campaigns were waged against the Tibetans, the Turks, and the Khitan. During this early stage of Hsüan Tsung's reign, which lasted until about 721, the court continued to be dominated by the examination-recruited ministers who had served the empress Wu, Chung Tsung, and Jui Tsung. The authority of the throne, the efficient functioning of the regular bureaucracy, and the finances of the state were largely restored. Moreover, the members of the Imperial clan were gradually excluded from any effective political role, while the power of the palace officials and members of consort families was strictly limited to avoid any further palace coups.

A period of wide-ranging reforms in administration began in 720. The whole structure of central government was changed in such a way as to concentrate more and more authority in the hands of the chief ministers. At the same time there was a marked resurgence of the influence of the old aristocracy at court, and the period 721–737 was one of continuous political tension between the aristocrats and the examination-recruited professional bureaucrats. The aristocratic faction had its chief centres of power in the censorate, which was in general control of the functioning of the bureaucracy, and in the various new irregularly established commissions that implemented a series of sweeping financial reforms. The population was effectively reregistered, bringing vast numbers of taxpayers onto the rolls and a sharp increase in revenue; the coinage was improved, and the transportation system was reformed so effectively that the Emperor no longer had to move the court between Ch'ang-an and Lo-yang periodically to avoid famine. The empire's revenues increased, enabling the Emperor to establish along the northern frontiers an ever-growing permanent military establishment (by the end of his reign numbering 600,000 men), without overburdening the population.

The political influence of these financial experts grew even greater in the latter part of Hsüan Tsung's reign, and after 737, Li Lin-fu, chief representative of the aristocratic interest, became virtual dictator and the aristocratic party was firmly entrenched in power. From about 740 onward the Emperor's actual control of affairs began to decline. The reforms, which theretofore had mostly been necessary for greater administrative efficiency, now tended more and more to destroy the balance of political power. The chief ministers formally acquired unprecedented power and prestige as heads of the government. At the same time the Emperor began to erect around him a complex palace administration, including two private secretariats (used to bypass the normal channels for drafting legislation) and palace treasuries handling the Emperor's private purse. The financial experts, too, devoted more and more of their attention to purely exploitative measures designed to pay for court extravagance and the Emperor's personal needs.

Decline of Hsüan Tsung's control

Moreover, after 737, the vast regional commands established earlier in the reign to control the northern border and to provide a necessary network of strategic command now began to develop widespread powers in other fields and to acquire territorial authority. By the late 740s some of these generals had grown immensely powerful and began to intervene in court politics. Most important of them was Li Lin-fu's protégé An Lu-shan, who controlled the northeast and had an army of 180,000 troops. The central government had no standing armies under its own command to rival the forces of these military governors.

Meanwhile Hsüan Tsung had withdrawn more and more. Always a great patron of the arts—he had founded Imperial music academies to provide court musicians and had patronized poets, painters, and writers—he now became deeply involved in the study of Taoism, from whose founder the T'ang royal house claimed to be descended.

He also began to suffer from family problems. Hsüan Tsung had produced 59 children, of whom 48 survived to maturity. Many of these had been married into the great aristocratic families, thus helping to cement the authority of the latter. The Emperor kept very strict control over his children and their innumerable descendants, but he nevertheless came under the powerful influence of at least two of his many consorts. The first was Wu Hui-fei, who had great influence from the early 720s until her death in 737; she was influential in the rise of Li Lin-fu, and shortly before her death she was involved in an incident that resulted in the death of the crown prince and three other princes who stood in the way of her own eldest son's becoming heir to the throne. The eventual heir, however, was another prince (the future Su Tsung), who was opposed to Li Lin-fu. From 746 onward there were several alleged succession plots, which led to purges involving many prominent ministers and generals.

The Emperor also came under the influence of another favourite, the consort Yang Kuei-fei, who was originally the wife of his own son Li Mao but who had been taken into the Imperial harem. During the later years of his reign, Hsüan Tsung was completely infatuated with her and heaped honours on members of her family. One of these relatives, her cousin Yang Kuo-chung, rose rapidly to rival even Li Lin-fu in power and, on the latter's death in 752, replaced him as the dominant chief minister. There had already been some tensions between Yang Kuo-chung and An Lu-shan. With the removal of his patron at court, and the increasing hostility of Yang Kuo-chung, An Lu-shan began building up his provincial power base in readiness for an armed confrontation. This began at the end of 755. An Lu-shan's forces swiftly struck into the northeastern provinces, and, by the summer of 756, they were approaching Ch'ang-an. Hsüan Tsung, accompanied only by a small group of intimate courtiers, Yang Kuei-fei, and a few troops, fled to take refuge in Szechwan, the power base of the Yang clan. They had reached Ma-wei when the soldiers mutinied, killed Yang Kuo-chung, and forced Hsüan Tsung to have Yang Kuei-fei killed.

Shortly afterward, the heir apparent, who had fled separately to Ling-wu, west of the capital, proclaimed himself

Abdication

emperor. Hsüan Tsung, who heard of this only some time after it had occurred, acquiesced and abdicated formally in his favour. He lived in retirement until his death in 762.

Although Hsüan Tsung's reign ended in political disaster and personal tragedy, it was a period of internal stability, good government, and prosperity, an era of confidence during which real progress was made in every field. The sudden end of this period not only changed the political system completely but it was also a dramatic, traumatic experience for the men of the time. In the next decades the confident pride of Hsüan Tsung's age was replaced by self-questioning, by withdrawal from public affairs, and by a new spirit of social and political criticism.

BIBLIOGRAPHY. The only satisfactory account of this period in a Western language is E.G. PULLEYBLANK, *The Background of the Rebellion of An Lu-shan* (1955), which deals primarily with the political events of the latter part of the reign. There is no standard biography in any language. OTTO FRANKE, *Geschichte des chinesischen Reiches*, vol. 2 (1936), gives an account based upon the standard Chinese official histories. HOWARD S. LEVY has discussed Hsüan Tsung's succession to the throne in "How a Prince Became Emperor: The Accession of Hsüan-tsung," *Sinologica*, 6:101–121 (1961). The financial and administrative reforms of the period are dealt with in D.C. TWITCHETT, *Financial Administration Under the T'ang Dynasty* (1963).

(D.C.T.)

T'ang T'ai Tsung

The second emperor and real founder of the T'ang dynasty, which ruled China from AD 618 to 907, was Li Shih-min, whose posthumous title was T'ang T'ai Tsung (Pin-yin romanization Tang Tai Zong). The new dynasty proved long and enduring and became the model for its successors down to modern times. This continuity was largely the work of T'ai Tsung, who was as competent as an administrator and ruler as he was triumphant as a general. His achievement was to re-found the united Chinese Empire on a basis that endured under successive dynasties for more than 1,000 years.

The second son of Li Yüan, a military governor under the Sui dynasty, T'ai Tsung was born in 600 a member of an aristocratic family of Chinese and Tatar descent. As early as 618 he instigated the revolt of his father against the failing Sui dynasty and then led the armies of the new T'ang regime to complete victory over 11 rivals. In 626 he succeeded his father as emperor and reigned until his death in 649. China had been divided into two rival empires, north and south, Tatar and Chinese, for 272 years until reunited in 590 by the short-lived Sui dynasty. The new unity collapsed under the second Sui emperor, Yang Ti; and it was in the ensuing anarchy that the future T'ai Tsung, in a series of brilliant campaigns, established the T'ang dynasty and restored the unity of the Chinese empire.

Establishment of the T'ang dynasty

The restoration of unity was a considerable feat of arms. At the fall of the Sui dynasty almost all the provinces set up separate regimes that were in conflict with each other and that aimed, individually, to win supreme mastery of the empire. The south seceded under a claimant of a former southern dynasty. The Turks, meanwhile, then the dominant tribe among the nomads beyond the Great Wall (extending from the Yellow Sea roughly west to Central Asia), raided China, set up their own puppet emperors in border regions, and planned more massive invasions.

In the six years between 618 and 624, T'ai Tsung defeated all rivals and repelled the Turks. The south was reconquered; many of the contesting leaders fought to the death, though some joined the T'ang and obtained rank and title. The unity of the empire was popular with the mass of the people, for it alone ensured peace and absence of internal strife. But T'ai Tsung was less skillful at court intrigue. He provoked the jealousy of his less distinguished elder brother, the crown prince Chien-ch'eng, who intrigued to have T'ai Tsung dismissed from his command and plotted to assassinate him. T'ai Tsung, supported by his generals and advisers, forestalled the attempt and ambushed his brother and followers at the palace gate, shooting them down with arrows. His father abdicated, but this violent solution to a political problem was copied in later generations, proving detrimental to governmental stability. T'ai Tsung also failed to find a suitable son as successor, being bound by custom and political pressure to choose one born to the senior empress, not to any concubine. His first choice, the crown prince Ch'eng-ch'ien, proved to have an unstable character; he plotted against his father and had to be demoted. The second choice, Prince Chih, was docile, weak, and incapable, becoming after T'ai Tsung's death, as Emperor Kao Tsung, the instrument of his ambitious and capable empress, Wu Tse T'ien.



T'ang T'ai Tsung, portrait by an unknown artist. In the National Palace Museum, Taipei.
By courtesy of the National Palace Museum, Taipei

Domestic policies

As a ruler T'ai Tsung was eminently successful. His authority was never disputed, and his reforms reinvigorated and transformed old institutions. The system of public examinations for the civil service was among the most important of his innovations; in time, it transformed the Chinese governing class from a dominant aristocracy to a bureaucracy recruited from a much wider field of literate gentry. This was indeed T'ai Tsung's purpose; to end the contests of aristocratic families, such as his own, which had made the period of division between north and south one of incessant dynastic instability. The army was converted from levies raised by great aristocrats into a professional army, stationed on the frontiers and commanded by officers who had no eminent social background. Hence it was a stabilizing influence in the earlier T'ang period and remained the model system for later ages. T'ai Tsung also promoted education, setting up a state college, which might almost be called a university, to train future scholars for the civil service. His thorough administration included reorganization of provincial government, finance, taxation, and land tenure. The government of the T'ang empire was then the most developed and most sophisticated system anywhere in the world.

Posterity, with the exception of Communist historians, has hailed T'ai Tsung as the greatest emperor to occupy the Chinese throne. His career was certainly a most formative influence on later Chinese history and civilization. An accomplished patron of learning, he personally edited

the history of the division between north and south, and his calligraphy, an art much esteemed in China, has been recorded on stone inscriptions and used as a model in schools for more than 1,000 years. The middle T'ang period, the 8th century, is regarded as one of the ages of highest cultural achievement in Chinese history, and the conditions for this development were created by the political stability and reforms established under T'ang T'ai Tsung.

BIBLIOGRAPHY. C.P. FITZGERALD, *Son of Heaven: A Biography of Li Shih-min, Founder of the T'ang Dynasty* (1933).

(C.P.F.)

Tanks and Armoured Vehicles

Tanks are the principal type of tracked military armoured vehicle. Essentially mobile platforms for direct-fire weapons, they make the weapons mounted in them more effective by virtue of their cross-country mobility and the protection that they provide for their crews. Tank weapons have ranged from a single machine gun of small calibre to present-day combinations of one or more machine guns with a large gun of 75- to 152-millimetre calibre. A few experimental tanks have been armed with even larger guns and some tanks are now supplied with guided missiles.

Other major tracked armoured vehicles include: personnel carriers, originally conceived only for transporting infantrymen into battle, but since also used for fighting, although they are less powerfully armed than tanks; self-propelled guns, which are sometimes unarmoured and wheeled; armoured cars, which often resemble the lighter reconnaissance types of tanks but run on wheels; and other specialized vehicles, including armoured recovery vehicles, bridgelayers, and armoured bulldozers, all of which are principally intended to assist tanks, and amphibious armoured vehicles that are used for assault landing operations.

HISTORY OF ARMOURED VEHICLES IN WARFARE

Earliest developments. Ideas for armoured fighting vehicles may be traced as far back as the 2nd millennium BC, when horse-drawn war chariots were used in the Middle East by the Egyptians, Hittites, and others as mobile platforms for fighting with bows and arrows. Other concepts, including protection of armoured vehicles, can be traced back through wheeled siege towers and battering rams in the Middle Ages to similar devices used by the Assyrians in the 9th century BC. More recent forms of the protected vehicle idea may be found in battle cars proposed in 1335 by Guido da Vigevano, by Leonardo da Vinci in 1484, and by others down to J. Cowan, who secured a patent in England in 1855 for a wheeled, armoured vehicle based on the steam tractor.

It was only at the beginning of this century, however, that armoured vehicles began to take practical form. By then, the essential mechanical elements had been developed. In particular, the traction engine and the motor car became available for military application. The first self-propelled armoured vehicle was produced in 1900 in England when John Fowler & Company armoured one of their steam traction engines for hauling supplies in the South African War of 1899–1902. The first use of the motor car as a gun carrier was in 1899, in England, when a Maxim machine gun was mounted on a powered four-wheeled vehicle with a bicycle seat and handle bar steering (quadricycle). The inevitable next step was a vehicle that was both armed and armoured. Such a vehicle was constructed and exhibited in London in 1902. In the latter part of 1899 a machine-gun-armed motor car was built in the United States and in 1902 a partially armoured machine-gun car was exhibited in Paris by the Société Charron, Girardot et Voigt, which two years later developed a fully armoured car with a turret. By 1904 a turreted armoured car was also built in Austria by the Austro-Daimler Company.

First self-propelled armoured vehicle

To complete the evolution of the basic elements of the modern armoured vehicle, it remained only to adopt tracks as an alternative to wheels and this became inevitable with the appearance of the tracked agricultural tractor. There was no incentive for this development, however, until after the outbreak of World War I.

Proposals for tracked armoured vehicles failed to arouse the interest of military authorities. Suggestions for an armoured tracked gun made in 1903 in France and five years later in Britain were both denied consideration. Similarly, a design for a tracked armoured vehicle put forward in 1911 was rejected, in turn, by both the Austro-Hungarian and German general staffs, and in 1912 the British War Office turned down still another design.

World War I. The outbreak of World War I in 1914 radically changed the situation. Its opening stages of mobile warfare accelerated the development of armoured cars, numbers of which were quickly improvised in Belgium, France, and Britain. Trench warfare, which ended the usefulness of armoured cars, brought forth new proposals for tracked armoured vehicles. Most of these resulted from attempts to make armoured cars capable of moving off the roads, over broken ground, and through barbed wire. The first tracked armoured vehicle was improvised in July 1915, in Britain, by mounting an armoured car body on a Killen-Strait tractor. The vehicle was constructed by the Armoured Car Division of the Royal Naval Air Service, whose ideas, backed by the First Lord of the Admiralty, Winston S. Churchill, resulted in the formation of an Admiralty Landships Committee. A series of experiments by this committee led in September 1915 to the construction of the first tank, called "Little Willie." A second model, or "Big Willie," quickly followed; it was designed to cross wide trenches. "Big Willie" was accepted by the British Army, which ordered 100 tanks of this type in February 1916.

First tracked armoured vehicle

Simultaneously but independently, tanks were also being developed in France. Like the very first British tank, the first French tank (the Schneider) amounted to an armoured box on a tractor chassis; 400 were ordered in February 1916. But French tanks were not used until April 1917, whereas British tanks were first sent into action on September 15, 1916. Only 49 were available and their success was limited. On November 20, 1917, however, 474 British tanks were concentrated at Cambrai and achieved a spectacular breakthrough. These tanks, however, were too slow (four miles per hour) and had too short a range (20–40 miles) to exploit the breakthrough. In consequence, demand grew for a lighter, faster type of tank and in 1918 the 14-ton Medium A appeared with a speed of eight miles per hour and a range of 80 miles. After 1918, however, the most widely used tank was the French Renault F.T., a light and slow 6-ton vehicle designed for close infantry support (see Figure 1, right).

When World War I ended in 1918, France had produced 3,870 tanks, and Britain 2,636. Most French tanks survived into the postwar period; these were the Renault F.T., much more serviceable than their heavier British counterparts. Moreover, the Renault F.T. fitted well with traditional ideas about the primacy of the infantry; the French Army adopted the doctrine that tanks are a mere auxiliary to infantry. France's lead was followed in most other countries; the U.S. and Italy both assigned tanks to infantry support and copied the Renault F.T. The U.S. copy was the M1917 light tank, and the Italian the Fiat 3000. The only other country to produce tanks by the end of the war was Germany, which built about 20.

Interwar developments. The Renault F.T. remained the most numerous tank in the world into the early '30s. Aware of the need for more powerful vehicles, if only for leading infantry assaults, the French Army took the lead in developing well-armed tanks. The original 1918 French Schneider and St. Chamond tanks already had 75-millimetre guns while the heavier British tanks were at best armed with 57-mm guns. After the war the French built ten 68-ton 2C tanks with the first turret-mounted 75-mm guns and continued to develop 75-mm-gun tanks, notably the 30-ton Char *B* of 1936.

In the meantime, Britain took the lead, technically and tactically, in developing the mobility of tanks. Even before World War I had ended, work had started on the Medium D with a maximum speed of 20 mph. The

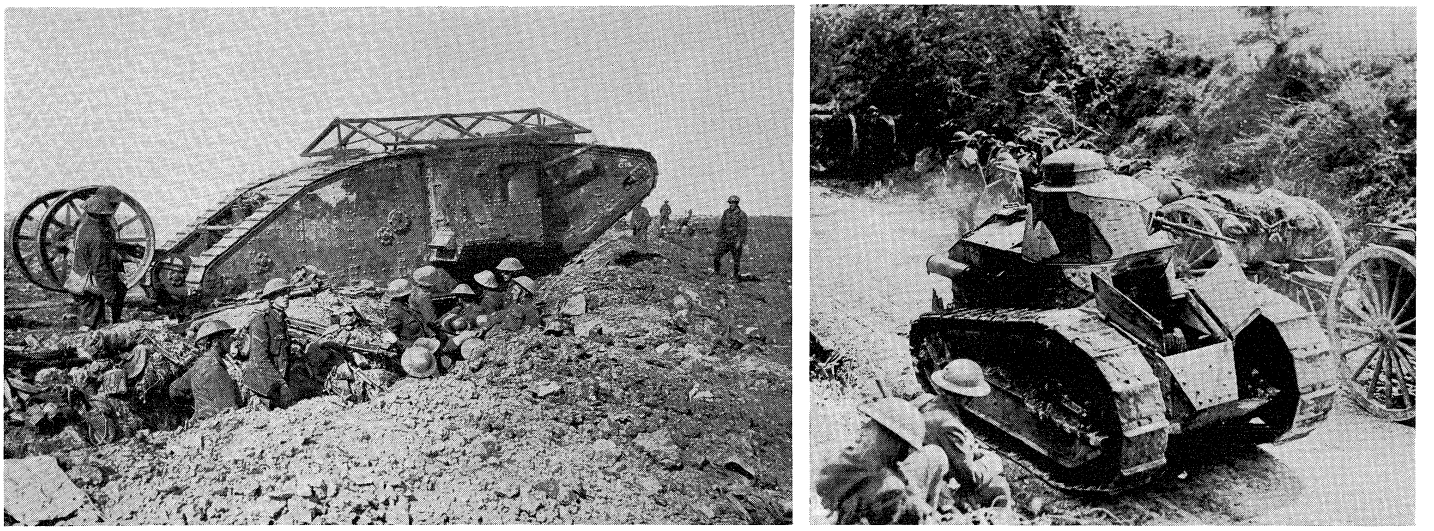


Figure 1: *Tanks of World War I.*
(Left) British Mark I tank with anti-bomb roof and "tail," 1916. (Right) French Renault F.T. light tank, 1918.
By courtesy of the Imperial War Museum, London; photographs, Camera Press—Pix from Publix

British Army ordered 160 of the new Vickers Medium tanks between 1923 and 1928. They were virtually the only tanks the British Army had until the early '30s and the only tanks to be produced in quantity anywhere in the world during the mid-'20s. Capable of 20 mph, the Vickers Mediums stimulated the Royal Tank Corps to develop mobile tactics. Various experiments during the '20s and early '30s resulted in general adoption of two separate categories of tanks. Mobile tanks were intended for the role performed earlier by horse cavalry. Slower but more heavily armoured tanks provided infantry support.

Tankettes

Before this division into mobile and slow tanks had crystallized, several different designs were tried. The British Independent tank of 1925 with five turrets started a trend toward many turrets among heavy tanks. Another trend setter was a small turretless tankette originated in Britain by Maj. G. le Q. Martel and J. Carden in the mid-'20s, and a slightly heavier, turreted two-man light tank. The number of light tanks and tankettes grew rapidly after 1929, as several countries started to produce armoured vehicles. The Soviet Union was by far the most important producer; on a much smaller scale Poland, Czechoslovakia, and Japan entered the field in 1930–31. Concurrently, tank production started up again in France and Italy. As tank production grew and spread among nations, the value of tankettes and light tanks armed only with machine guns decreased and heavier models armed with 37- to 47-mm guns for fighting other tanks began to displace them. An early example was the Vickers-Armstrongs 6-ton model of 1930, copied on a large scale in the Soviet Union (as the T-26). The most successful example was the BT, also built in large numbers in the Soviet Union. Fastest tank of its day, the BT was based on the designs evolved in the U.S. by J.W. Christie, who in 1928 built an experimental model capable of 42.5 mph. Christie's vehicles could run on wheels after the removal of tracks, and, far more significant, had road wheels independently suspended. This feature enabled them to move over broken ground faster than tanks with the usual arrangement of wheels.

Although they were relatively well armed and mobile, tanks of the T-26 and BT type were underarmoured (plates 10 to 15 millimetres thick) and were not, therefore, suitable for close infantry support. This was clearly demonstrated in 1937 during the Civil War in Spain where T-26 and BT tanks were used by the Communist forces. Even before this time, it had become clear that tanks that moved at the slow pace of the infantry and were therefore exposed to the full effect of anti-tank guns had to be thickly armoured. This realization led, in the mid-'30s, to such infantry tanks as the French R-35 with 40-mm and the British A.11 with up to 60-mm armour.

Apart from being lightly armoured, the Soviet BT, the equivalent British cruiser tanks, and the German Pz. III also required support from more heavily armed tanks if they were to engage in fighting of any intensity. The need for more powerful 75-mm gun tanks was clearly recognized in Germany, leading to 1934 design of the Pz. IV. The problem was realized less clearly in the Soviet Union, even though the T-28 and T-35 multiturret tanks with 76-mm guns were first built there in 1932–33. But the Russians recognized more quickly than others the need for the next step, which was to replace all the light–medium tanks armed with 37- to 47-mm guns by medium tanks armed with 75- or 76-mm guns. Thus, in 1939, while the Germans were still developing the Pz. III from a 37-mm to a 50-mm gun version, the Russians were already concentrating on the T-34 medium tank with a 76-mm gun.

In Britain even 40-mm gun tanks were scarce on the eve of World War II, and all but 80 of the 1,148 British tanks produced between 1930 and 1939 were still armed only with machine guns. Italy was even worse off, with only 70 M/11 tanks with 37-mm guns while the rest of its total of 1,500 were L/3 tankettes. Except for a few experimental models, the U.S. had only about 300 machine-gun-armed light tanks. Most of the 2,000 tanks produced in Japan were equally lightly armed. By comparison, France had a more powerful tank force—2,677 modern tanks, of which, however, only 172 were the Char *B*, armed with 75-mm guns. The largest force was the U.S.S.R.'s, which, as a result of a massive production program started 1930–31, had about 20,000 tanks by 1939, considerably more than the rest of the world put together.

World War II. The most effective tank force proved to be the German, composed in 1939 of 3,195 vehicles, including 211 Pz. IV's. What made the German tanks so formidable was that instead of being divided between various infantry and cavalry tank units they were all concentrated in the panzer divisions. The successes of the panzer divisions during the first two years of World War II led the major armies to reorganize most of their tanks into similar formations; this resulted in a dramatic increase in production, as shown in the following table:

German panzer divisions

Tank Production During World War II

year	country				
	Germany	Britain	U.S.	U.S.S.R.	Japan
1939	249	969	...	...	462
1940	1,460	1,399	331	2,794	1,023
1941	3,256	4,841	4,052	6,590	1,024
1942	4,278	8,611	24,997	24,668	1,165
1943	5,966	7,476	29,497	20,000	776
1944	9,161	...	17,565	17,000	342

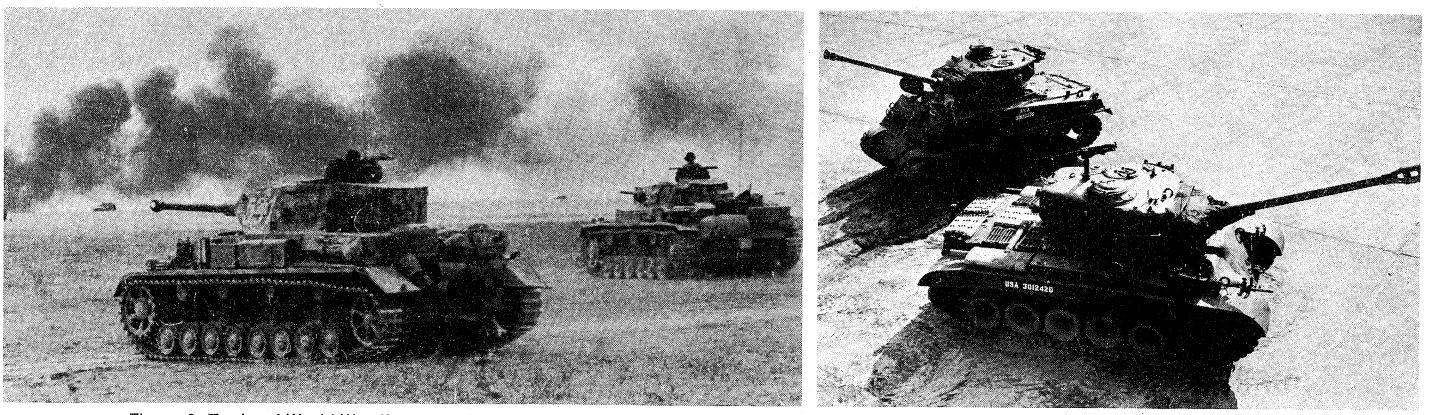


Figure 2: *Tanks of World War II.*
(Left) German Pz. IV (foreground) and Pz. III (background) tanks, 1942. (Right) U.S. Army M26 Pershing tank with 90-millimetre gun (foreground) and M4 Sherman tank with 76-millimetre gun.
By courtesy of (right) the U. S. Army; photograph, (left) Ullstein

The campaigns of 1940 and 1941 in which armoured forces played such an important role also intensified development of tanks and other armoured vehicles. The German Pz. IV and Soviet T-34 were rearmed in 1942 with longer barrelled, higher velocity guns; soon afterward, these began to be displaced by more powerfully armed tanks. In 1943 the Germans introduced the Panther medium tank with a long 75-mm gun having a muzzle velocity of 3,070 feet per second (ft/sec) compared with 1,260 ft/sec for the original Pz. IV and 2,460 ft/sec for its 1942 version (see Figure 2, left). The 43-ton Panther weighed almost twice as much as its predecessor and was correspondingly better armoured. The Germany also introduced the still more powerful Tiger tank, armed with an 88-mm gun. Its final version (Tiger II), at 68 tons, was to be the heaviest tank used during World War II. To oppose it, the Russians brought out the JS, or Stalin, heavy tank that appeared in 1944 armed with a 122-mm gun. Its muzzle velocity was lower than that of the German 88-mm guns, however, and it weighed only 46 tons. At about the same time the T-34 was rearmed with an 85-mm gun.

In contrast to the breakthrough role of the earlier heavy tanks, the Tiger and JS tanks functioned chiefly to support basic medium tanks by destroying enemy tanks at long range. German and Soviet armies also developed other heavy vehicles for this purpose, such as the 128-mm-gun Jagdtiger and the 122-mm-gun ISU, which, in effect, were turretless tanks. In addition, all armies developed lightly armoured self-propelled antitank guns and the U.S. Army developed a specialized category of tank destroyers, which resembled self-propelled guns in being relatively lightly armoured but, like tanks, had rotating turrets.

Turretless tank

The turretless tank type of vehicle originated with the Sturmgeschutz or assault gun introduced in 1940 by the German Army for infantry support but subsequently transformed into more versatile vehicles particularly suited for destroying enemy tanks. No such vehicles were produced in Britain or the U.S. To the end of World War II, however, the British Army retained a specialized category of infantry tanks, such as the Churchill, and of cruiser tanks, such as the Crusader and Cromwell. The former were well armoured, and the latter were fast, but none was well armed compared to German and Soviet tanks. As a result, during 1943 and 1944, British armoured divisions were mostly equipped with U.S.-built M4 Sherman medium tanks (see Figure 2, right). The M4 was preceded by the mechanically similar M3 medium, also armed with a medium-velocity 75-mm gun, but mounted in the hull instead of the turret, because this could be put into production more quickly when tanks were urgently required in 1940 and 1941. Production of the M4 began in 1942 (it replaced the stop-gap M3); eventually 49,234 were built, making it the principal tank of U.S. and other allied armoured forces. Successful when first introduced, it was, by 1944, no longer adequately armed and should have been replaced by a new medium tank. Several factors kept it in production, including a contemporary obsession with numbers inspired by the usual overestimates of enemy strength. Furthermore, the U.S. Army, like the British, believed the fallacious doctrine that armoured divisions should confine themselves to exploitation of infantry breakthroughs and did not, therefore, need powerfully armed tanks. Only toward the end of the war did the U.S. Army introduce a few M26 Pershing heavy tanks with a 90-mm gun comparable to that of the original German Tiger. Similarly, the British Army introduced the prototypes of the Centurion tank with a 76-mm gun comparable to that of the German Panther. Otherwise, U.S. and British tanks were well behind the German and Soviet tanks in their gun power.

POSTWAR TANK DESIGN

Tanks in the nuclear age

After World War II it was generally recognized that all tanks must be well-armed to fight enemy tanks. This finally ended the division of tanks into undergunned categories of specialized infantry and cavalry tanks, which the British Army retained longer than any other. Still not fully recognized were the advantages of concentrating tanks in fully mechanized formations; the British and U.S. armies continued to divide tanks between the armoured divisions and the less mobile infantry divisions. After World War II, tanks also suffered from one of the periodic waves of pessimism about their future. New antitank weapons, such as rocket launchers and recoilless rifles, and the mistaken belief that the value of tanks lay primarily in their armour protection caused this attitude. The Soviet Army, however, maintained large armoured forces; the threat they posed to western Europe as the Cold War became more intense, together with the havoc created by Soviet-built T34/85 tanks during the invasion of South Korea in 1950, provided a new impetus to development.

Development of tactical nuclear weapons in the mid-'50s changed the situation again. These new weapons dictated employment of mobile forces capable of dispersed operation in small self-contained units that could make up for the low troop density on the battlefield by their mobility and weapon power and that could better evade destruction by keeping the situation fluid. Nuclear weapons consequently encouraged the use of armoured forces because of their cross-country mobility and relatively high ratio of weapon to manpower. Moreover, tanks and other armoured vehicles, by virtue of the protection that they provide against blast and radioactivity, can operate fairly close to nuclear explosions and can cross areas that have been subjected to nuclear strikes without undue risk to their crews. Armoured units survive enemy nuclear weapons better and can better exploit the devastating effects of nuclear strikes. Thus tanks and other armoured vehicles have become even more important with the advent of tactical nuclear weapons.

The British Centurion. The most significant feature of tank development since World War II has been the emphasis placed on armament. Progress in this field is best illustrated by the British Centurion, which started in 1945

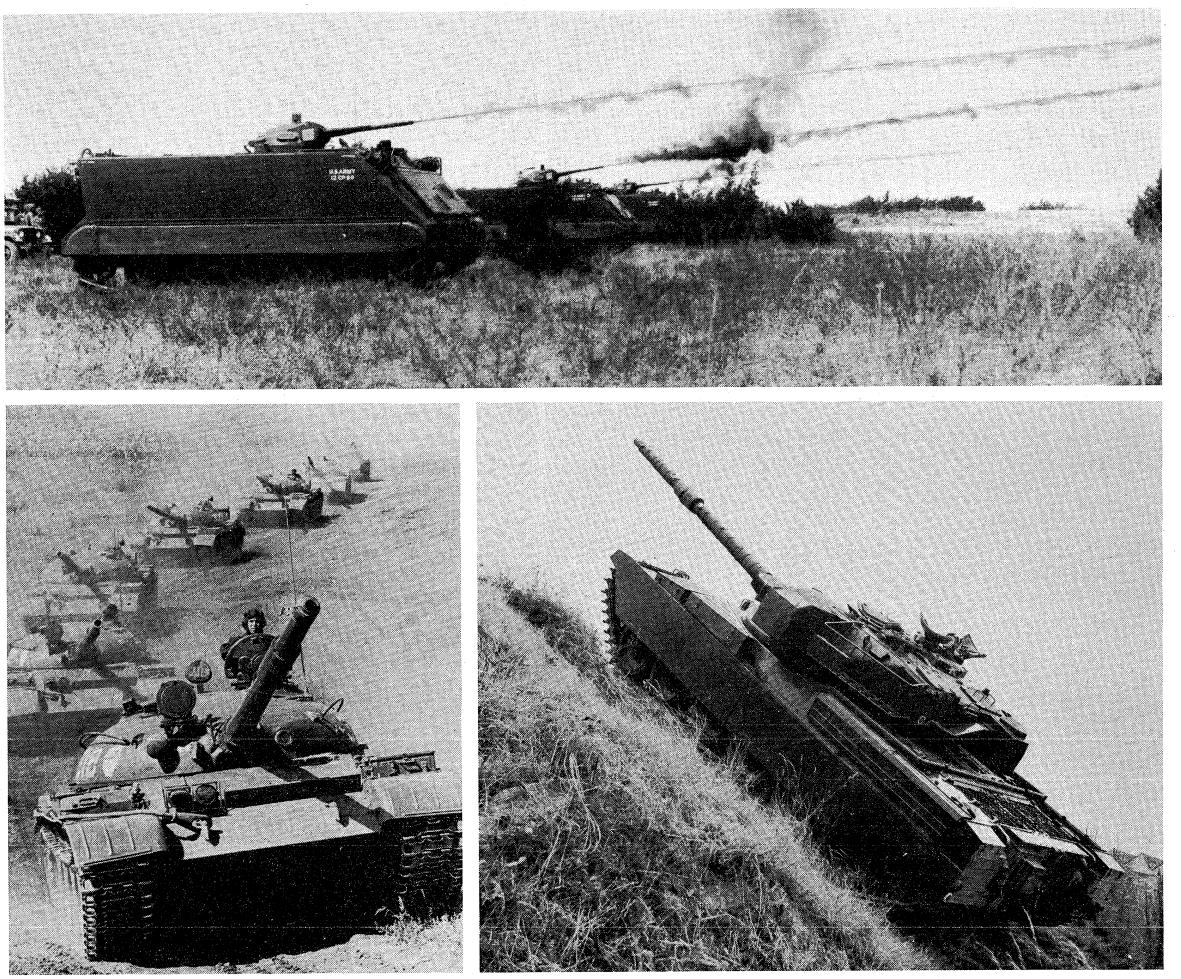


Figure 3: *Tanks and armoured vehicles since World War II.*
(Top) U.S. Army M132 armoured personnel carriers, mounted with flame throwers. (Left) Soviet T-62 tanks armed with 115-millimetre guns. (Right) British Chieftain tank with 120-millimetre gun.
By courtesy of (top) the U.S. Army, (right) Central Office of Information, London, Crown copyright reserved; photograph, (left) Tass—Sovfoto

with a 76-mm gun but in 1948 was rearmed with an 83.4-mm, 20-pounder gun firing newly developed, tungsten-carbide-cored, armour-piercing ammunition with a muzzle velocity of 4,800 ft/sec. As a result, the 83.4-mm gun could penetrate armour almost twice as thick as that which the most powerful 88-mm gun of the German Tiger penetrated. In 1959 the Centurion was again upgunned, this time with a 105-mm weapon.

As a result of its outstanding performance the Centurion was adopted not only by Britain and major Commonwealth countries but also by Sweden, Switzerland, and later Israel; it was also delivered under the U.S. military aid program during the 1950s to The Netherlands and Denmark, the total exported from Britain amounting to more than 2,500. In the meantime the British Army had developed an even more powerful tank, the Conqueror, which was armed with a 120-mm gun firing armour-piercing ammunition, introduced in 1954 as a heavy gun tank to support the Centurions. Unfortunately its thick armour gave the Conqueror a weight of 65 tons in contrast to the Centurion's 50 tons; when the Centurions were rearmed with the longer ranged 105-mm guns, the Conqueror was withdrawn from service (1958).

Soviet tanks. At about the same time the Soviet Army also began to withdraw some of its heavy tanks, the JS-3, and later the T-10, both armed with 122-mm guns. The basic Soviet tank for many years after World War II was the T-34/85, of which 11,778 had been produced in 1944. It was finally replaced by the T-54, which evolved from the T-34/85 through the T-44 but had a 100-mm gun in a much better profiled turret, a crew of four instead of five, a transversely mounted engine, and other improvements. Introduced around 1953, the T-54 became during the late '50s the standard tank of the Communist bloc, manufactured not only in the U.S.S.R. but also in Poland and Czechoslovakia. A copy of the T-54 was also produced in China, which exported some to Pakistan in 1966 while the U.S.S.R. supplied more than 2,000 to Egypt, Syria, Iraq, and Algeria after 1955 and exported some to Finland, Afghanistan, and Cuba. In 1963, however, the T-54 and its improved version, the T-55, began to be replaced in the Soviet Army by the current T-62 (see Figure 3, left). The T-62, externally similar to the T-54, has a much more powerful, high-velocity 115-mm gun; it represents concentration on a single type of battle tank as well-armed as contemporary technology will allow.

U.S. tanks. At first the U.S. Army also subscribed to a policy of developing heavy as well as medium tanks. Formulated in 1946 this policy led to the design of the 90-mm-gun T42 medium and the 120-mm-gun T43 heavy tanks. Designated as the M103, a few T43s were built, between 1951 and 1954, mainly for the U.S. Marine Corps. Meanwhile the T42 was supplanted in 1948 by the M46 medium tank. This was basically the earlier M26 heavy tank with a new 810 horsepower air-cooled engine and cross-drive transmission. After the conflict in Korea in 1950, the turret of the T42, which for the first time incorporated a stereoscopic range finder, was mounted on the hull of the M46 to produce the M47 medium. This, in turn, was succeeded in 1952 by the M48, which was similar to the M47 except for a ballistically superior turret and hull and a four-man instead of five-man crew.

The M48 became virtually the only battle tank of the U.S. Army during the late '50s, while the M46 and M47

tanks were transferred to the French and Italian armies. More M47 tanks were later supplied to Belgium, West Germany, Greece, Spain, Turkey, and several other countries, followed still later by deliveries of the M48, which was being replaced in the U.S. Army by the 47-ton M60. The M60 was essentially the M48 with an air-cooled diesel in place of the gasoline engine and a British-designed 105-mm gun instead of the 90-mm gun of the earlier American tanks.

The 105-mm British tank gun was also adopted for the Swiss Pz. 61, the first tank ever produced in Switzerland. One of its characteristics is its relatively light weight of 36.5 tons. However, an even lighter tank, the 34.5-ton Type 61 was produced in Japan, the first after the demilitarization of 1945; it was armed with the same 90-mm gun as the U.S. M48.

France and Germany. After World War II the French Army started developing a single battle tank, the AMX-50, armed at first with a 100-mm gun and then a 120-mm gun. After building a number of prototypes, however, the French concluded that the disadvantages of such heavy, 50-ton tanks outweighed their advantages and developed instead a lighter battle tank, the AMX-30. This 36-ton tank is armed with a French-designed 105-mm gun, which fires an unusual type of armour-piercing projectile with a shaped charge mounted in ball bearings to prevent spinning and thereby retain its performance. The AMX-30, first produced in 1966, was originally designed to a specification agreed to in 1957 by France, Germany, and Italy. Italy subsequently withdrew from the agreement, however, and adopted the U.S. M60, some of which have since been built in Italy. France and Germany, on the other hand, adopted their own alternative designs, the German being the Leopard. The first prototypes of the Leopard were built in 1961, the first tanks built in Germany since 1945. The first of 1,845 Leopards ordered by the German Army were built in 1965; subsequently Leopards were ordered by the Belgian, Norwegian, and Dutch armies to replace their earlier U.S. or British tanks. The Leopard is armed with a British 105-mm tank gun and in spite of being four tons heavier than the AMX-30 it has the same high power-to-weight ratio of 20 horsepower per ton due to its 830-hp diesel.

The German Leopard

Chieftain. A tank having the same main armament and general characteristics as the Leopard was developed at about the same time in Britain by the Vickers Company but for the Indian and not the British Army. Called the Vijayanta, production began in 1966 in India where no tank had been built before. But in contrast to the Leopard, as well as the AMX-30, Pz. 61, and the U.S. M47, M48, and M60, the Vickers Vijayanta does not have an optical range finder. Instead it has a 0.5-inch ranging machine gun, in addition to the usual rifle-calibre coaxial machine gun. The ranging-machine-gun system was originally introduced on the upgunned 105-mm Centurions and adopted in the later British Chieftain. The Chieftain (Figure 3, right), which first came into service in 1966, is armed with a high-velocity 120-mm gun and weighs 51 tons. In the early '70s it was the most heavily armed and armoured battle tank in service anywhere in the world. It also embodies several novel features, including a supine position for the driver to lower the height of the hull, a two-stroke, opposed-piston diesel engine and, recently, a laser range finder.

Sweden's S-tank. The most original design is, however, the Swedish S-tank, developed since 1956 and adopted by the Swedish Army in 1964. This turretless vehicle has its 105-mm gun fixed to the hull so that the gun can only be elevated by altering the pitch of the hull. This is accomplished by means of an adjustable hydropneumatic suspension. To turn the gun in a different direction, the crew must turn the entire tank. Elimination of the turret and the fixed gun mounting considerably reduced the silhouette of the S-tank and facilitated installation of a loading mechanism, which made it possible to reduce the size of the crew from four to three. The consequent reduction in tank size has kept the weight down to 36.5 tons without sacrificing armour.

The S-tank also has an original engine installation consisting of a diesel and a gas turbine coupled to a common output shaft. All other battle tanks since the U.S. M48 and the British Centurion and an increasing number of lighter armoured vehicles are powered by diesel, rather than gasoline engines, because of their greater overall efficiency and range.

MBT-70. Another significant departure from general practice is represented by the U.S. M60A1E1 and M60A1E2 tanks, which have a 152-mm gun-launcher. The gun-launcher fires Shillelagh guided missiles as well as conventional projectiles but its barrel is too short to fire high-velocity armour-piercing shots; this reduces its flexibility in attacking enemy tanks. This shortcoming has been corrected in a more recent version of the 152-mm gun-launcher, which can fire armour-piercing ammunition as well as high explosive shells and missiles and which was mounted in the MBT-70. Development of the MBT-70 was started by the U.S. and Germany in 1963; in addition to the 152-mm gun-launcher it incorporated several advanced features, including a 1,475 hp diesel, which gave a power-to-weight ratio of almost 30 hp per ton, and an adjustable hydropneumatic suspension instead of the customary torsion-bar suspension. The MBT-70 was also provided with an automatic loader, which reduced its crew to three men. For the first time the crew was concentrated in the turret, which became an environmental control capsule protecting from radiological, chemical, and biological effects. Location of the driver in the turret alongside instead of below other crew members reduced the height of the MBT-70. This necessitated, however, a counter-rotating cupola for the driver to keep him pointed forward irrespective of the rotation of the turret, thus complicating the driving controls. All the sophistication incorporated in the MBT-70 made it very expensive. Production cost per tank was estimated in 1969 at $750,000 to $850,000, compared with $200,000 to $250,000 for the M60, Leopard, and Chieftain. As a result, Germany withdrew from its joint development in 1970 and the U.S. Army in 1972 abandoned its more austere XM803 version.

The U.S. gun-launcher

The most controversial feature of the MBT-70/XM803 was its gun-launcher. Its Shillelagh missiles gave it an advantage over enemy tanks at long ranges where tank guns are rather inaccurate, but they were highly questionable on cost-effectiveness grounds at the usual tank combat ranges of less than 2,000 metres, since one missile costs about $3,000, or the equivalent of 20 armour-piercing rounds. The gun-launcher is also a compromise between the requirements of gun and missile systems; it necessitates an overlarge gun to accommodate the large diameter missile and at the same time it prevents the missiles being launched without the vehicle exposing itself. Consequently, at an early stage the Germans wanted to rearm the MBT-70 with a 120-mm gun and favoured the development of special vehicles to fire missiles only. These are exemplified by the Jagdpanzer Rakete, a tank destroyer armed with SS-11 or Hot missiles.

OTHER ARMOURED VEHICLES

Airborne and amphibious vehicles. In 1941, the U.S. Army commenced development of the T9, the first air-transportable light tank. The first tanks actually flown into battle, however, were British Tetrarchs, a few of which were landed by glider in Normandy on D-Day in 1944. After 1946, the French Army built the AMX-13, a 14.3-ton air-transportable vehicle (75-mm gun), used chiefly as a tank destroyer and reconnaissance vehicle. Less successful was the M41 76-mm-gun light tank, constructed in 1950 by the U.S. Army. At 22.8 tons, this vehicle proved too heavy for contemporary transport aircraft. The latest U.S. light tank is the M551 Sheridan, equipped with a 152-mm gun-launcher.

The Sheridan was designed to meet two military needs, that of reconnaissance for a highly mobile vehicle, and that of airborne operations for a light but effective self-propelled weapon. Usually these two requirements have been met by separate vehicles. Thus the Soviet Army has the ASU-85, an 85-mm turretless assault gun, for its airborne forces, and the PT-76 amphibious light tank for

its reconnaissance units. The development of amphibious tanks dates back to the Medium D of 1919. Though all amphibious tanks have sacrificed armament for light weight and have been bulky to achieve sufficient buoyancy, the overall mobility of tanks such as the PT-76 made them effective reconnaissance vehicles.

Since the middle of World War II many other tanks have been made amphibious by collapsible flotation screens, which, when erected, allow even relatively heavy tanks to float. This method was first used with M4 medium tanks during the D-Day landings in Normandy in 1944. More recently, flotation screens have been permanently installed on the S-tank, the Vickers Vijayanta, the M551 Sheridan, and the Scorpion light reconnaissance tank introduced in 1969 by the British Army.

An alternative method to crossing rivers is "submerged fording," first tried in 1940 with the British A.9 cruiser and the German Pz. III and IV. Since World War II, provision for submerged fording has been built into several tanks including the Soviet T-54 and T-62 and the Leopard and the AMX-30.

Landing Vehicle Tracked

An entirely different problem is posed by amphibious landings from the open seas; to solve it the U.S. Marine Corps developed the Landing Vehicle Tracked, or LVT. Originally built in 1941 as an unarmoured cargo carrier, the LVT quickly acquired armour. Two types evolved: an armoured amphibious personnel and cargo carrier, and a turreted amphibious gun-vehicle for close fire support during landing operations. Altogether 18,620 LVTs were built during World War II; these played a prominent role in the Pacific campaigns from Guadalcanal onward. After World War II, LVTs were successfully used in Korea, notably for the 1950 Inchon landing. Two new models were built between 1951 and 1957: a 39-ton LVTP5 amphibious carrier, capable of carrying as many as 37 men, and an LVTH6 armed with a turret-mounted 105-mm howitzer. Both are still in service but a new LVTPX12 was developed during the late '60s to replace them. It incorporates several improvements, the most important being a boatlike hull with a stern instead of bow loading ramp and two water-jet propulsion units that greatly improve its performance in comparison with that of the earlier LVTs (propelled in water as well as on land by means of their tracks). At the same time the LVTPX12 has retained the seagoing qualities of the earlier LVTs, which could negotiate rough seas and Pacific surf in contrast to other amphibious vehicles intended primarily for crossing inland water obstacles. The use of water jet propulsion units in the LVTPX12, however, has been preceded by their use in several amphibious reconnaissance vehicles, including the Soviet PT-76.

Self-propelled guns. Development of self-propelled guns has also continued since World War II, except for self-propelled antitank guns. These became largely superfluous when it was recognized that all tanks needed to be sufficiently well armed to fight enemy tanks. Turretless assault guns, much favoured during World War II by the German and Soviet armies, have also largely disappeared except for a few specialized models such as the Soviet ASU-85 and the 90-mm-gun Jagdpanzer Kanone of the German Army. The general trend among the remaining self-propelled guns has been either toward lightly but completely armoured models with guns mounted in turrets with all-round traverse, like most tanks, or toward partially armoured models that are essentially tracked chassis with guns mounted on top. Examples of completely armoured guns are the 155-mm M109 self-propelled howitzer used by the U.S. and several west European armies, and the British 105-mm Abbot self-propelled gun. Partially armoured models include the U.S.-built M107 175-mm gun, the M110 8-inch howitzer, and the French 155-mm howitzer on the AMX-13 chassis. Though the Soviet Army has not produced either type of self-propelled gun, continuing to rely on tractor-towed guns, it has developed several armoured self-propelled launchers for its tactical missiles that make use of PT-76 and IS tank chassis.

Armoured personnel carriers. Since World War II armoured personnel carriers have become the next most important armoured vehicle after battle tanks. Though a few experimental models were built in Britain at the end of World War I, development of armoured carriers did not really begin until they were adopted for the panzer division infantry at the beginning of World War II. Germany's example was quickly followed by the U.S. which, by the end of the war, had produced 41,000 carriers. Both the German and U.S. carriers of World War II were of the half-track type and provided only light protection; nevertheless, they represented a major advance on the earlier method of transporting infantry into battle in unarmoured trucks. Moreover, the panzer grenadiers used them effectively as combat vehicles and fought from them on the move, thus greatly increasing the mobility of the infantry on the battlefield.

After World War II the U.S. Army led in developing fully tracked carriers with all-round protection (see Figure 3, top). The first postwar carrier was the M44 armoured utility vehicle of 1945. Though unduly large, this vehicle demonstrated the advantages of its type. Next came the M75, first built in 1948. Essentially a scaled-down M44 with a similar box body, it carried 12 instead of 27 men. Adopted in 1953, the M75 became the first tracked armoured personnel carrier used in quantity. The war in Korea, in which a few M75s successfully carried infantry under fire, added impetus to armoured personnel carrier development. When the 1955 tactical nuclear weapons tests in Nevada demonstrated that armoured carriers permitted infantry to operate relatively close to nuclear explosions, armoured carriers were demanded for infantry as well as armoured divisions.

The M75 in Korea

By 1955 the M59 began to replace the M75. Though similar, the newer M59 was less expensive and able to swim across calm inland waters. In 1960 the M59 was supplanted by the current M113, lower and considerably lighter due partly to the contemporary development of aluminum armour. The M113 was in fact the first aluminum-armoured vehicle to be put into large-scale production. Since its appearance an increasing number of the lighter vehicles have been built using aluminum armour, including the M109 155-mm self-propelled howitzer, M551 Sheridan, LVTPX12, Scorpion, and the Fox armoured car. Battle tanks, however, have continued to use steel armour, which is more effective against large-calibre armour-piercing projectiles and does not have to be as thick as equivalent aluminum armour.

As well as being lighter, the M113 was also less expensive than its predecessors, costing only $22,000, a significant factor in its large-scale adoption, first by the U.S. Army and then by several other armies. By 1970 about 40,000 M113 carriers had been produced, making the type the most numerous armoured vehicle outside the Soviet bloc. M113 carriers have been used on a large scale in the war in Vietnam, often as combat vehicles, though they were not designed for the role and are at a disadvantage in spite of the addition of roof-mounted machine guns with shields. The shortcomings of the M113 stem from the original U.S. Army doctrine that infantry must dismount from carriers to close with the enemy. As a result the M113 and the earlier U.S. carriers were designed as transport rather than fighting vehicles.

The first attempt to produce a tracked armoured carrier from which the infantry could fight on the move was made by the French Army. The AMX-VTT production began in 1958; from within, its occupants could fire their own hand weapons and a machine gun mounted in a turret. A further step forward was taken when the German Army ordered the HS-30 carrier for its panzer grenadiers. This offered better protection and had a turret with a 20-mm automatic cannon, which enabled it to fight other light armoured vehicles.

By the mid-'60s the need to replace the original type of carrier by what came to be known as the mechanized infantry combat vehicle, or MICV, was generally recognized and the U.S. Army developed its first vehicle of this type. The experimental XM701 had a 20-mm gun turret and provision for its occupants to fire their own weapons from within. A basically similar vehicle, the Marder, was adopted by the German Army.

Like the British, the Soviet Army originally relied mainly on wheeled armoured carriers. The early BTR-152s were little more than armoured six-wheeled trucks but the more recent BTR-60P is a much more advanced amphibious vehicle with eight-wheel drive and a water jet-propulsion unit. The first Soviet tracked carrier, the BTR-50P, introduced around 1955, was little more than the PT-76 tank without its turret; but in 1967 the Soviet Army revealed a much more advanced tracked amphibious armoured personnel carrier or mechanized infantry combat vehicle. This has a turret with a medium-calibre gun and a machine gun and is also fitted with antitank guided missiles. In consequence it represents a versatile fighting unit, which, in some circumstances, could rival tanks as the basic armoured vehicle.

BIBLIOGRAPHY. P. CHAMBERLAIN and C. ELLIS, *British and American Tanks of World War II* (1969), an extensive collection of photographs and descriptions of British and U.S. tanks built during WW II; A. DUVIGNAC, *Histoire de l'armée motorisée* (1948), an account of the development of French tanks and other military vehicles from early attempts to 1937–38; J.F.C. FULLER, *Tanks in the Great War, 1914–1918* (1920), a contemporary account of the development and employment of British tanks and tank units during WW I, with some discussion of French and U.S. tanks; O.H. HACKER *et al.*, *Heigls Taschenbuch der Tanks*, 2 vol. (1935), a classic work, although inaccurate on a number of points, containing a fully illustrated catalogue of armoured vehicles of all countries to 1934–35; R.J. ICKS, *Tanks and Armored Vehicles* (1945), an account of developments to 1943–44, similar in principle to Hacker *et al.*, but less extensive; R.M. OGORKIEWICZ, *Design and Development of Fighting Vehicles* (1968), the only book in English treating the more technical aspects of the design and development of armoured vehicles, and *Armoured Forces* (1970), a comprehensive account of the development of armoured forces, their organization, and the different types of vehicles employed to 1956–57, with revisions to 1969; F.M. VON SENGER UND ETTERLIN, *German Tanks of World War II* (Eng. trans. 1969), an extensive collection of illustrations and descriptions of German armoured vehicles, 1926–45, and *Taschenbuch der Panzer* (1969), an illustrated catalogue of armoured vehicles used throughout the world, 1960–70; B.T. WHITE, *British Tanks and Fighting Vehicles 1914–1945* (1970), photographs and descriptions of British armoured vehicles, and *Tanks and Other Armoured Fighting Vehicles 1900–1918* (1970), descriptions of the earliest armoured vehicles.

(R.M.O.)

Tanzania

The United Republic of Tanzania is a sovereign state located in eastern Africa. It was formed in 1964 by the union of the Republic of Tanganyika on the East African mainland with the offshore state of Zanzibar, comprising the islands of Zanzibar and Pemba. The total area of the republic is 364,945 square miles (945,203 square kilometres). The area on the mainland is 363,728 square miles, of which nearly 22,805 square miles consists of water. The island of Zanzibar has an area of 640 square miles and the island of Pemba an area of 380 square miles. The two islands are separated from the mainland by a channel, which, at its narrowest, is about 22 miles wide. Farther south lies the island of Mafia, with an area of 200 square miles; it is administered from the mainland. The United Republic has a total population of over 14,000,000, of whom about 13,600,000 live in the mainland and about 400,000 on Zanzibar. The capital is the port city of Dar es Salaam.

To the north, Tanzania is bounded by Kenya and Uganda, which converge on Lake Victoria; about half of the lake lies within Tanzania; to the northwest lie the republics of Rwanda and Burundi. To the west, the waters of Lake Tanganyika (of which about half belongs to Tanzania) separate the republic from Zaire. To the southwest Tanzania is bordered by Zambia and Malawi. Since the 1960s part of the Malawi–Tanzania boundary has been in dispute, with Malawi claiming the eastern shore of Lake Nyasa and Tanzania claiming a border through the lake to join the Tanzania–Malawi–Mozambique tripoint in the southeast end. To the south the Ruvuma River is the land boundary between Tanzania and Mozambique.

In the 20th century the mainland territory has been, successively, a German colony, a British-administered League of Nations mandate, a United Nations trust territory under British administration, and, finally—since 1961—an independent state within the Commonwealth of Nations. Tanzania is one of the larger African countries and has a primarily agricultural economy. During the period of British colonial administration, the limited development funds available for East African territories were often devoted to Kenya or to Uganda rather than to Tanzania, so that today the country has more to do to catch up with these two neighbours, with which it is now linked through membership in the East African Community. Cash crops grown for export include sisal, cotton, and coffee; diamonds also are exported. Under the leadership of President Julius Nyerere a program of African socialism has been launched, of which an important feature is the establishment of *ujamaa vijijini* (cooperative village communities). The Tanzam railway, 1,100 miles long, scheduled to be completed in the mid-1970s, was being constructed (with aid from the People's Republic of China) to transport Zambian copper to the port of Dar es Salaam, thereby lessening land-locked Zambia's dependence upon white-administered southern African states, while opening up parts of Tanzania for development. For associated physical features, see EAST AFRICAN LAKES; EAST AFRICAN MOUNTAINS; for historical aspects, see EAST AFRICA, HISTORY OF.

Developmental efforts

THE LANDSCAPE

Relief features. Although most of Tanzania consists of plain and plateau, it has some spectacular relief features, including Africa's highest mountain, Kilimanjaro (19,340 feet, or 5,895 metres high), and Lake Tanganyika (1,174 feet, or 358 metres, below sea level) which forms part of the world's greatest rift-valley system. The coastal area is low-lying and is narrow, except where rivers flow into it from the west.

The mountain system is sprawled in a twisting pattern, extending north to southwest. In the southwest, the tail of the pattern is formed by the Ufipa Plateau, the Mbeya Range, and Rungwe Mountain. Running northward in elongated form are the Southern Highlands, and the Ukaguru and Nguru mountains. Near the coast and aligned from southeast to northwest are the Usambara and Pare mountains; the highland complex culminates in the snowclad peak of Kilimanjaro, west of which lies Meru (14,979 feet, or 4,565 metres) and the perfect rim of the Ngorongoro Crater (volcanic). The Mbulu-Kondoa ranges and the Kiboriani Mountains complete the pattern where they meet the Ukaguru Range. The Uluguru Mountains stand 120 miles west of Dar es Salaam and east of the Rubehorn Mountains; overlooking Lake Nyasa in the south are the Kipengere and Livingstone mountains.

Northward from Lake Nyasa the rift system bifurcates. The Western Rift is well defined, especially by the deep, linear form of Lake Tanganyika. The Eastern Rift is well defined only in the north and central parts, where faulting is complex. Ol Doinyo Lengai, west of Kilimanjaro and south of the small Lake Natron, is the only active volcano. Directly or indirectly, the rifts have produced depressions such as the Pangani Valley (which extends southeastward from Kilimanjaro to the Indian Ocean) and the Malagarasi Swamp (in the northwestern part of the Central Plateau), formed numerous volcanic features, altered the pattern of drainage, and caused shallow lakes. Lake Victoria, however, which ranks as the third largest lake in the world, is not part of the Rift System.

The rivers

The largest river, the Rufiji, together with its tributaries, drains most of southern Tanzania. Like the snowfed Pangani, the Wami, the Ruvu, and the Ruvuma rivers, it also empties into the Indian Ocean. The Kagera and several small rivers drain into Lake Victoria, while the Bubu, the Wembere, and some other rivers flow into interior basins.

Soils. Tanzania possesses a greater variety of soils than any other African country. On a regional basis, eluvials (soils formed from the decomposition of rock and found in their place of origin) and illuvials (soil material that has been leached out or washed away from

MAP INDEX

Political subdivisions

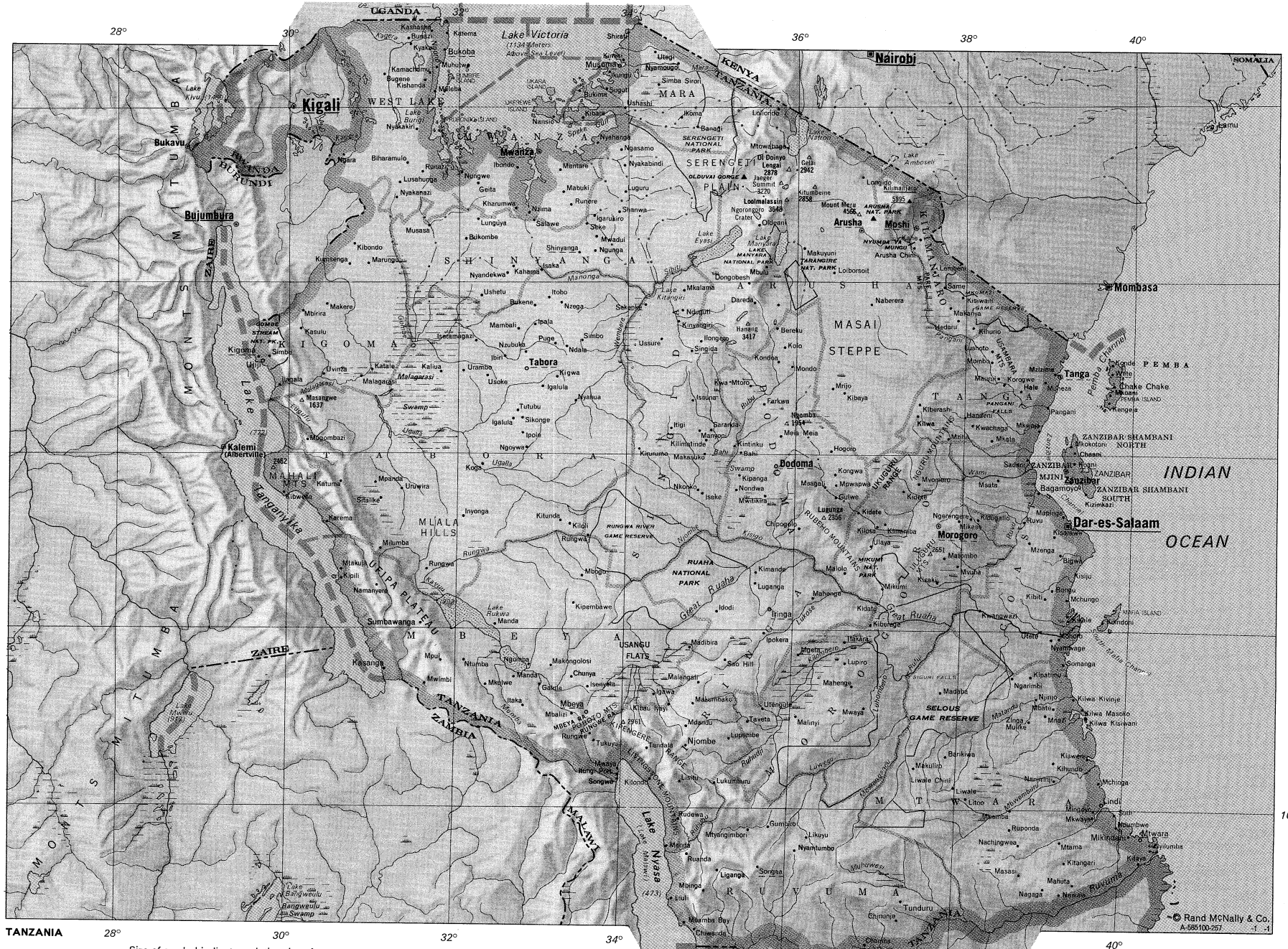
Arusha..... 4·00s 36·15e
Coast..... 7·00s 39·00e
Dodoma..... 6·00s 36·00e
Iringa..... 9·00s 35·00e
Kigoma..... 4·30s 30·30e
Kilimanjaro..... 3·45s 37·45e
Mara..... 1·45s 34·00e
Mbeya..... 8·30s 33·00e
Morogoro..... 8·00s 37·00e
Mtwara.....10·00s 39·00e
Mwanza..... 2·45s 32·45e
Pemba..... 5·10s 39·48e
Ruvuma.....11·00s 36·00e
Shinyanga..... 3·45s 33·00e
Singida..... 5·30s 34·30e
Tabora..... 6·00s 32·00e
Tanga..... 5·00s 38·15e
West Lake..... 2·00s 31·30e
Zanzibar Mjini... 6·10s 39·12e
Zanzibar Shambani North..... 5·55s 39·20e
Zanzibar Shambani South..... 6·15s 39·28e

Cities and towns

Arusha..... 3·22s 36·41e
Arusha Chini.... 3·35s 37·20e
Bagamoyo..... 6·26s 38·54e
Bahi..... 5·59s 35·19e
Banagi..... 2·16s 34·51e
Bereku..... 4·27s 35·44e
Bigwa..... 7·13s 39·09e
Biharamulo..... 2·38s 31·20e
Bugene..... 1·35s 31·08e
Bukene..... 4·14s 32·53e
Bukima..... 1·48s 32·25e
Bukoba..... 1·20s 31·49e
Bukombe..... 3·31s 32·03e
Bunazi..... 1·13s 31·24e
Bungu..... 7·38s 39·03e
Chaani..... 5·57s 39·18e
Chake Chake.... 5·15s 39·46e
Chamba.....11·35s 36·58e
Chinunje.....11·19s 37·19e
Chipogolo..... 6·52s 36·02e
Chiwanda.....11·22s 34·54e
Chunya..... 8·32s 33·25e
Dareda..... 4·13s 35·33e
Dar es Salaam... 6·48s 39·17e
Dodoma..... 6·11s 35·45e
Dongobesh..... 4·04s 35·23e
Farkwa..... 5·24s 35·36e
Galula..... 8·36s 33·02e
Geita..... 2·52s 32·10e
Gulwe..... 6·30s 36·29e
Gumbiro.....10·16s 35·39e
Hale..... 5·17s 38·36e
Handeni..... 5·26s 38·01e
Hedaru..... 4·30s 37·54e
Hogoro..... 5·57s 36·27e
Ibiri..... 4·56s 32·33e
Ibondo..... 2·38s 32·40e
Idodi..... 7·47s 35·11e
Ifakara..... 8·08s 36·41e
Igalula..... 5·14s 33·00e
Igalula..... 5·38s 32·38e
Igarukiro..... 3·08s 33·31e
Igawa..... 8·35s 34·28e
Ikoma..... 2·04s 34·37e
Ikungu..... 1·34s 33·40e
Ilagala..... 5·12s 29·50e
Ilongero..... 4·40s 34·52e
Inyonga..... 6·43s 32·04e
Ipala..... 4·30s 32·53e
Ipokera..... 8·03s 35·41e
Ipole..... 5·47s 32·44e
Iringa..... 7·46s 35·42e
Isaka..... 3·54s 32·56e
Iseke..... 6·25s 35·01e
Isenyela..... 8·36s 33·30e
Iseramagazi..... 4·40s 32·09e
Issuna..... 5·23s 34·46e
Itaka..... 8·52s 32·47e
Itigi..... 5·42s 34·29e
Itobo..... 4·10s 33·01e
Itungi Port..... 9·35s 33·56e
Kahama..... 3·50s 32·36e
Kahe..... 3·30s 37·26e
Kalema..... 1·12s 31·50e
Kaliua..... 5·04s 31·48e
Kamachumu..... 1·35s 31·37e
Karema..... 6·49s 30·26e
Kasanga..... 8·28s 31·09e
Kashasha..... 1·44s 31·37e
Kasulu..... 4·34s 30·06e
Katale..... 4·59s 31·03e
Katuma..... 6·10s 30·34e
Kengeja..... 5·25s 39·44e
Kharumwa..... 3·12s 32·39e
Kibara..... 2·09s 33·27e
Kibau Iyayi..... 8·52s 34·32e
Kibaya..... 5·18s 36·34e
Kiberashi..... 5·23s 37·26e
Kiberege..... 7·57s 36·52e
Kibiti..... 7·44s 38·57e
Kibondo..... 3·35s 30·42e
Kibwesa..... 6·28s 29·57e
Kidatu..... 7·42s 36·57e
Kidete..... 6·25s 37·16e
Kidete..... 6·39s 36·42e
Kidugallo..... 6·47s 38·12e
Kigoma..... 4·52s 29·38e
Kigwa..... 5·10s 33·08e
Kihundo..... 9·25s 38·59e
Kihurio..... 4·28s 38·04e
Kikale..... 7·50s 39·12e
Kilimatinde..... 5·51s 34·57e
Kilindoni..... 7·55s 39·39e
Kiloli..... 6·50s 33·23e
Kilondo..... 9·46s 34·21e
Kilosa..... 6·50s 36·59e
Kilwa..... 5·32s 37·31e
Kilwa Kisiwani... 8·58s 39·30e
Kilwa-Kivinje.... 8·45s 39·24e
Kilwa Masoko.... 8·56s 39·31e
Kimamba..... 6·47s 37·08e
Kimande..... 7·22s 35·30e
Kinesi..... 1·28s 33·52e
Kintinku..... 5·53s 35·14e
Kinyangiri..... 4·27s 34·37e
Kipanga..... 6·14s 35·21e
Kipatimu..... 8·29s 38·56e
Kipembawe..... 7·39s 33·24e
Kipili..... 7·26s 30·36e
Kirurumo..... 5·53s 34·11e
Kisaki..... 7·28s 37·36e
Kisarawe..... 6·54s 39·04e
Kishanda..... 1·42s 31·34e
Kisiju..... 7·24s 39·20e
Kisiwani..... 4·08s 37·57e
Kiswere..... 9·26s 39·33e
Kitangari.....10·39s 39·20e
Kitaya.....10·39s 40·10e
Kitunda..... 6·48s 33·13e
Kizimkazi..... 6·27s 39·28e
Koani..... 6·08s 39·17e
Koga..... 6·14s 32·25e
Kolo..... 4·44s 35·50e
Konde..... 4·57s 39·45e
Kondoa..... 4·54s 35·47e
Kongwa..... 6·12s 36·25e
Korogwe..... 5·09s 38·29e
Kumsenga..... 3·47s 30·25e
Kwachaga..... 5·38s 38·08e
Kwa Mtoro..... 5·14s 35·26e
Kwangwazi..... 7·47s 38·15e
Kyaka..... 1·16s 31·25e
Lembeni..... 3·47s 37·37e
Liganga.....10·41s 35·15e
Likuyu.....10·20s 36·14e
Lindi.....10·00s 39·43e
Lisitu..... 9·39s 34·39e
Litoo..... 9·54s 38·24e
Liuli.....11·05s 34·38e
Liwale..... 9·41s 38·01e
Liwale Chini..... 9·41s 38·01e
Loiborsoit..... 3·52s 36·26e
Loliondo..... 2·03s 35·37e
Longido..... 2·44s 36·41e
Luganga..... 7·31s 35·32e
Luguru..... 2·55s 33·58e
Lukumburu..... 9·45s 35·09e
Lunguya..... 3·23s 32·24e
Lupembe..... 9·15s 35·15e
Lupiro..... 8·23s 36·40e
Lusahunga..... 2·52s 31·15e
Lushoto..... 4·47s 38·17e
Mabuki..... 2·59s 33·11e
Madaba..... 8·40s 37·47e
Madibira..... 8·12s 34·49e
Mahenge..... 7·38s 36·16e
Mahenge..... 8·41s 36·43e
Mahuta.....10·52s 39·27e
Makanya..... 4·20s 37·51e
Makasuko..... 6·00s 34·56e
Makere..... 4·17s 30·25e
Makongolosi..... 8·24s 33·09e
Makuliro..... 9·35s 37·26e
Makumbako..... 8·51s 34·50e
Makuyuni..... 3·33s 36·06e
Malagarasi..... 5·06s 30·50e
Malangali..... 8·34s 34·51e
Malinyi..... 8·56s 36·08e
Malolo..... 7·18s 36·35e
Mambali..... 4·33s 32·41e
Manda..... 7·58s 32·26e
Manda..... 8·30s 32·44e
Manda.....10·28s 34·35e
Mantare..... 2·43s 33·13e
Manyoni..... 5·45s 34·50e
Mapinga..... 6·36s 39·04e
Marungu..... 3·44s 30·48e
Masasi. .....10·43s 38·48e
Matombo..... 7·03s 37·46e
Maurui..... 5·07s 38·23e
Mbalizi..... 8·56s 33·22e
Mbamba Bay....11·17s 34·46e
Mbate..... 8·52s 39·10e
Mbemba.....10·03s 38·36e
Mbeya..... 8·54s 33·27e
Mbinga.....10·56s 35·01e
Mbirira..... 4·21s 30·10e
Mbogo..... 7·26s 33·26e
Mbulu..... 3·51s 35·32e
Mchinga..... 9·44s 39·42e
Mchungo..... 7·42s 39·17e
Mdandu..... 9·09s 34·42e
Meia Meia..... 5·49s 35·48e
Mgeta..... 8·19s 36·08e
Mikese..... 6·46s 37·54e
Mikindani.....10·17s 40·07e
Mikumi..... 7·24s 36·59e
Milumba..... 7·06s 31·04e
Mingoyo.....10·06s 39·38e
Mkalama..... 4·07s 34·38e
Mkata..... 5·47s 38·17e
Mkoani..... 5·22s 39·39e
Mkokotoni..... 5·52s 39·15e
Mkulwe..... 8·35s 32·19e
Mkwaja..... 5·47s 38·51e
Mkwaya.....10·06s 39·40e
Mnazi..... 8·54s 39·06e
Mohoro..... 8·08s 39·10e
Mombo..... 4·53s 38·17e
Mondo..... 4·59s 35·54e
Morogoro..... 6·49s 37·40e
Moshi..... 3·21s 37·20e
Mpanda..... 6·22s 31·02e
Mpui..... 8·21s 31·50e
Mpwapwa..... 6·21s 36·29e
Mrijo..... 5·10s 36·15e
Msagali..... 6·21s 36·18e
Msata..... 6·20s 38·23e
Mtakuja..... 7·22s 30·37e
Mtama.....10·18s 39·22e
Mtowabaga..... 2·30s 35·53e
Mtwara.....10·16s 40·11e
Mtyangimbori...10·16s 35·31e
Mugombazi..... 5·50s 30·14e
Muheza..... 5·10s 38·47e
Muhutwe..... 1·33s 31·42e
Muleba..... 1·49s 31·40e
Musasa..... 3·21s 31·33e
Musoma..... 1·30s 33·48e
Mvomero..... 6·20s 37·25e
Mvuha..... 7·12s 37·51e
Mwadui..... 3·33s 33·36e
Mwanza..... 2·31s 32·54e
Mwaya..... 8·55s 36·50e
Mwaya..... 9·33s 33·57e
Mwimbi..... 8·39s 31·40e
Mwitikira..... 6·31s 35·39e
Mzenga..... 6·56s 38·43e
Mziha..... 5·54s 37·47e
Mzizima..... 5·13s 39·02e
Naberera..... 4·12s 36·56e
Nachingwea.....10·23s 38·46e
Nagaga.....10·54s 39·07e
Namanga..... 2·33s 36·46e
Namanyere..... 7·31s 31·03e
Nanjirinji..... 9·39s 39·04e
Nansio..... 2·08s 33·03e
Ndala..... 4·46s 33·16e
Nduguti..... 4·18s 34·42e
Ndumbwe.....10·14s 39·58e
Newala.....10·56s 39·18e
Ngara..... 2·28s 30·39e
Ngarimbi..... 8·28s 38·36e
Ngasamo..... 2·33s 33·53e
Ngerengere..... 6·45s 38·07e
Ngomba..... 8·23s 32·53e
Ngoywa..... 5·56s 32·48e
Ngunga..... 3·41s 33·34e
Njinjo..... 8·48s 38·54e
Njombe..... 9·20s 34·46e
Nkonko..... 6·20s 34·58e
Nondwa..... 6·26s 35·20e
Ntumba..... 8·20s 32·05e
Nungwe..... 2·46s 32·01e
Nyahanga..... 2·23s 33·33e
Nyahua..... 5·24s 33·19e
Nyakabindi..... 2·38s 33·59e
Nyakakiri..... 2·15s 31·28e
Nyakanazi..... 3·00s 31·15e
Nyamougo..... 1·29s 34·33e
Nyamtumbo.....10·29s 36·02e
Nyamwage..... 8·08s 39·00e
Nyandekwa..... 3·55s 32·30e
Nyilumba.....10·29s 40·20e
Nzega..... 4·13s 33·11e
Nzima..... 3·03s 32·48e
Nzubuka..... 4·45s 32·50e
Oldeani..... 3·21s 35·33e
Pangani..... 5·26s 38·58e
Puge..... 4·55s 33·07e
Pugu..... 6·54s 39·07e
Ruanda.....10·33s 34·57e
Rudewa.....10·06s 34·39e
Runazi..... 2·47s 31·28e
Runere..... 3·06s 33·16e
Rungwa..... 6·57s 33·31e
Rungwa..... 7·21s 31·40e
Rungwe..... 9·10s 33·36e
Ruponda.....10·15s 38·42e
Ruvu..... 6·48s 38·39e
Sadani..... 6·03s 38·47e
Salawe..... 3·19s 32·52e
Same..... 4·04s 37·44e
Sao Hill..... 8·20s 35·12e
Saranda..... 5·43s 34·59e
Seke..... 3·20s 33·31e
Sekenke..... 4·16s 34·10e
Shanwa..... 3·10s 33·46e
Shinyanga..... 3·40s 33·26e
Shirati..... 1·08s 33·59e
Sikonge..... 5·38s 32·46e
Simba Sirori..... 1·44s 34·13e
Simbo..... 4·40s 33·27e
Simbo..... 4·53s 29·44e
Singida..... 4·49s 34·45e
Sitalike..... 6·38s 31·08e
Somanga..... 8·24s 39·17e
Songea.....10·41s 35·39e
Songwe..... 9·42s 33·56e
Sudi.....10·06s 39·57e
Suguti..... 1·44s 33·39e
Sumbawanga.... 7·58s 31·37e
Tabora..... 5·01s 32·48e
Tandala..... 9·23s 34·14e
Tanga..... 5·04s 39·06e
Taveta..... 9·01s 35·37e
Tukuyu..... 9·15s 33·39e
Tunduru.....11·07s 37·21e
Tutubu..... 5·30s 32·41e
Ujiji..... 4·55s 29·41e
Ulaya..... 7·04s 36·54e
Urambo..... 5·04s 32·03e
Uruwira..... 6·27s 31·21e
Ushashi..... 2·00s 33·57e
Ushetu..... 4·10s 32·16e
Usoke..... 5·06s 32·30e
Ussure..... 4·39s 34·23e
Utegi..... 1·20s 34·35e
Utengule..... 8·57s 35·50e
Utete..... 7·59s 38·47e
Uvinza..... 5·06s 30·22e
Wete..... 5·04s 39·43e
Zanzibar..... 6·10s 39·11e
Zinga Mulike.... 9·09s 38·44e

Physical features and points of interest

Amboseli, Lake.. 2·37s 37·08e
Arusha National Park..... 3·05s 36·42e
Bahi Swamp..... 6·05s 35·10e
Bubu, *river*..... 6·03s 35·19e
Bumbire Island..... 1·40s 31·53e
Burigi, Lake..... 2·05s 31·15e
Eyasi, Lake..... 3·40s 35·05e
Gelai..... 2·33s 36·05e
Gombe, *river*..... 4·38s 31·40e
Gombe Stream National Park... 4·42s 29·37e
Great Ruaha, *river*..... 7·56s 37·52e
Hanang, *mountain*..... 4·26s 35·24e
Ikimba, Lake..... 1·28s 31·30e
Jaeger Summit, *peak*..... 2·52s 35·47e
Kagera, *river*..... 0·57s 31·47e
Kavuu, *river*..... 7·40s 31·46e
Kilimanjaro, *mountain*..... 3·04s 37·22e
Kilombero, *river*.. 8·31s 37·22e
Kipengere Range..... 9·10s 34·15e
Kisigo, *river*..... 7·03s 35·50e
Kitangiri, Lake... 4·05s 34·19e
Kitumbeine, *volcano*..... 2·44s 36·16e
Lake Manyara National Park... 3·30s 36·25e
Livingstone Mountains..... 9·45s 34·20e
Loolmalassin, *volcano*..... 3·03s 35·49e
Lugufu, *river*..... 5·10s 29·46e
Lugunga, mountain..... 6·47s 36·19e
Luhombero, *river*..... 8·24s 37·12e
Lukose, *river*..... 7·28s 36·31e
Luwegu, *river*..... 8·31s 37·23e
Mafia Island..... 7·50s 39·50e
Mahali Mountains..... 6·12s 29·50e
Malagarasi, *river*..... 5·12s 29·47e
Malagarasi Swamp..... 5·00s 31·10e
Manonga, *river*... 4·08s 34·12e
Manyara, Lake... 3·35s 35·50e
Mara, *river*..... 1·31s 33·56e
Masai Steppe.... 4·45s 37·00e
Masangwe, mountain..... 5·28s 30·05e
Matandu, *river*.. 8·45s 39·19e
Mbarangandu, *river*..... 8·57s 37·24e
Mbeya Range, *mountain range*..... 8·50s 33·20e
Mbwemburu, *river*..... 9·29s 39·39e
Mikumi National Park..... 7·12s 37·05e


TANZANIA
Size of symbol indicates relative size of town
Elevations in metres
Rand McNally & Co.
A-565100-257
0 50 100 150 200 km
0 50 100 150 mi
Nairobi
Kigali
Bujumbura
Bukavu
Mwanza
Arusha
Moshi
Mombasa
Tanga
Dodoma
Dar-es-Salaam
Morogoro
Zanzibar
Tabora
Kalemi (Albertville)
Sumbawanga
Iringa
Mbeya
Njombe
Songea
Lake Victoria
Lake Tanganyika
Lake Nyasa
Lake Rukwa
Lake Kivu
Lake Eyasi
Lake Manyara
Lake Natron
Lake Amboseli
Lake Rukwa
Lake Mweru
Lake Bangweulu
Serengeti National Park
Ngorongoro Crater
Kilimanjaro
Mount Meru
Ruaha National Park
Selous Game Reserve
Mikumi Nat. Park
Arusha Nat. Park
Lake Manyara National Park
Tarangire Nat. Park
Gombe Stream Nat. Pk.
Rungwa River Game Reserve
Kilombero Game Controlled Area
Usangu Flats
Masai Steppe
Mahali Mts.
Mlala Hills
Ufipa Plateau
Rungwe
Kyela
Pemba
Zanzibar
Mafia Island
Uganda
Kenya
Tanzania
Zaire
Zambia
Malawi
Mozambique
Somalia
Rwanda
Burundi
Indian Ocean

MAP INDEX (continued)

Mkomazi Game Reserve, *wildlife refuge*... 4·10s 38·10e
Mlala Hills..... 6·47s 31·45e
Momba, *river*..... 8·28s 32·40e
Meru, Mount, *mountain*..... 3·14s 36·45e
Muhuwesi, *river*..11·16s 37·58e
Natron, Lake..... 2·25s 36·00e
Ngomba, *mountain*..... 5·43s 35·52e
Ngono, *river*..... 1·08s 31·35e
Ngorongoro Crater..... 3·10s 35·35e
Nguru Mountains..... 6·00s 37·30e
Njombe, *river*.... 6·56s 35·06e
Nyasa, Lake.....10·50s 34·30e
Nyumba ya Mungu, *dam*... 3·40s 37·30e
Ol Doinyo Lengai, *volcano*. 2·45s 35·54e
Olduvai Gorge... 2·58s 35·22e
Pangani Falls, *waterfalls*..... 5·21s 38·39e
Pangani, *river*.... 5·26s 38·58e
Pare Mountains. 4·00s 37·45e
Pemba Channel.. 5·10s 39·20e
Pemba Island.... 7·31s 39·25e
Poroto Mountains..... 9·00s 33·45e
Ruaha National Park..... 7·32s 34·40e
Rubeho Mountains..... 6·55s 36·30e
Rubondo Island.. 2·20s 31·52e
Rufiji, *river*..... 8·00s 39·20e
Ruhudji, *river*.... 8·52s 36·01e
Ruhuhu, *river*....10·31s 34·34e
Rukwa, Lake..... 8·00s 32·25e
Rungwa, *river*.... 7·36s 31·50e
Rungwa River Game Reserve, *wildlife refuge*... 6·54s 34·10e
Rungwe Range, *mountain range*. 9·10s 33·50e
Ruvu, *river*..... 6·23s 38·52e
Ruvuma, *river*....10·29s 40·28e
Selous Game Reserve, *wildlife refuge*... 9·10s 37·10e
Serengeti National Park... 2·20s 34·50e
Serengeti Plain.. 2·50s 35·00e
Sibiti, *river*..... 3·49s 34·46e
Siguri Falls, * *waterfalls*..... 8·31s 37·23e
South Mafia Channel..... 8·10s 39·40e
Speke Gulf..... 2·20s 33·15e
Tanganyika, Lake 7·00s 30·00e
Tarangire National Park... 4·00s 36·00e
Ufipa Plateau ... 8·00s 31·30e
Ugalla, *river*..... 5·08s 30·42e
Ukara Island..... 1·50s 33·03e
Ukerewe Island.. 2·03s 33·00e
Ukuguru Range, *mountain range*. 6·20s 37·10e
Uluguru Mountains..... 7·10s 37·40e
Usambara Mountains..... 4·45s 38·30e
Usangu Flats, Plain..... 8·30s 34·15e
Victoria, Lake.... 1·30s 33·00e
Wami, *river*..... 6·08s 38·49e
Wembere, *river*.. 4·10s 34·11e
Zanzibar, *island*.. 6·10s 39·20e
Zanzibar Channel 6·00s 39·00e

its place of origin) both exist. Soils in a repetitive sequence that is determined by relief and drainage are known as catenas (connected series). In the humid and highland areas, brown volcanic soils and ferrisols (iron-bearing soils) located on crystalline rocks are the most productive soils. Areas receiving annual rainfall in excess of 53 inches (such as Bukoba and Tukuyu) have productive yellow-brown and yellow-red soils of low fertility. In the north and northeast plateau country, soils are fertile but lose much moisture because of a high rate of evaporation and transpiration from plants. Over the Central Plateau and in the south podsolic soils (infertile, acidic forest soils) occur on sandstone and crystalline rock. The fertile soils of the river basins require management, such as drainage. Scattered over many parts are rocky outcrops, debris, and stony shallow soils.

Climate. Tanzania experiences an equatorial type of climate, generally modified by altitude. Seasonal changes are more marked in rainfall than in temperature.

A low-pressure zone, apparently accompanying the sun's annual movement, draws two major airstreams into the region. Between March and September, moisture-laden southeast airstreams sweep across the Indian Ocean before moving over the mainland. At other seasons northeast airstreams, originating in Asia, are drier.

The variation in the mean monthly temperature is small, amounting to 4° F (2° C) at Mwanza on the southern shore of Lake Victoria, to 7° F (4° C) at Dar es Salaam on the coast and to 9° F (5° C) at Dodoma in central Tanzania; diurnal variations can be great. Temperatures vary considerably with altitude. Daily maximums near the coast exceed 90° F (32° C); on the highlands, minimum temperatures may fall below 68° F (20° C).

The seasonal rainfall pattern also varies considerably in different regions. The two rainy seasons in the north (March to May and October to December) contrast with the one in the south (December to April). Rainfall totals depend on aspect and altitude. Places on the western side of Lake Victoria receive more rainfall than those on the eastern side. The windward and leeward sides of the Uluguru Mountains receive rainfall of about 101 inches and between 23 and 31 inches, respectively. Areas bordering the lakes and the coast receive more rainfall than the Central Plateau.

Vegetation. Woodland, called miombo, extends over nearly 30 percent of Tanzania. Typically, the ground is covered with grass, from which rise trees of the *acacia*, *Brachystegia* (a genus of trees of the pea family having 30 tropical African species), and *Isoberlinia* (a genus of trees of the pea family having six tropical African species) groups. Similarly, wooded grassland, of which less than 50 percent has tree cover, presents a parklike appearance. Several types of forest are found in small, discontinuous stretches. Above the forest line are intermontane grasslands, and even alpine deserts. Semidesert vegetation is found on the leeward side of the Pare, Usambara, and Kilimanjaro mountains. In other areas, reed and grass swamps contrast with the thick mangrove forests of the coastal creeks. The salt pans around Lake Natron are barren.

Animal life. The animal life of Tanzania is unique in abundance and variety; it includes the five big game animals—elephant, rhinoceros, buffalo, lion, and leopard. The plains teem with hoofed animals such as the zebra, the giraffe, and various species of antelope—the gazelle, the eland, the kudu, the dik-dik, and the wildebeeste. Less common are the chimpanzee and such species of antelope as the topi, the oryx, the puku, and the oribi. The carnivores include the leopard, various cats, the hyena, and the wild dog. Hippopotamuses and crocodiles inhabit riverbanks and lakeshores. Nearly 1,500 varieties of birds, including the ostrich, are common, as are many species of snakes, lizards, and insects.

Big game animals

Traditional regions. Tanzania can be divided into four general regions—the coastal area, the highlands, the lake region, and the Central Plateau.

The coastal area is climatically hot and humid, and, as already mentioned, is low-lying and narrow, except where rivers enter it from the west. The cultural and economic mode of life is distinctive due to overseas contacts, Islāmic influence, and the Swahili language.

The highlands, which are cooler and which receive adequate rainfall, are generally densely populated and can be subdivided into northern and southern regions. In the north, the Chagga, Arusha, Pare, and related peoples have been exposed to external influences; producing both cash and subsistence crops, they have progressed economically and socially. In contrast, in the south, the Nyakyusa and Ngoni peoples have suffered from the inaccessibility of their area; their economy remains traditional.

The Lake Victoria region is a zone of contrasts and is generally thickly peopled. In the west-northwest, the Haya people now have to migrate because the production of coffee and bananas has fallen. The Sukuma and Nyamwezi peoples, however, are well established, raising cattle, and growing cotton and food crops.

In the dry, monotonous Central Plateau, independent pastoralists, best exemplified by the Masai and related groups, contrast with the Gogo, who eke out a harsh existence based on cattle and grains.

The landscape under human settlement. *Rural settlement.* The majority of people live in dispersed farmsteads. In the fishing and farming economy of the coastal area, villages with irregular streets consisting of mud and pole houses and their compounds are set amidst coconut palms and fields planted with cashews, cassavas, and other crops. Individual homesteads are densely scattered in the Kilimanjaro and Meru area. The lush vegetation of banana and coffee trees shelters a lower tier of food crops. Maize is grown in the plains, and cattle are stall fed. Cylindrically shaped huts, surmounted with cone-shaped roofs, are being replaced by modern houses. The Sukuma have a village area within which houses may be dispersed, each homestead having at its disposal a range of different soils. The homestead compound has concentric encircling hedges around the garden, the houses (built for the head of the homestead and his wives), the grain stores, and the cattle pen. In the central and drier parts of the north, the houses are thinly distributed and are semi-permanent. The Gogo have squat, rectangular, flat-roofed houses, grouped to form a courtyard in which cattle are kept. Most traditional forms of settlements are rapidly undergoing change; the house form is becoming uniformly rectangular.

Urban settlement. Tribal and traditional constraints are eased in the towns. Dar es Salaam (which in 1970 had

Dar es Salaam

a population of 350,000) is the largest, most picturesque, and most cosmopolitan of the towns of Tanzania. The business section is dominated by Asians. Disparity between the rich and the poor is still apparent but is less along racial lines than in the past. Suburbs for the low-income group are characterized by overcrowded six-room "Swahili" houses. The bungalows of the rich are dispersed in scenic locations. The port of Tanga, with a population of 61,000, is located opposite the island of Pemba. It is similar to Dar es Salaam but retains many more buildings of the German period; its fortunes are tied to sisal. The old towns of Bagamoyo, Pangani, and Kilwa, with ancient buildings and narrow lanes, are falling into decay. Arusha, with a population of 32,000, is a new town; it is the capital of the East African Community and the centre of some of the most famous game parks in East Africa. Mwanza (population 35,000), Kigoma (21,000), Musoma (15,000), and Bukoba (8,000) are all lake ports. Other important towns are Moshi (27,000), Morogoro (25,000), Dodoma (24,000), Iringa (22,000), Tabora (21,000), and Mtwara (20,000).

Zanzibar and Pemba. *Relief.* Low-lying Pemba, whose highest point reaches an altitude of 311 feet, and Zanzibar, which reaches 390 feet, are islands whose structure consists of coralline rocks. The west and northwest of Zanzibar consist of several ridges rising above 200 feet, but nearly two-thirds of the south and east are low-lying. Pemba appears hilly because the level central ridge has been gullied and eroded by streams draining into numerous creeks. On Zanzibar Island short streams drain mostly to the north and west. The few streams in the east disappear into the porous coralline rock.

Soils. Among the ten types of soils recognized in Zanzibar are fertile sandy loams and deep-red earths, which occur on high ground; on valley bottoms, less fertile gray and yellow sandy soils are found. The eight soil types in Pemba include brown loams; pockets of infertile sands are found on the plains.

Climate. Zanzibar and Pemba have rainfall of 60 inches and 80 inches, respectively. The rainfall is highest in April and May, lowest in November and December. Humidity is high. The average temperature is 81° F (27° C) in Zanzibar and 79° F (26° C) in Pemba; the annual temperature ranges are small.

Vegetation. Long human occupation has resulted in the clearance of most of the forests, which have been replaced with coconuts, cloves, bananas, citrus, and other crops. On the eastern side of the islands, especially on Zanzibar, there is bush (scrub).

Animal life. Although there is some difference between the animal life of the two islands, it is generally similar to that on the mainland. Animal life common to both islands includes monkeys, civet cats, and mongooses. More than 100 species of birds have been recorded in Zanzibar.

Rural settlements. The rural settlements—and the life in them—changed drastically after the nationalization of land, which was followed by agricultural reforms. By 1970 large plantations of cloves and coconuts, once almost exclusively the property of fewer than 50 Arab families, had been redistributed in three-acre units to about 24,000 peasant families. The production of food crops, especially rice, is being encouraged. Fishing villages are still important in the east.

Urban settlements. The city of Zanzibar, with a population of 68,000, is still primarily a Muslim town, although the distinctive mode of life and culture, reminiscent of an Eastern commercial centre, has almost disappeared since the downfall of the Arab oligarchy in 1964. The hub of civic life is moving from Stone Town with its narrow lanes to a new town with modern buildings and amenities that is being built at Ngambo, the former African quarter. Kilimani, Bambi, and Chaani are being developed into new rural towns; a similar change is taking place in Pemba at Mkoani, Chake Chake, and Wete.

Zanzibar city

PEOPLE AND POPULATION

Linguistic groups. For official purposes both Swahili and English are commonly used. Swahili originated between Kilwa, on what is now the southern coast of Tanzania, and Lamu, on what is now the northern coast of Kenya. During the early 19th century it gradually spread westward as a language of trade to become the lingua franca of Tanzania. The language has a Bantu base, and its vocabulary has been enlarged with additions from Arabic, English, Hindi, and Portuguese.

Four of the main language groups in Africa are represented in Tanzania. More than 95 percent of the people speak Bantu languages, of which about 100 are distinguishable. The chief Bantu languages are Kisukuma, Kinyamwezi, Kimakonde, Kihaya, Kichagga, Kigogo, Kiha, Kihehe, Kinyakyusa, and Kiluguru. In north and north central Tanzania are three other main language families. Nilotic languages are spoken by the Masai, Tatoga, and Luo peoples. The languages spoken by Iraqw, Gorowa, Alawa, Ngomvia, and Mbugu (in the Usambaras) belong to the Cushitic grouping. Only the Sandawe and the Hadza speak the Khoisan, or Click language, common among the Bushmen and Hottentots of southern Africa. Arabic is widespread on the coast. The main languages spoken by the Asians are Gujarati, Hindi, Punjabi, and Urdu, whereas the Goanese speak Konkani.

Ethnic groups. Tanzania is a multitribal state in which historically, culturally, and politically no single group has dominated. During the Stone Age, bands of hunter-gatherers of the Bushmen type inhabited parts of the country. The Sandawe are vestiges of this early group. About 1,000 BC southern Cushites, represented by the Burungi, Gorowa, Iraqw, and Mbugu peoples, began to expand from Ethiopia into the highlands and Rift Valley of what is now Tanzania. Early in the first and second millennium, Bantu speakers from the west or southwest penetrated into the country. As they were more numerous and were also agriculturalists familiar with iron tools, they gradually absorbed the hunters and gatherers or forced them to retreat to less desirable areas. As each group settled, customs, social organization, and languages became distinct. Today only the Sukuma number over 1,000,000. The Nyamwezi, Hehet Nyakyusa, Haya Chaga (Chagga), Gogo, and Ha each number over 300,000. The Ngoni, Nyakyusa, and Yao groups represent the Bantu groups from the south that have retreated northward. The Nilotics, chiefly consisting of the Luo, Tatoga, and the Massai, form a distinctive group.

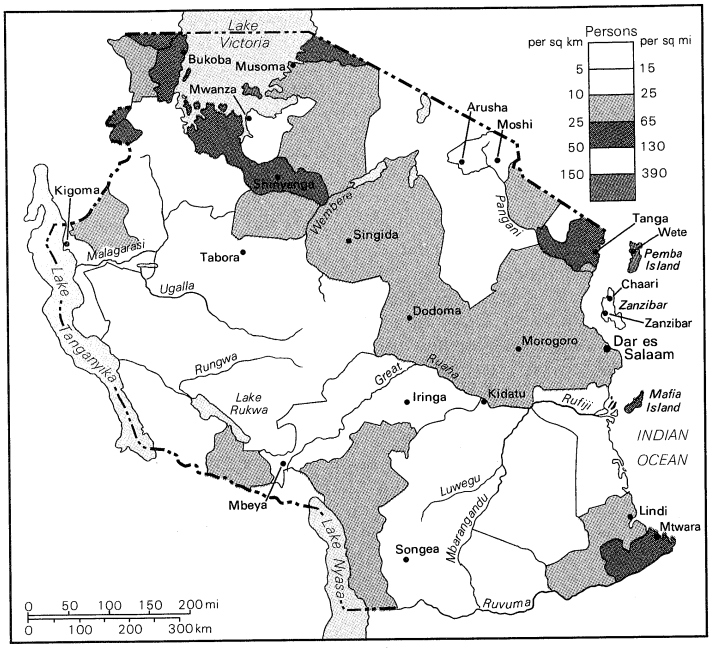
There are also Asian and European minorities. The Asians are a very diverse group, consisting of Ismāʿīlīs, Bohrās, Sikhs, Punjabis, Jains, and Goanese. The largest group of Europeans are English speaking. Others include Germans, Scandinavians, Italians, and Greeks. The Arabs also form a distinct minority group.

Religious groups. The number of people professing traditional (mostly animistic) beliefs is gradually declining. During the last population census, 31 percent claimed to be Christians of various denominations. Muslims (31 percent) are most widespread in the coastal areas, where they form 63 percent of the urban population. Among adherents of Islām, both the Sunnī and Shīʿah sects are represented. Among Asian communities, the majority of Muslims are Ismāʿīlī Khōjās, who recognize the Aga Khan as their leader. Other Asians adhere to Hinduism, Jainism, and Buddhism, while the Goanese are Roman Catholic, and the small Parsi community practices Zoroastrianism.

Demography. *Birth rates and death rates.* At the 1967 census the population of mainland Tanzania was about 11,960,000, representing an increase of 3,000,000, or 35 percent, since 1957. Between 1957 and 1967 mortality dropped from an estimated 24 to 25 per thousand to 22, while the birth rate increased from 45 to 46 to 47 per thousand. Life expectancy, which was estimated to be between 35 and 40 years, is currently between 40 and 41 years. The population growth is close to 3 percent a year.

Population growth

Population characteristics. One of the main demographic characteristics of Tanzania is the youthfulness of its population; 44 percent are less than 15 years old, and only 3 percent are more than 75 years old. Another characteristic is the sex ratio of 95 males to 100 females. Regional variations however, are striking. In the Tanga,



Population density of Tanzania.

Pangani, and Mzizima districts the ratios are about 120 males to 100 females, but in the Kibondo, Kasulu, and Njombe districts they are about 80 males to 100 females.

Emigration and immigration. Temporary emigration from southern Tanzania to Central Africa and the mines in the Republic of South Africa has now ceased. Internal migration is, however, still widespread because of inequalities in economic opportunities. Most of the southern regions of the country experience a relatively high degree of emigration.

The immigration of large numbers of Asians, including Arabs, and of Europeans, which was a feature of the colonial period has, in the case of the Asians, been reversed. International immigrants are mainly refugees from southern Africa, of whom 75 percent come from Mozambique alone. According to the 1967 census, during the preceding decade the Coast Region—especially the city of Dar es Salaam—attracted the most people (about 114,000), followed by Tabora (70,000), Arusha (59,000) and Tanga (49,000), while Kigoma, Iringa, Singida, and Zanzibar had the least number of immigrants.

Population distribution. Most of the well-populated areas of Tanzania are located on the periphery of the country—around the shores of Lake Victoria, in the highlands, and in other well-watered areas. The population is very sparse in the central and western parts, partly because of infestation by the deadly tsetse fly and—in the north and central parts—because of very dry conditions. Only about 6 percent of the people live in urban areas.

In 1967 the population density on the mainland was 34.7 per square mile, compared with 25.6 in 1957. Two of the most densely peopled regions are Mwanza and Kilimanjaro, which have densities of 140 and 127 per square mile, respectively. At the other extreme, the Ruvuma and Tabora regions have less than 18 persons per square mile. Local variations in densities are considerable. In the Masai, Monduli, Nachingwea, and Mpanda districts, densities are as low as 5 persons per square mile, in contrast with 442.1 in the Ukerewe District.

Demographic trends. With better child care and improvements in health facilities, mortality is expected to continue to decline, life expectancy to increase, and the rate of population increase to accelerate. Sex ratios are likely to normalize. With the emphasis on *ujamaa* (commune) villages, and with the establishment of better links with the south, population is expected to become more clustered and the pattern of its distribution to alter.

Zanzibar: people and population. *Languages.* Swahili is the principal language in Zanzibar and Pemba. The classical dialect is Kiunguja. Arabic is also important because of long-established Islāmic tradition, past Arab overlordship, and the presence of a large Arab-speaking minority. Three Bantu languages are spoken on the islands—Kipemba on Pemba, Kitumbatu in southern Pemba and northeastern Zanzibar, and Kihadimu mostly in southern and eastern Zanzibar. There are several dialects surviving in local areas. Among the Asian communities, the chief languages are Gujarati, Kutchi, and Hindustani. English, taught in schools, is widely used.

Tanzania, Area and Population

	area*		population	
	sq mi	sq km	1967 census	1972 estimate
Tanganyika				
Regions				
Arusha	31,700	82,103	610,000	720,000
	32,654	84,574		
Coast	13,020	33,722	784,000	949,000
	13,020	33,722		
Dodoma	15,950	41,311	709,000	792,000
	15,950	41,311		
Iringa	21,950	56,851	690,000	802,000
	22,750	58,923		
Kigoma	14,300	37,037	473,000	508,000
	17,400	45,066		
Kilimanjaro	5,100	13,209	653,000	766,000
	5,100	13,209		
Mara	8,400	21,756	544,000	644,000
	11,400	29,526		
Mbeya	32,100	83,139	969,000	1,112,000
	34,800	90,132		
Morogoro	28,200	73,038	685,000	754,000
	28,200	73,038		
Mtwara	31,950	82,751	1,041,000	1,167,000
	31,950	82,751		
Mwanza	7,600	19,684	1,056,000	1,200,000
	13,850	35,872		
Ruvuma	23,650	61,254	393,000	449,000
	24,800	64,233		
Shinyanga	19,600	50,764	899,000	1,018,000
	19,600	50,764		
Singida	19,050	49,340	458,000	489,000
	19,050	49,340		
Tabora	47,100	121,989	563,000	662,000
	47,750	123,673		
Tanga	10,350	26,807	771,000	877,000
	10,350	26,807		
West Lake	11,100	28,749	659,000	725,000
	15,300	39,627		
Zanzibar				
Regions				
Pemba	380	984	164,000	§
	380	984		
Zanzibar Mjini (urban)	93	241	95,000	§
	93	241		
Zanzibar Shambani (rural) North	186	482	56,000	§
	186	482		
Zanzibar Shambani (rural) South	358	928	39,000	§
	358	928		
Total Tanzania	342,140†	886,139	12,313,000‡	14,028,000
	364,945†	945,203		

*Where two figures are given, the first is the land area, the second the total area. †Converted area figures do not add to total given because of rounding. ‡Figures do not add to total given because of rounding. §The total mid-1972 population estimate for Zanzibar and Pemba was 394,000.
Source: Official government figures.

Ethnic groups. There are several groups of Africans, nearly a third of whom are recent arrivals from the mainland. Indigenous Bantu groups, consisting of the Pemba in Pemba and the Hadimu and Tumbatu in Zanzibar, have absorbed the settlers who came from Persia in the 10th century. These groups, and some of the descendants of slaves, call themselves Shirazi. The Afro-Shirazi group comprises the majority of people on the islands. There are also small enclaves of Comorians and Somalis. Arab settlements were also established early, and intermarriage with the local people took place. Later Arab arrivals came from Oman and constituted an elite. The poorer recent immigrants from Oman are known as Manga. The genealogy of the Afro-Arab groups is blurred. The Asians, now forming a very small minority, may be divided into Muslim and non-Muslim groups.

The Afro-Shirazi group

Religion. Almost the whole of the Arab and the African peoples of Zanzibar—about 96 percent—profess the Islāmic faith. Traditional African beliefs are also found existing in conjunction with Islām. Among Muslims, the Sunnī sect is preferred by the indigenous people. The

small Ibāḍī sect was once important because it was adhered to by members of the Arab elite. Other Islāmic groups are confined to the Asians, and include the Ismāʿīlīs, Ithnā ʿAsharīs, Sunnīs, and Bohrās. The non-Muslims, who are in a minority, are mainly Hindus, although the few Goanese are all Catholic. African converts to Christianity are few.

Demography. The population of Zanzibar and Pemba in 1967 was about 354,000. The total growth during the decade 1957 to 1967 was 18 percent. Zanzibar Island, with 190,000 inhabitants, was the more populous, while Pemba, with 427 persons per square mile was more densely peopled. With an average of 347 persons per square mile, the overall density of population is high.

The ratio of males per 100 females has declined from 111 in 1957 to 102 in 1967. In the Magharib rural area of Zanzibar, however, it continues to remain high, at 124. Also, in contrast to the mainland, a comparatively high proportion—23 percent—of the population is urbanized. It is estimated that between 1957 and 1967 the crude death rate of from 20 to 21 per thousand changed little; life expectancy increased from between 42 and 43 to between 43 and 44 years, and infant mortality declined from 160 to 140 per thousand.

THE NATIONAL ECONOMY

If international comparisons are made, Tanzania is a poor country. Per capita gross domestic product was $65 U.S. in the 1960s. (The monetary unit is the Tanzanian shilling [TSh.] with TShs. 7.14 equalling $1 U.S., and TShs. 18.615 equalling £1 sterling as of December 1970). The country is noted for its primary agricultural products, being one of the world's foremost suppliers of sisal, cashew nuts, and cloves, and—to a lesser extent—cotton, pyrethrum, and tobacco. The country is large and its resource base is wide. Regional cooperation through a common market with Kenya and Uganda has aided commerce. Trade and communication links have been established with other African countries.

Importance of agriculture to the economy

Resources. *Mineral resources.* Diamonds and salt are important. Prospecting for other minerals is continuing. The Nyanzian rocks south and east of Lake Victoria are gold bearing; to the west, tin and tungsten deposits have been located. Mica is found in central and eastern Tanzania. In the south, deposits of lead, copper, silver, and gold have been found. In the southwest, large quantities of coal and iron have been proved. In several of the volcanic pipes (eruptive passages of volcanoes) diamonds are to be found. Deposits of limestone, gypsum, salt, kaolin (china clay), and rock phosphate occur in the coastal and lake regions.

Biological resources. Tanzania has substantial biological resources. There are a great variety of grasslands that are not fully utilized because of infestation by the tsetse fly, which is deadly to cattle. The forests contain tropical hardwoods, such as mahogany and teak; softwoods are found in mountain areas. Great concentrations of animal life still survive in several parts of the country, especially in the north. The exploitation of coastal fisheries has barely begun. Several lakes—including Lake Victoria—and the rivers are important sources of fish.

Power resources. The ample energy resources of Tanzania include quantities of charcoal and firewood, large resources of coal in the south (amounting to more than 304,000,000 tons), and adequate waterpower.

Sources of national income. *Agriculture, forestry, and fishing.* Tanzania is primarily an agricultural country, with nearly 90 percent of the people gaining a living from the land. In 1970 the contribution of agriculture, hunting, forestry, and fishing to the gross domestic product was over 39 percent. The major crops produced in 1969 were sisal, cashew nuts, sugar, lint cotton, coffee, wheat, and tea. Tobacco and pyrethrum are also produced. Substantial quantities of staple grains, fruits, and vegetables are produced for local consumption.

Annual fish production was estimated to be 140,000 tons; the fish were mainly caught in lakes and rivers, especially Lake Victoria.

Only a fraction of the forest production is recorded. Forests with indigenous trees, mostly hardwoods, cover 37,000,000 acres, but softwood production is increasing. Among the main timbers being exploited are *Podocarpus* (a genus of evergreen trees with a soft, pulpy fruit), camphorwood, mahogany, and mangrove.

Forestry

There are close to 11,000,000 head of cattle, and there is an export trade in meat and meat products, hides, and skins. Many goats, sheep, and poultry are raised.

Mining and quarrying. Mineral exports include diamonds, of which those from Mwadui have proved the most valuable since the 1940s. Gold production is declining, whereas the production of salt from Uvinza near Lake Tanganyika and the coastal area is increasing. Gemstones, including the rare Tanzanite, are important. Around Dar es Salaam quarrying is undertaken on a large scale. With the development of transport and industry, it is anticipated that the large quantities of coal (284,000,000 tons at Ruhuhu and 20,000,000 tons at Songwe), iron (78,000,000 tons at Liganga), phosphate, kaolin (2,000,000 tons), and salt will be utilized.

Manufacturing. The contribution of the manufacturing sector to the gross domestic product in 1970 was 8 percent. While textile production is increasing, food processing still remains the largest manufacturing industry. Ninety-five percent of the manufactured goods were produced for local consumption. The main manufacturing centres are Dar es Salaam (food, brewing, textiles, shoes, petroleum refining, rubber cement, and motor assembly), Tanga (twine, furniture, food processing, and fertilizers) Arusha and Moshi (food and drink, textiles and furniture) and Mwanza (food and drink and textiles).

Energy. The largest installed hydroelectric complex is on the Pangani River. The Hale and Pangani Falls generating stations and two thermal stations at Dar es Salaam provide power for the coastal grid supplying Dar es Salaam, Tanga, and Morogoro. A new station at Kidatu, which will ultimately have a capacity of 200 megawatts, is to be completed by 1975. One of the forerunners of multi-purpose dams is the Nyumba ya Mungu on the Pangani River. Thirty towns have electricity for domestic and industrial use.

Hydroelectric plants

The capacity of the petroleum refinery at Dar es Salaam is 750,000 tons. Large quantities of petroleum products are exported to Zambia through the Tanzama Pipeline, which is 1,080 miles long.

Financial services. The Bank of Tanzania replaced the East African Currency Board in 1966. Its main functions are the control of banking and credit, issuance of currency, management of foreign exchange, and those of a government bank and fiscal agency. The National Bank of Commerce replaced the 12 nationalized banks in 1967. It maintains numerous branches and rural agencies.

Other financial services are provided by the National Insurance Company, the Permanent Housing and Finance Company, the Tanganyika Development Finance Company the National Provident Fund, the Tanzania Post Office Savings Bank, the National Co-operative Bank, the National Development Credit Agency, and the National Lotteries.

Foreign trade. In recent years the pattern of trade has altered. Between 1962 and 1970 the value of sisal, coffee, and cotton exports dropped from 58 percent to 39 percent, and the value of imports of consumer goods dropped from 48 percent to 33 percent, while imports of capital goods increased from 26 percent to 40 percent. Total trade more than doubled. There is a growing trade with socialistic countries. Trade with Kenya and Uganda, while relatively large, is commonly marked by a deficit.

Tourism. Tanzania's eight national parks are among the finest in Africa. The Serengeti and Ruaha parks have a combined area of over 10,000 square miles. Gombe is noted for chimpanzees, and the Ngorongoro Park occupies the immense crater already mentioned. Hotel accommodations have increased greatly in the north and on the coast near Dar es Salaam. About 40,000 visitors came to Tanzania in 1968, and it is anticipated that this figure will soon double.

Management of the economy. *The private sector.* The private sector has diminished in importance, although

the bulk of the retail trade is still in private hands. African retailers, since the early 1960s, have outnumbered the previously predominant Asian traders by three to one. Other important businesses in private hands include the importation of motor vehicles, transportation, farming, construction, and numerous small industries. Private industrial companies continue to participate in joint ventures with the government.

Nationalization

The public sector and the role of the government. At the time of independence, in 1961, there was a modest public participation in the economy; the proportion has dramatically increased since 1967. The main targets of nationalization were the banks, large importing and exporting firms, and eight large milling firms with associated food-manufacturing industries. The National Insurance Company was brought under public ownership and given the monopoly of insurance, and the government received majority shares in the sisal industry and the wholesale trade. Major governmental capital expenditure for the financial year 1970 to 1971 was in the National Development Corporation of Tanzania, the Tanzania Tourist Corporation, the National Agriculture and Food Corporation, the Tanzania Tea Authority, the Tanzania Electric Supply Company, and the State Trading Corporation.

Taxation. Direct and indirect taxes are collected by three separate bodies. The East African Community is responsible for collecting income and surtaxes, import tax, excise duty, and a transfer tax levied on some manufactured goods from the partner states. The central government collects personal tax, corporation tax, export tax on crops, indirect sales tax, and estate and other taxes. Local government is responsible for taxes collected for vehicle registration, liquor licenses, and market dues.

Trade unions and employer associations. In 1964 the National Union of Tanganyika Workers (NUTA) was established as the sole body authorized to represent workers. Membership is not mandatory, although where more than 50 percent of an establishment are members, the minister responsible for labour can be requested to order the rest to join the union. Official strikes were declared illegal in 1962. There is also a Federation of Tanganyika Employers.

Ujamaa vijijini. Tanzania's policy of *ujamaa vijijini* was adopted in 1967, and official backing given to ideology and strategy for development at a grass-roots level for the diffusion of economic and political power after other cooperative efforts had failed. By 1970, more than 1,200 villages had been established. The main features of life in the villages include collective production, equality of opportunity, and self-reliance. The government is committed to support these collective peasant enterprises by supplying productive and social amenities.

Contemporary economic policies. The major means of production and exchange are intended to be under the control of the government, so that all resources of the country may be utilized to eliminate poverty, ignorance, and disease. Much emphasis has been placed on rural development and agriculture, which the government plans to revolutionize through *ujamaa* (the Tanzanian version of socialism). Toward this end, massive investments in development are projected in the second five-year development plan (1969 to 1974).

The second five-year plan

Problems and prospects. To offset dependence on primary commodities, which are subject to fluctuation in world prices as well as to fluctuations in output, production is being diversified.

The future direction that the economy will take depends to a great extent upon the success of the *ujamaa* program.

Transportation. Tanzania maintains regional transport links with Kenya and Uganda within the framework of a common authority over rail, shipping, and harbour development. Since 1965 similar links with Zambia have helped to end the isolation of southern Tanzania. Parts of a major road running from Dar es Salaam to Lusaka in Zambia, via Iringa and Mbeya, have been completed.

The general program for road construction aims at providing seven first class roads—four running east to west and three running north to south; parts of these roads have been completed. The number of feeder roads is also to be increased. Between 1961 and 1971 the length of paved roads increased from about 250 miles to about 1,400 miles, and the number of vehicles increased from 40,000 to 80,000.

The railways

The core of the railway system consists of the 218-mile German-built railway from Tanga to Moshi, and the 830-mile east-to-west Central Line from Dar es Salaam to Kigoma. The two lines were not linked until 1963. During the British administration the Tanga Railway was joined to the Kenya railway system and was also extended to Arusha, while the Central Line was linked with the lake port of Mwanza. Construction of the 1,100-mile Tanzam Railway from Dar es Salaam to Zambia, due to completion by 1977, began in 1970.

Dar es Salaam, the facilities of which include six deep-water berths, is the main port; its hinterland stretches to Burundi, Zaire, and Zambia. Tanga to the north and Mtwara to the south are smaller but are significant for the export of sisal and cashew nuts, respectively. Regular shipping services operate on Lake Victoria and Lake Tanganyika; Lake Victoria has a train–ferry operating between the lake ports.

The national airline, which operates both domestic and international services, is owned jointly with Kenya and Uganda. Dar es Salaam was the only international airport until, to promote the tourist trade, another international airport was completed at Kilimanjaro in 1970. Most international airlines call at Dar des Salaam. Scheduled domestic flights connect about 20 smaller airports. Nearly 200 small landing strips dot the country.

ADMINISTRATION AND SOCIAL CONDITIONS

The structure of government. *The constitutional framework.* The Interim Constitution, established in 1965, gives Tanzania two authorities—the United Republic and Zanzibar. The former has rights over the mainland as well as over a number of union matters, which include the constitution, external affairs, defense, police, emergency powers, citizenship, immigration, external trade and borrowing, most taxes and communications, currency, and the public services of the United Republic. The Zanzibar authority deals with nonunion affairs and with internal matters pertaining to the former People's Republic of Zanzibar. The Zanzibar government has its own executive and legislature consisting of a Revolutionary Council. The government is headed by the president of Zanzibar.

The presidency

The executive power of the United Republic is vested in the president, who is head of state and commander in chief of the armed forces. He is advised by a cabinet of ministers, including the two vice presidents. One of the vice presidents must be from Zanzibar; the second is the leader of the National Assembly and the president's principal assistant.

The National Assembly and the president constitute Parliament. A large majority of members of Parliament, who represent constituencies and institutions, are elected, but the 32 members of the Zanzibar Revolutionary Council and up to 20 others from Zanzibar are nominated by the president. Regional commissioners are ex officio members. Other provisions of the constitution concern the maintenance of an independent judiciary, the establishment of a Permanent Commission of Enquiry, union financial matters, and armed forces.

Regional, state, and local government. In organization for the administration of mainland Tanzania there are 17 regions and the city of Dar es Salaam. Each region is administered by a regional commissioner responsible to the central government and appointed by the president. Regions are divided into districts under an area commissioner. The commissioners, being political appointees, help to link the administration with the party.

There are both urban and rural local governments. The former Native Authority Councils under chiefs (who have had no official power since 1961) have been replaced by nearly 60 district councils that are elected and dominated by Tanzania's single party (see below). Only the large towns have municipal or town councils.

Elections. Tanzania has universal suffrage. Elections are held at all levels, from the ten-household cells (see below) to the National Assembly. For the National Assembly, the National Executive Council of the national party presents two names from each constituency from which the electorate makes its choice. General elections are held every five years. Constituencies are homogenous areas grouping about 100,000 people within each district. There were 120 constituencies in 1970.

The Tanganyika African National Union

The party. Tanzania became a one-party state in 1965, with the Tanganyika African National Union (TANU) at its helm. In Zanzibar, the sister organization is the Afro-Shirazi Party. TANU is a mass party with membership open to all adults older than 18 years. The party has a number of sections (such as a women's organization and a parents' association) and institutional affiliates. The basic organ of the TANU is a cell, which consists of ten households under an elected leader who represents the cell at the annual conference of the party's local branch. Each such conference has an executive body. TANU also has organs at the district and regional level. The national organs consist of committees and conferences, including the National Executive Committee.

De facto political developments. Tanzania has chosen a socialistic form of government. To prevent elites, members of Parliament and high-income civil servants, under the leadership code, may not receive two or more salaries or own shares in companies. More than 70 percent of the eligible voters registered for the 1970 elections. The presidential incumbent, Julius Nyerere, received over 98 percent of the votes in Zanzibar and 94 percent on the mainland.

Justice. Under the constitution, the judiciary is granted complete independence from political and executive influences. Judges are appointed by the president in consultation with the chief justice. A unified judiciary system was initiated in 1962 when the former native and local courts dealing with customary laws became primary courts. There are more than 800 primary courts, where magistrates, each assisted by two local assessors, and using Swahili, have jurisdiction over customary and Islāmic law and certain sections of the penal code. Qualified lawyers, acting as resident magistrates in district courts, hear appeals from primary courts and also have regional jurisdiction in most criminal and civil cases. The High Court, consisting of eight puisne (junior) judges and the chief justice, have original jurisdiction over constitutional matters, serious crimes including murder, and civil claims for large sums, and they also hear appeals from lower courts. The final court of appeal is the Court of Appeal for East Africa. The tenets of law are made up of English, Islāmic, and customary laws. Customary laws, often unwritten, are being restated.

In Zanzibar and Pemba, Muslim courts deal with such personal matters as marriage, divorce, and inheritance. The Revolutionary Council assesses justice after the decision of the courts.

Armed forces. The main defense force consists of an army of 10,000 men. Four infantry battalions are supported by 12 light tanks, other armed vehicles, and a few artillery units. Elementary military instruction is given during national service. The navy has four small patrol boats and about 100 men. The air force, which is unarmed, consists of ten transport planes and seven trainers.

Administration. *Education.* The education system has three levels—primary school (seven years), secondary school (four to six years), and university. There are also professional, vocational, and religious schools. Secondary school education has been greatly expanded in recent years, and primary school education has also increased. There are about 4,500 primary schools and more than 100 secondary schools. In 1970 the constituent part of the University of East Africa located at Dar es Salaam was established as a national institution, the University of Dar es Salaam.

Educational orientation

Following the views expressed in *Education for Self-Reliance,* written by President Nyerere in 1967, the education system was reoriented away from preparation for university education and toward the goal of community service. The entrance age to primary schools was raised to seven years, and practical instruction was emphasized.

Health services. There is a countrywide network of medical services operated by the government, by local authorities, and by missions. There are more than 120 hospitals run jointly by the government and missions, but for the majority of the people the 1,600 dispensaries are the only source of medical assistance. Public medical services are socialized, but there are also private practitioners. There is one doctor for every 22,000 people. The national expenditure on health has been greatly increased. Emphasis is placed mainly on preventive medicine, which stresses nutrition, environmental sanitation, maternal and child health, control of communicable diseases, and the establishment of rural health centres.

Housing. There is a program for the construction of low-income housing. The major expenditure of the National Housing Corporation established in 1962, has been on slum clearance in Dar es Salaam, where about 4,000 low-cost and about 400 medium-cost housing units had been constructed by 1971.

Police services. Tanzania has a police force of 8,000 men and women under the control of an inspector general. In addition to the regular police force, there are railway, mine, and marine police units. The police are generally unarmed and are assisted by a special constabulary and, on occasions, by civilian volunteers. In addition, there are four special paramilitary units.

Social conditions. *Wages.* Only 9 percent of the economically active labour force is in the wage economy. The highest average wages are gained in commerce, the lowest in agriculture. Minimum wages are set by law.

Environmental diseases

Health conditions. The main environmental diseases are malaria, sleeping sickness, schistosomiasis (a parasitic disease), and onchocerciasis (a parasitic disease that may cause blindness). Communicable diseases include smallpox, poliomyelitis, leprosy, tuberculosis, dysenteries, and enteric fevers. Defective nutrition is also a major problem. Significant improvements in health have resulted from implementation of control measures.

CULTURAL LIFE AND INSTITUTIONS

Tanzania has a remarkable cultural milieu. One of the earliest evidences of human ancestry, dating back 1,250,000 years, was discovered at Olduvai Gorge, just north of the tip of Lake Eyasi in the Eastern Rift area. Rock paintings, mounds, and other evidence of prehistoric human settlement continue to be found. Transoceanic contacts with Arabia, India, and China, and later with Europe, have left their mark. Of even greater significance in the heritage of the country are the variety of traditions, languages, and customs that exist. The national language, Swahili, is perhaps one of the most versatile languages in black Africa as well as one of the most widely distributed.

The arts. Among the plastic arts practiced are traditional abstract Makonde wood carving, the more common Zaramo carving of animals and people, and the fabrication of weaponry by the Masai. Zanzibar is renowned for its Arab chests and elaborately carved and studded doors. Elsewhere, crafts include the making of baskets, mats, pots, skin work, drums, and musical instruments. Every tribe has its dances and folklore. Swahili poetry and prose writing, once neglected, now flourish. The prose and poetry of Shaaban Robert (who died in 1962) merit attention, as do President Julius Nyerere's translations of Shakespearean plays into Swahili.

Press and broadcasting. Much has been done to give the press a national identity and to widen its circulation. The important Swahili dailies are *Uhuru* and *Ngurumo.* Several other smaller Swahili and vernacular newspapers are published throughout the country. The English language newspaper *The Standard-Tanzania* has the second largest circulation. Although nationalized, it operates independently and commercially. It also publishes a weekly, the *Sunday News.* The *Nationalist,* another English language daily, is the party newspaper. East African dailies printed in Nairobi also have a wide circulation.

The Tanzania Information Service is responsible for the

Radio broadcasting

reception of news from abroad, through agreement with Reuters and other agencies, and for the dissemination of local news. Because of the size of the country, radio broadcasting is very important. Radio Tanzania broadcasts in Swahili and English, with special programs for schools and farmers. Public listening posts and private sets assure widespread reception. Programs in English, Ambo (Ovambo), Zulu, Herero, and Portuguese are broadcast to southern Africa to assist African liberation movements there.

PROSPECTS FOR THE FUTURE

The development of Tanzania as an agricultural country suffers as a result of falling prices for primary products and rising prices for manufactured goods. To remedy the situation, efforts are being made to diversify export crops and markets.

In the decade since independence, stock has been taken of national resources, and priorities and strategies for change have been redefined. There has been a shift away from massive investment in a few settlement schemes to the establishment of numerous socialistic villages (*ujamaa vijijini*); from industrial development to emphasis on rural development; from dependence on foreign aid to a greater degree of self-reliance. Since 1967, the major means of production have been nationalized, and the emphasis is now on socialist planning.

During the first five-year plan (1964–69), it was intended that economic growth take place at the rate of 6.7 percent a year, but growth took place at 5 percent. The proposed investment for the second plan (1969–74) aims to achieve an annual growth rate of 6.5 percent as a result of applying the five principles of social equality: *Ujamaa*, self-reliance, economic transformation, social transformation, and economic links with African countries.

Politically, public participation in the affairs of the country has been made possible through popular elections. Tanzania's foreign policy is based on defending basic human rights, and this defense is typified by the fight against apartheid (the policy of racial segregation of South Africa) and colonialism. Tanzania's other interests include maintaining unity and cooperation among African states, and they have also found expression in the East African Community as well as in the establishment of a practical and mutually beneficial relationship with Zambia.

BIBLIOGRAPHY. R. OLIVER *et al.* (eds.), *History of East Africa*, 3 vol. (1963–68), gives an authoritative historical account of Tanzania during the colonial era, in an East African context. Resources are regionally covered in E.W. RUSSELL (ed.), *The Natural Resources of East Africa* (1962). The two companion volumes, *Tanganyika: A Review of Its Resources and Their Development* (1955) and *A Handbook of Tanganyika*, 2nd ed. (1958), both ed. by J.P. MOFFETT, mostly cover developments during the British period. A contemporary semi-official handbook with a comprehensive though not detailed coverage of the United Republic is *Tanzania Today* (1968). Spatial aspects of resources and development are found in the official *Atlas of Tanzania* (1967); and the more comprehensive *Tanzania in Maps*, ed. by LEONARD BERRY (1972). Various government ministries publish annual reports. For post-independent economic development, the volumes for the two *Five-Year Plan(s) for Economic and Social Development* (1964–69) and (1969–74), issued by the Government Printer, are important documents. The *Economic Survey and Annual Plan* annually supplements the current plan. President JULIUS NYERERE, *Freedom and Unity* (1967), *Freedom and Socialism* (1968), and *Education for Self Reliance* (1967), contain important policy statements. For Zanzibar, M.F. LOFCHIE, *Zanzibar: Background to Revolution* (1965), is important. A useful summary of society and politics is found in J. MIDDLETON and J. CAMPBELL, *Zanzibar: Its Society and Its Politics* (1965). The official account of development is contained in *The Fruits of the Revolution in the Islands: The Story of the Progress Made in Zanzibar and Pemba During the Three Years Following the Revolution of January 12th, 1964*, published by the Government Printer (1967). General statistical information may be found in the *Statistical Abstracts* (1964–), issued by the CENTRAL STATISTICAL BUREAU.

(A.C.M.)

Taoism

Taoism, along with Confucianism, is one of the two major indigenous religio-philosophical traditions that have shaped Chinese life for more than 2,000 years.

In the broadest sense, a Taoist attitude toward life can be seen in the accepting and yielding, the joyful and carefree sides of the Chinese character, an attitude that offsets and complements the moral and duty-conscious, austere and purposeful character ascribed to Confucianism. Taoism is also characterized by a positive, active attitude toward the occult and the metaphysical (theories on the nature of reality), whereas the agnostic, pragmatic Confucian tradition considers these issues of only marginal importance, although the reality of such issues is, by most Confucians, not denied.

More strictly defined, Taoism includes: the ideas and attitudes peculiar to the *Lao-tzu* (or *Tao-te Ching;* "Classic of the Way of Power"), the *Chuang-tzu*, the *Lieh-tzu*, and related writings; the Taoist religion, which is concerned with the ritual worship of the Tao; and those who identify themselves as Taoists. (See also TAOISM, HISTORY OF; TAOIST LITERATURE; and CONFUCIANISM.)

NATURE AND SIGNIFICANCE

Taoist thought permeates Chinese culture, including many aspects not usually considered Taoist. In Chinese religion, the Taoist tradition—often serving as a link between the Confucian tradition and folk tradition—has generally been more popular and spontaneous than the official (Confucian) state cult and less diffuse and shapeless than folk religion.

Taoist philosophy and religion have found their way into all Asian cultures influenced by China, especially those of Vietnam, Japan, and Korea. Various religious practices reminiscent of Taoism in such areas of Chinese cultural influence indicate early contacts with Chinese travellers and immigrants that have yet to be elucidated.

Distinctions between Taoist philosophy and religion

Both Western Sinologists and Chinese scholars themselves have distinguished—since Han times (206 BC–AD 220)—between a Taoist philosophy of the great mystics and their commentators (*Tao-chia*) and a later Taoist religion (*Tao-chiao*). This theory—no longer considered valid—was based on the view that the "ancient Taoism" of the mystics antedated the "later Neo-Taoist superstitions" that were misinterpretations of the mystics' metaphorical images. The mystics, however, should be viewed against the background of the religious practices existing in their own times. Their ecstasies, for example, were closely related to the trances and spirit journeys of the early magicians and shamans (religious personages with healing and psychic transformation powers). Not only are the authors of the *Tao-te Ching*, the *Chuang-tzu* (book of "Master Chuang"), and the *Lieh-tzu* (book of "Master Lieh") not the actual and central founders of an earlier "pure" Taoism later degraded into superstitious practices but they can even be considered somewhat on the margin of older Taoist traditions. Therefore, because there has been a nearly continuous mutual influence between Taoists of different social classes—philosophers, ascetics, alchemists, and the priests of popular cults—the distinction between philosophical and religious Taoism in this article is made simply for the sake of descriptive convenience.

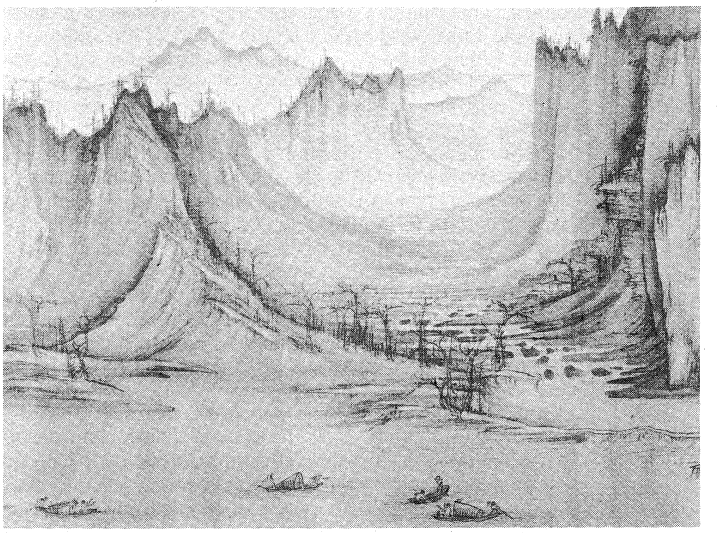
There is also a tendency among scholars today to draw a less rigid line between what is called Taoist and what is called Confucian. The two traditions share many of the same ideas about man, society, the ruler, Heaven, and the universe—ideas that were not created by either school but that stem from a tradition prior to either Confucius or Lao-tzu.

Viewed from this common tradition, orthodox Confucianism limited its field of interest to the creation of a moral and political system that fashioned society and the Chinese empire; whereas Taoism, inside the same world view, represented more personal and metaphysical preoccupations.

In the case of Buddhism—a third tradition that influenced China—fundamental concepts such as the nonexistence of the individual ego and the illusory nature of the

physical world are diametrically opposed to Taoism. In terms of overt individual and collective practices, however, competition between these two religions for influence among the people—a competition in which Confucianism had no need to participate because it had state patronage—resulted in mutual borrowings, numerous superficial similarities, and essentially Chinese developments inside Buddhism, such as the Ch'an (Zen) sect. In folk religion, since Sung times (960–1279), Taoist and Buddhist elements have coexisted without clear distinctions in the minds of the worshippers.

Certain concepts of ancient agrarian religion have dominated Chinese thought uninterruptedly from before the formation of the philosophic schools until the first radical break with tradition and the overthrow of dynastic rule at the beginning of the 20th century, and they are thus not specifically Taoist. The most important of these concepts are: the solidarity of nature and man; that is, the interaction between the universe and human society; the cyclical character of time and the universal rhythm and the law of return; and the worship of ancestors, the cult of Heaven, and the divine nature of the sovereign.



"Fishing in a Mountain Stream," detail of an ink drawing on silk by Hsü Tao-ning, 11th century. The drawing suggests the Taoist concept of harmony of the universe and man's relative role in the universal order. In the William Rockhill Nelson Gallery and Atkins Museum of Fine Arts, Kansas City, Missouri.

By courtesy of the Nelson Gallery-Atkins Museum, Kansas City, Missouri (Nelson Fund)

BASIC CONCEPTS OF TAOISM

The meaning of Tao

Man and the universe. Tao is a key term in all Chinese thought, including the Confucian. In common usage it means "way," also *the* Way to be followed and, by extension, a code of behaviour and a doctrine. The philosophical notion of Tao expressed the religious sympathy and complete solidarity that unites nature and man. In the magical arts that influenced Taoism, Tao designated the magical feat of bringing Heaven and Earth, the sacred powers and man, into communication with each other. In this sense, Tao was an art and a power, specifically the power of the magician and the king. Tao also was the universal force in the cosmos, expressed, in the human realm, in the power of the exemplary ruler. The "Way of the King" (*wang tao*) consisted in periodical ritual actions intended to restore order in nature by imitating the "Way of Heaven" (*t'ien tao*).

This is what Lao-tzu called the Tao that can be named. He used the name Tao also, however, for the ineffable reality, experienced in ecstasy, that lies at the origin of the universe and behind or within appearances. This permanent Tao (*ch'ang tao*) comes nearest to what other religions and mysticisms call the Absolute.

The microcosm–macrocosm concept. The conception of the universe common to all Chinese philosophy is neither materialistic nor animistic (a belief system centring on soul substances); it can be called magical or even alchemical. The universe is viewed as a hierarchically organized mechanism in which every part reproduces the whole. Man is a microcosm (small universe) corresponding rigorously to this macrocosm (large universe); his body reproduces the plan of the cosmos. Between man and universe there exists a system of correspondences and participations that the ritualists, philosophers, alchemists, and physicians have described but certainly not invented. This originally magical feeling of the integral unity of mankind and the natural order has always characterized the Chinese mentality, and the Taoists especially have elaborated upon it. The five organs of the body and its orifices and the dispositions, features, and passions of man correspond to the five directions, the five holy mountains, the sections of the sky, the seasons, and the elements (*wu-hsing*), which in China are not material but more like five fundamental phases of any process in space-time. Whoever understands man thus understands the structure of the universe. The physiologist knows that blood circulates because rivers carry water and that the body has 360 articulations because the ritual year has 360 days. In religious Taoism the interior of the body is inhabited by the same gods as those of the macrocosm. An adept often searches for his divine teacher in all the holy mountains of China until he finally discovers him in one of the "palaces" inside his head.

Return to the Tao. The law of the Tao as natural order refers to the continuous reversion of everything to its starting point. Anything that develops extreme qualities will invariably revert to the opposite qualities: "Reversion is the movement of the Tao" (*Lao-tzu*). All being issues from the Tao and ineluctably returns to it; Undifferentiated Unity becomes multiplicity in the movement of the Tao. Life and death are contained in this eternal transformation from Non-Being into Being and back to Non-Being, but the underlying primordial unity is never lost.

The eternal transformation from Non-Being to Being to Non-Being

For society, any reform means a type of return to the remote past; civilization is considered a degradation of the natural order, and the ideal is the return to an original purity. For the individual, wisdom is to conform to the rhythm of the universe. The Taoist mystic, however, not only adapts himself ritually and physiologically to the alternations of nature but creates a void inside himself that permits him to return to nature's origin. Lao-tzu, in trance, "wandered freely in the origin of all beings." Thus, in ecstasy he escaped the rhythm of life and death by contemplating the universal return. "Having attained perfect emptiness, holding fast to stillness, I can watch the return of the ever active Ten Thousand Beings." The number 10,000 symbolizes totality.

Change and transformation. All parts of the universe are attuned in a rhythmical pulsation. Nothing is static; all beings are subjected to periodical mutations and transformations that represent the Chinese view of creation. Instead of being opposed with a static ideal, change itself is systematized and made intelligible, as in the theory of the five phases (*wu-hsing*) and in the 64 hexagrams of the *I Ching* (*Book of Changes*), which are basic recurrent constellations in the general flux. An unchanging unity (the permanent Tao) was seen as underlying the kaleidoscopic plurality.

Chuang-tzu's image for creation was that of the activity of the potter and the bronze caster: "to shape and to transform" (*tsao hua*). These are two phases of the same process: the imperceptible Tao shapes the universe continuously out of primordial chaos; the perpetual transformation of the universe by the alternations of Yin and Yang, or complementary energies (seen as night and day or as winter and summer), is nothing but the external aspect of the same Tao. The shaping of the Ten Thousand Beings by the Supreme Unity and their transformation by Yin and Yang are both simultaneous and perpetual. Thus, the saint's ecstatic union is a "moving together with the Tao; dispersing and concentrating, his appearance has no consistency." United with the permanent Tao, the saint's outer aspect becomes one of

ungraspable change. Because the gods can become perceptible only by adapting to the mode of this changing world, their apparitions are "transformations" (*pien-hua*); and the magician (*hua-jen*) is believed to be one who transforms rather than one who conjures out of nothing.

The social ideal of primitivism. Any willful human intervention is believed to be able to ruin the harmony of the natural transformation process. The spontaneous rhythm of the primitive agrarian community and its unself-conscious symbiosis with nature's cycles is thus the Taoist ideal of society.

The ideal society

In the ideal society there are no books; the *Lao-tzu* (*Tao-te Ching*) itself would not have been written but for the entreaty of the guardian of the pass Yin Hsi, who asked the "Old Master" to write down his thoughts. In the Golden Age, past or future, knotted cords are the only form of records. The people of this age are "dull and unwitting, they have no desire; this is called uncarved simplicity. In uncarved simplicity the people attain their true nature."

Chuang-tzu liked to oppose the Heaven-made and the man-made; that is, nature and society. He wanted man to renounce all artificial "cunning contrivances" that facilitate his work but lead to "cunning hearts" and agitated souls in which the Tao will not dwell. Man should equally renounce all concepts of measure, law, and virtue. "Fashion pecks and bushels for people to measure by and they will steal by peck and bushel." He blamed not only the culture heroes and inventors praised by the Confucians but also the sages who shaped the rites and rules of society.

> That the unwrought substance was blighted in order to fashion implements—this was the crime of the artisan. That the Way (Tao) and its Virtue (*te*) were destroyed in order to create benevolence and righteousness—this was the fault of the sage.

Even "coveting knowledge" is condemned because it engenders competition and "fight to the death over profit."

Religious goals of the individual. The Confucian saint (*sheng*) is viewed as a ruler of antiquity or a great sage who taught men how to return to the rites of antiquity. The Taoist sainthood, however, is internal (*nei sheng*), although it can become manifest in an external royalty (*wai wang*) that brings the world back to the Way by means of quietism: variously called "non-intervention" (*wu-wei*), "inner cultivation" (*nei yeh*), or "art of the heart and mind" (*hsin-shu*).

Whereas worldly ambitions, riches, and (especially) discursive knowledge scatter the person and drain his energies, the saint "embraces Unity" or "holds fast to the One" (*pao i*); that is, he inspires to union with the Tao in a primordial undivided state underlying consciousness. "Embracing Unity" also means that he maintains the balance of Yin and Yang within himself and the union of his spiritual (*hun*) and vegetative (*p'o*) souls, the dispersion of which spells death; Taoists usually believed there were three *hun* and seven *p'o*. The spiritual soul tends to wander (in dreams), and any passion or desire can result in loss of soul. To retain and harmonize one's souls is important for physical life as well as for the unification of the whole human entity. Cleansed of every distraction, the saint creates inside himself a void that in reality is plenitude. Empty of all impurity, he is full of the original energy (*yüan ch'i*), which is the principle of life that in the ordinary man decays from the moment of birth on.

Proof of sainthood

Because vital energy and spirituality are not clearly distinguished, old age in itself becomes a proof of sainthood. The aged Taoist sage became a saint because he had been able to cultivate himself throughout a long existence; his longevity in itself was the proof of his saintliness and union with the Tao. Externally he had a healthy, flourishing appearance and inside he contained an ever-flowing source of energy that manifested itself in radiance and in a powerful, beneficial influence on his surroundings, which is the charismatic efficacy (*te*) of the Tao.

The mystic insight of Chuang-tzu made him scorn those who strove for longevity and immortality through physiological practices. Nevertheless, physical immortality was a Taoist goal probably long before and alongside the unfolding of Taoist mysticism. The adept of immortality had a choice among many methods that were all intended to restore the pure energies possessed at birth by the infant whose perfect vital force Lao-tzu admired. Through these methods, the adept became an immortal (*hsien*) who lived 1,000 years in this world if he so chose and, once satiated with life, "ascended to heaven in broad daylight." This was the final apotheosis of the Taoist who had transformed his body into pure Yang energy.

Symbolism and mythology. Taoists prefer to convey their ecstatic insights in images and parables. The Tao is low and receiving as a valley, soft and life-giving as water, and it is the "mysterious female," the source of all life, the Mother of the Ten Thousand Beings. Man should become weak and yielding as water that overcomes the hard and the strong and always takes the low ground; he should develop his male and female sides but "prefer femininity," "feed on the mother," and find within himself the well that never runs dry. Tao is also the axis, the ridgepole, the pivot, and the empty centre of the hub. The sage is the "useless tree" or the huge gourd too large to be fashioned into implements. A frequent metaphor for the working of the Tao is the incommunicable ability to be skillful at a craft. The skilled artisan does not ponder on his action, but, in union with the Tao of his subject, he does his work reflexively and without conscious intent.

Preservation and transformation of ancient Chinese myths

Much ancient Chinese mythology has been preserved by the Taoists, who drew on it to illustrate their views. A chaos (*hun-tun*) myth is recorded as a metaphor for the undifferentiated primal unity; the mythical emperors (Huang Ti and others) are extolled for wise Taoist rule or blamed for introducing harmful civilization. Dreams of mythical paradises and journeys on clouds and flying dragons are metaphors for the wanderings of the soul, the attainment of the Tao, and the identity of dream and reality.

Taoists have transformed and adapted some ancient myths to their beliefs. Thus, the Queen Mother of the West (Hsi Wang Mu), who was a mountain spirit, pestilence goddess, and tigress, became a high deity—the Fairy Queen of all immortals.

PHILOSOPHICAL TAOISM

Philosophical Taoism is developed mainly in the book *Lao-tzu* (or *Tao-te Ching*) and in the apparently later work of Chuang-tzu (late-4th century–*c.* 300 BC).

Discursive knowledge, for Lao-tzu and Chuang-tzu, had only the negative role of demonstrating the relativity of all knowledge that is not gained in ecstatic vision. Solitary meditation, "sitting in forgetfulness" (*tso wang*), is viewed as the sole way to true knowledge and power. For Chuang-tzu, "the Tao cannot be affirmed to exist or not to exist. The name Tao is merely an artifice for practical purposes . . . neither speech nor silence can convey it. Let us banish both speech and silence." Nevertheless, these problems were discussed unendingly. Tao was used intellectually as a "principle of universal intelligibility." Chuang-tzu's injunction "vomit your intelligence" expressed only the Taoist ridicule for dialectical argumentation and any kind of abstract realism.

The Tao cannot be grasped, but it can be received. Under a master "teaching without words," the adept goes through a cathartic process of emptying his mind of all passions, notions, and distinctions ("the cleansing of the heart," *hsin chai*) until it becomes a "mirror of Heaven and Earth reflecting the multiplicity of things." He becomes inhabited by the Tao and finally reaches an ecstatic trance in which "his body has the appearance of dead wood and his soul the insensibility of dead ashes," while he floats, not knowing "whether he carries the wind or whether the wind carries him," in time and space—or, rather, in an eternal now: "strolling at the origin of all things."

The teachings of Lao-tzu. The *Tao-te Ching* is a short anthology of paradoxical statements circumscribing the nature of the Tao and of maxims for the ruler who governs through the power (*te*) of the Tao.

Variations on the meaning of Tao

Cosmology. What Lao-tzu calls the "permanent Tao" in reality is nameless. The name (*ming*) in ancient Chinese thought implied an evaluation assigning an object its place in a hierarchical universe. The Tao is outside these categories.

> It is something formlessly fashioned, that existed before Heaven and Earth; . . . Its name (*ming*) we do not know; Tao is the byname that we give it. Were I forced to say to what class of things it belongs I should call it Immense.

Tao is the "imperceptible, indiscernible," about which nothing can be predicated but that latently contains the forms, entities, and forces of all particular phenomena: "It was from the Nameless that Heaven and Earth sprang; the Named is the mother that rears the Ten Thousand Beings, each after its kind." The Nameless (*wu-ming*) and the Named (*yu-ming*), Not-Being (*wu*) and Being (*yu*), are interdependent and "grow out of one another."

Not-Being (*wu*) and Tao are not identical; *wu* and *yu* are two aspects of the permanent Tao: "in its mode of being Unseen, we will see its mysteries; in the mode of the Seen, we will see its boundaries."

Not-Being does not mean Nothingness but rather the absence of perceptible qualities; in Lao-tzu's view it is superior to Being. It is the Void (that is, empty incipience) that harbours in itself all potentialities and without which even Being lacks its efficacy.

Emptiness realized in the mind of the Taoist who has freed himself from all obstructing notions and distracting passions makes the Tao act through him without obstacle. An essential characteristic that governs the Tao is spontaneity (*tzu-jan*), the what-is-so-of-itself, the self-so, the unconditioned. The Tao, in turn, governs the universe: "The ways of Heaven are conditioned by those of the Tao, and the ways of Tao by the Self-so."

This is the way of the saint who does not intervene but possesses the total power of spontaneous realization that is at work in the universe; of his accomplishments "everyone, throughout the country, says 'It happened of its own accord' (*tzu-jan*)."

Man and society. The power acquired by the Taoist is *te*, the efficacy of the Tao in the realm of Being, which is translated as "virtue." Lao-tzu viewed it, however, as different from Confucian virtue:

> The man of superior virtue is not virtuous, and that is why he has virtue. The man of inferior [Confucian] virtue never strays from virtue, and that is why he has no virtue.

The "superior virtue" of Taoism is a latent power that never lays claim to its achievements; it is the "mysterious power" (*hsüan te*) of Tao present in the heart of the sage—"the man of superior virtue never acts (*wu-wei*), and yet there is nothing he leaves undone."

Wu-wei is not an ideal of absolute inaction nor a mere "not-overdoing." It is an action so well in accordance with things that its author leaves no trace of himself in his work: "Perfect activity leaves no track behind it; perfect speech is like a jade worker whose tool leaves no mark." It is the Tao that "never acts, yet there is nothing it does not do." There is no true achievement without *wu-wei* because every deliberate intervention in the natural course of things will sooner or later turn into the opposite of what was intended and will result in failure.

The sage who practices *wu-wei* lives out of his original nature before it was tampered with by knowledge and restricted by morality; he has reverted to infancy (that is, the undiminished vitality of the newborn state); he has "returned to the state of the Uncarved Block (*p'u*)." *P'u* is uncut, unpainted wood, simplicity. Society carves this wood into specific shapes for its own use and thus robs the individual piece of its original totality. "Once the uncarved block is carved, it forms utensils (that is, instruments of government); but when the Sage uses it, he would be fit to become Chief of all Ministers. This is why the great craftsman (ruler) does not carve (rule)."

The *Tao-te Ching* as a handbook for rulers

The *Tao-te Ching* was meant as a handbook for the ruler. He should be a sage whose actions pass so unnoticed that his very existence remains unknown. He imposes no restrictions or prohibitions on his subjects; "so long as I love quietude, the people will of themselves go straight. So long as I act only by inactivity, the people will of themselves become prosperous." His simplicity makes the Ten Thousand Beings passionless and still and peace follows naturally. He does not teach them discrimination, virtue, or ambition because "when intellect emerges, the great artifices begin. When discord is rife in families, 'dutiful sons' appear. When the State falls into anarchy, 'loyal subjects' appear." Thus, it is better to banish wisdom, righteousness, and ingenuity, and the people will benefit a hundredfold.

> Therefore the Holy Man rules by emptying their hearts (minds) and filling their bellies, weakening their wills and strengthening their bones, ever striving to make the people knowledgeless and desireless.

The word people in this passage more likely refers not to the common people but to those nobles and intellectuals who incite the ruler's ambition and aggressiveness.

War is condemned but not entirely excluded: "Arms are ill-omened instruments," and the sage uses them only when he cannot do otherwise. He does not glory in victory; "he that has conquered in battle is received with rites of mourning."

The teachings of Chuang-tzu. Whereas the *Lao-tzu* is addressed to the sage-king, the *Chuang-tzu* is the earliest surviving Chinese text to present a philosophy for private life, a wisdom for the individual. Chuang-tzu is said to have preferred the doctrine of Lao-tzu over all others; many of his writings strike the reader as metaphorical illustrations of the terse sayings of the "Old Master."

Whereas Lao-tzu in his book as well as in his life (in legend) was concerned with Taoist rule, Chuang-tzu, some generations later, rejected all participation in society. He compared the servant of state to the well-fed decorated ox being led to sacrifice in the temple and himself to the untended piglet blissfully frolicking in the mire.

No Chinese philosopher, however, has been able to divert his interest completely from political problems. Chuang-tzu's answers are those of the mystic. The ruler should seek the Tao and do nothing else. The mere presence of a holy man makes a country prosper: "By concentrating his spirit, he can protect creatures from sickness and plague and make the harvest plentiful."

Ideas of knowledge and language. Characteristic of Chuang-tzu are his ideas of knowledge and language developed under the stimulus of his friend and opponent, the philosopher Hui Shih.

Because, in the Taoist view, all beings and everything are fundamentally one, opposing opinions can arise only when people lose sight of the Whole and regard their partial truths as absolute. They are then like the frog at the bottom of the well who takes the bit of brightness he sees for the whole sky. The closed systems—*i.e.*, the passions and prejudices into which petty minds shut themselves—hide the Tao, the "Supreme Master" who resides inside themselves and is superior to all distinctions.

The relativity of moral and epistemological notions

Thus, Chuang-tzu's holy man fully recognizes the relativity of notions like good and evil and true and false. He is neutral and open to the extent that he offers no active resistance to any would-be opponent, whether it be a person or an idea. "When you argue, there are some things you are failing to see. In the greatest Tao nothing is named; in the greatest disputation, nothing is said."

The person who wants to know the Tao is told: "Don't meditate, don't cogitate. . . . Follow no school, follow no way, and then you will attain the Tao"; discard knowledge, forget distinctions, reach no-knowledge. "Forget" indicates that distinctions had to be known first. The original ignorance of the child is distinguished from the no-knowledge of the sage who can "sit in forgetfulness."

The mystic does not speak because declaring unity, by creating the duality of the speaker and the affirmation, destroys it. Those who speak about the Tao (like Chuang-tzu himself) are "wholly wrong. For he who knows does not speak; he who speaks does not know." Chuang-tzu was aware of the fact that, in speaking about it, he could do no more than hint at the way toward the all-embracing and intuitive knowledge.

Identity of life and death. Mystic realization does away with the distinction between the self and the world.

This idea also governs Chuang-tzu's attitude toward death. Life and death are but one of the pairs of cyclical phases, such as day and night or summer and winter. "Since life and death are each other's companions, why worry about them? All beings are one." Life and death are not in opposition but merely two aspects of the same reality, arrested moments out of the flux of the universal mutations of everything into everything. Man is no exception; "he goes back into the great weaving machine: thus all beings issue from the Loom and return to the Loom."

Viewed from the single reality experienced in ecstasy, it is just as difficult to distinguish life from death as it is to distinguish the waking Chuang-tzu from the dreaming butterfly. Death is natural, and men ought neither to fear nor to desire it. Chuang-tzu's attitude thus is one of serene acceptance.

The perfect man. Chuang-tzu's descriptions of the indescribable Tao, as well as of those who have attained union with the Tao, are invariably poetic. The perfect man has identified his life rhythm so completely with the rhythm of the forces of nature that he has become indistinguishable from them and shares their immortality and infinity, which is above the cycle of ordinary life and death. He is "pure spirit. He feels neither the heat of the brushlands afire nor the cold of the waters in flood"; nothing can startle or frighten him. Not that he is magically invulnerable (as the adepts of physical immortality would have it), but he is "so cautious in shunning and approaching, that nothing can do him injury."

The perfect man as pure spirit

"A man like this rides the clouds as his carriages and the sun and moon as his steeds." The theme of the spiritual wandering (*yüan yu*), which can be traced back to the shamanistic soul journey, crops up wherever Chuang-tzu speaks of the perfect man.

> Those who let themselves be borne away by the unadulterated energies of Heaven and Earth and can harness the six composite energies to roam through the limitless, whatever need they henceforth depend on?

These wanderings are journeys within oneself; they are roamings through the Infinite in ecstasy. Transcending the ordinary distinctions of things and one with the Tao, "the Perfect Man has no self, the Holy Man has no merit, the Sage has no fame." He lives inconspicuously among men, and whatever applies to the Tao applies to him.

Contributions of the "I Ching" and the school of Natural Change. *The idea of Yin and Yang.* Yin and Yang literally mean "dark side" and "sunny side" of a hill. They are mentioned for the first time in the *Hsi tz'u*, or "Appended Explanations" (*c.* 4th century BC), an appendix to the *I Ching (Book of Changes)*: "One [time] Yin, one [time] Yang, this is the Tao." Yin and Yang are two complementary, interdependent principles or phases alternating in space and time; they are emblems evoking the harmonious interplay of all pairs of opposites in the universe.

First conceived by musicians, astronomers, or diviners and then propagated by a school that came to be named after them, Yin and Yang became the common stock of all Chinese philosophy. The Taoist treatise *Huai-nan-tzu* (book of "The Master Huai-nan") describes how the one "Primordial Breath" (*yüan ch'i*) split into the light ethereal Yang breath, which formed Heaven; and the heavier, cruder Yin breath, which formed Earth. The diversifications and interactions of Yin and Yang produced the Ten Thousand Beings.

> The warm breath of Yang accumulated to produce fire, the essence of which formed the sun. The cold breath of Yin accumulated to produce water, the essence of which became the moon.

The idea of ch'i. Yin and Yang are often referred to as two "breaths" (*ch'i*). *Ch'i* means air, breath, or vapour —originally the vapour arising from cooking cereals. It also came to mean a cosmic energy. The Primordial Breath is a name of the chaos (state of Unity) in which the original life force is not yet diversified into the phases that the concepts Yin and Yang describe.

Cosmic and active energies

Every man has a portion of this primordial life-force allotted to him at birth, and his task is not to dissipate it through the activity of his senses but to strengthen, control, and increase it in order to live out his full span of life.

The idea of wu-hsing. Another important set of notions associated with the same school of Yin-Yang are the "five agents" or "phases" (*wu-hsing*) or "powers" (*wu-te*): water, fire, wood, metal, earth. They are also "breaths" (*i.e.*, active energies), the idea of which enabled the philosophers to construct a coherent system of correspondences and participations linking all phenomena of the macrocosm and the microcosm. Associated with spatial directions, seasons of the year, colours, musical notes, animals, and other aspects of nature, they also correspond, in the human body, to the five inner organs. The Taoist techniques of longevity are grounded in these correspondences. The idea behind such techniques was that of nourishing the inner organs with the essences corresponding to their respective phases and during the season dominated by the latter.

Eclectic philosophical thought. *Yang Chu and the Lieh-tzu.* Yang Chu (*c.* 400 BC) is representative of the early pre-Taoist recluses, "those who hid themselves" (*yin-shih*), who, in the *Analects* of Confucius, ridiculed Confucius' zeal to improve society. Yang Chu held that each individual should value his own life above all else, despise wealth and power, and not agree to sacrifice even a single hair of his head to benefit the whole world. The scattered sayings of Yang Chu in pre-Han texts are much less hedonistic than his doctrine as it is presented in the *Lieh-tzu* (book of "Master Lieh").

Lieh-tzu was a legendary Taoist master whom Chuang-tzu described as being able to "ride the wind and go soaring around with cool and breezy skill." In many old legends Lieh-tzu is the paragon of the spiritual traveller. The text named after him (of uncertain date) presents a philosophy that views natural changes and human activities as wholly mechanistic in their operation; neither human effort nor divine destiny can change the course of things.

Kuan-tzu and Huai-nan-tzu. In the several Taoist chapters of the *Kuan-tzu* (book of "Master Kuan"), another text of uncertain date, emphasis is placed on "the art of the heart (mind)"; the heart governs the body as the chief governs the state. If the organs and senses submit to it, the heart can achieve a desirelessness and emptiness that make it a pure receptacle of the "heart inside the heart," a new soul that is the indwelling Tao.

The *Huai-nan-tzu* is a compilation of essays written by different learned magicians (*fang-shih*) at the court of their patron, the Prince of Huai-nan. Although lacking in unity, it is a compendium of the knowledge of the time that had been neglected by the less speculative scholars of the new state Confucianism. The *Huai-nan-tzu* discusses the most elaborate cosmology up to that time, the position of man in the macrocosm, the ordering of society, and the ideal of personal sainthood.

Neo-Taoism. During the 3rd and 4th centuries AD, in the political and social chaos that followed the collapse of the Han dynasty, a new philosophy took shape. Termed "obscure learning" (*hsüan hsüeh*), it oscillated between "nihilistic revolt" and "mystical escapism." The word *hsüan*, meaning "dark," "abstruse," "mysterious," alludes to the first chapter of the *Tao-te Ching*, in which the Tao is described as "*hsüan* of the *hsüan*," "mystery of mysteries."

Nihilistic revolt and mystical escapism

Ho Yen (died AD 249) and Wang Pi (flourished 3rd century AD) wrote commentaries to the *Tao-te Ching* and the *I Ching* but honoured Confucius (who "embodied Not-Being" and dealt with Being) above Lao-tzu and Chuang-tzu (who constantly spoke of Not-Being because they were still entangled in Being). The commentators Hsiang Hsiu (flourished AD 250) and Kuo Hsiang (died AD 312) also criticized Chuang-tzu in their commentary (the *Chuang-tzu chu*, also called the Kuo-Hsiang commentary), but they saw no basic difference between the doctrine of the "great Sage" Confucius and the Taoist philosophers, and they used Taoist thought to reinterpret the sayings of Confucius. Their real problem was to reconcile the opposing claims of a life of public ser-

vice and the hermit's quietism. Their type of sage lived in the world and out of it at the same time; "although he moves up and down through ten thousand critical decisions, he remains tranquil and self-possessed" (Kuo Hsiang). Others compared the rhythm of office holding and retirement to the "naturalness" (*tzu-jan*) of the coming and going of the seasons.

"Naturalness" was the foremost law and ideal of the Neo-Taoists. The "thusness-in-itself" combined for them the ideas of unspoiled nature, spontaneity, and the Tao. The so-called Seven Sages of the Bamboo Grove, among them Hsi K'ang (flourished 3rd century) and Yüan Chi (flourished 3rd century), tried to avoid the dangers of the political scene and worshipped "naturalness" with wine, poetry, and music.

RELIGIOUS TAOISM

Long life and the vital force were common concerns of all Taoist traditions and inseparable from saintliness. The original state of the infant or even the embryonic state in which the vital force is still perfectly concentrated and undiminished was the ideal of the mystics, who sought it in trance, as well as of the religious Taoists, who developed countless formulas and practices to restore the infant's complete Yang energy to the physical body. This effort, however, made them shift away from the classical ideal of adaptation to the course of nature and attempt a reversal of the natural flow toward death. Thus, the ideal of Yin and Yang as complementary, which implied the alternation of life and death, changed into a desire for the victory of life-giving Yang over Yin.

Spirit and matter as a continuum of vital energies

Longevity and immortality. For the ancient Chinese in general, spirit and matter formed a continuum of more or less rarefied or crude vital energies, deriving from Heaven and Earth. Man was believed to be composed of several kinds of vital energies, to which the Western dichotomy of spiritual and material is not particularly relevant. The techniques of longevity were intended to prevent the scattering of these energies, which would result in death. They also sought to refine the coarse, perishable energies and transform the heavy mortal body into a light immortal one.

The Chinese ideogram for "immortal" (*hsien*) depicts a man and a mountain, suggesting a hermit; the older form of *hsien*, however, shows a man dancing around, flapping his sleeves like wings. To become immortal is to be "transformed into a feathered being." This image comes from the mythology of eastern Chinese tribes who claimed bird ancestors, worshipped bird deities, and held religious rites with bird dances performed on stilts. The affinity of the Taoist immortals to birds (crane, phoenix, magpie, stork, or raven) is a persistent theme in iconography and legend.

There are many categories of immortals. The highest are those who "ascend to heaven in broad daylight." There are also those who live in terrestrial paradises (on holy mountains or islands) for centuries without growing old and later appear disguised in this world to transmit their immortality formulas and magical powers to worthy adepts. Lower immortals do not reach paradise before dying an apparent death (*shih chieh*), leaving their sandals or their canes in the coffins to take on the appearance of their corpses.

Techniques of longevity. *Hygiene techniques.* The detailed correspondence between the human microcosm and the macrocosm are at the basis of all techniques to "nourish the vital force" (*yang-hsing*). The communication between the five inner organs (lungs corresponding to metal, heart to fire, spleen to the earth, liver to wood, and kidneys to water), all other body organs, and the outer world proceeds through the orifices (nose for the lungs, ears for the heart, mouth for the spleen, eyes for the liver, and the lower orifices for the kidneys). All these orifices are passageways for entry and exit of vital forces and have to be closely guarded. Because the orifices are sense organs and desires result in loss of vital force, the senses have to be carefully kept in balance lest disease be caused through overindulgence in any one of the corresponding desires.

Abstinence from certain foods and practice of gymnastics

All dietary regimens are intended to nourish the respective organs in right proportions with foods and medicinal herbs containing the energy (*ch'i*) corresponding in quality to their respective elements. A preliminary step in diet is complete abstinence from all cereals in order to starve and kill the "three worms," or the "three corpses," which are maleficent demons inside the body that work to hasten its decomposition.

In order to make all energies in the body reach their proper place and to maintain a continuous circulatory process, the adept practices gymnastics called "to conduct (the breath) and to stretch" (or to attract it to its proper place, *tao yin*). Chuang-tzu mentions (and ridicules) the *tao yin* adepts who practice "bear hangings and bird stretchings" in order to render their bodies flexible and make the energy circulate everywhere. Gymnastics as well as massage assist the circulation of breaths; they eliminate any internal obstructions, which can cause disease. They untie internal knots so that everything inside the organism can communicate freely.

Respiratory techniques and meditation. Chuang-tzu stated that ordinary people breathe through the throat but the saint breathes through the whole body, starting from the heels. The Taoist breathes not only atmospheric air but solar, lunar, and stellar emanations. He absorbs the characteristic energies of the five directions, guiding the green emanation of the east to the liver, the red emanation of the south to the heart, and so on. The *Ling pao wu fu Ching*, or "Classic of the Five Sacred Talismans" (a revealed sacred text of the 3rd or 4th century AD), gives a formula for absorbing the emanations of the sun at noon (the peak moment of Yang) and of the moon at midnight (the peak of Yin). Others inhale the emanations of dawn in spring, of noon in summer, of dusk in autumn, and so on. Others warn against the practice of this discipline in the latter part of the day, because the universe exhales dead air from noon to midnight. Another method taught how to "feed on air" by retaining breath and conducting it throughout the body. One who could hold his breath for the time of 1,000 respirations would become immortal.

About the time of the T'ang period (618–907) ordinary respiration (*wai ch'i*) was replaced by the respiration of inner breath (*nei ch'i*), or "embryonic breathing" (*t'ai hsi*). This inner breath was viewed as man's share of the primordial life breath contained in the lower of the three "cinnabar [mercuric sulfide] fields" of the body (centres located in the head, chest, and abdomen). This life breath is conducted, in a closed circuit like that of the embryo, through the body and directed by means of the "inner sight" (*nei-kuan*), an inward-turned vision of the eyes (considered a source of light). In case of sickness, the inner breath is conducted to the diseased organ and heals it.

The *T'ai-p'ing Ching*, or "Classic of the Great Peace," the first sacred text of religious Taoism (3rd century AD), mentions the ability of the interior vision to discern the gods residing in the five organs. Books like the two *Huang-t'ing Ching* ("Classic of the Yellow Court"), from the 3rd century AD, owe their popularity to the enumeration of thousands of gods in all parts of the body, the same gods as those residing in the macrocosm. Meditation establishes communication with these gods, either by making the external gods descend and visit their organs in the body and fortify it or by externalizing the inner gods and disposing them in a mandala (a symbol of the universe) around oneself. The latter practice is a preparation of religious ceremonies and recitation of sacred texts.

In many Taoist physiological practices an influence of Indian Yoga (a system of psychological and physiological meditation) is perceptible. Borrowings from Tantrism can be found in Taoist texts as early as the 7th century.

The "Way of the Yin"

Sexual techniques. The techniques of the "inner chamber" (*fang chung*) go back to antiquity. Their aim was to secure vitality, longevity, and male progeny. Specifically Taoist practices were the "Way of the Yin" of the immortal Yung-ch'eng (1st century AD), a technique "to make the semen return and repair the brain" (*huan ching pu*

nao). This technique consisted in the prevention of ejaculation during the sexual act and was thought to make the semen (*ching*, a potent mixture of all physical energies) circulate—mixed with breath—through the body from the lower to the upper "cinnabar field," there to vitalize the brain. Immortals are often depicted with a huge skull that is the storeroom of their Yang energy. Another idea was to blend seminal essence with breath in the "lower cinnabar field" and there to form the "mysterious embryo" (*hsüan t'ai*) of the new immortal body.

For the "repairing of the brain," the male adept also needed to absorb as much as possible of female Yin essence (and vice versa). It was therefore desirable to have intercourse with a succession of partners. This led to the much-decried communal "union of breaths" (*ho ch'i*), a highly ritualized ceremony that might have resembled more a group ordeal than an orgy.

The abstention from ejaculation exists equally in Tantric practices, which were known to the Taoists since T'ang times or earlier. As in Tantrism, the sexual terminology refers also to mental operations because only thought processes can make semen and breath circulate and "marry" to thus create the immortal man.

Alchemical techniques. Dietary and breathing techniques only prolong life so as to give time for the preparation of the elixir of immortality, which was composed of cinnabar (*tan*, for its red, or Yang, colour and its transmutability) and gold (*chin*, for its incorruptibility).

Alchemy evolved early in connection with metallurgy. The patron of alchemy was the Yellow Emperor (Huang Ti), who ascended to Heaven after the casting of a sacred metal crucible. This legend is alluded to in the alchemical recommendations given by Li Shao-chün, a 2nd-century-BC alchemist, to the emperor Han Wu Ti (*c.* 133 BC), in the *Shih-chi* (composed *c.* 85 BC), the earliest known reference to alchemy in any literature of East or West.

The "elixir of immortality"

Later, Taoist alchemists sought to produce liquid gold and pure (nine times transmuted) cinnabar, or a combination of both in an "elixir of immortality." "If these two substances are eaten," stated the alchemist Ko Hung (flourished 4th century AD), "they bring about a sublimation in the human body." Ko Hung's *Pao-p'u-tzu*, or "He Who Holds to Simplicity," distinguishes the art of "gold and cinnabar," that is, the compounding of the elixir, from the art of the "yellow and white," that is, the transmutation of baser metals into silver and gold. All of these practices were supposed to take place in the solitude of sacred mountains and were to be accompanied by preliminary fasting, purifications, and sacrifices to the patron gods. Also important was the right time for the operation and the notion of accelerating natural processes in time. Although Ko Hung, the most sober, inquisitive, and well-informed Taoist writer on alchemy at that time, was himself too poor to afford the ingredients of his formulas, he wrote of operative alchemy and laboratory work (*wai tan*, external alchemy).

This is not mentioned in the probably earlier and most popular alchemical treatise, the *Chou-i ts'an t'ung ch'i* by Wei Po-yang (of uncertain date, possibly 2nd century AD), which reads like the application of the doctrines of the *I Ching* to the procedures and principles of alchemy. Written in a highly abstract language, it has later been interpreted in terms of internal alchemy (*nei tan*).

The term internal alchemy refers to physiological or mental techniques of immortality that employ the language of the laboratory but in actuality refer to the human body. The human body becomes the alchemical crucible, the microcosm in which breath and semen, the "true" lead and the "true" mercury, circulate and fuse. The Yin and Yang energies rise and descend in the two main energy conduits *jen* and *tu* (already mentioned by Chuang-tzu), the two passageways between the upper and lower parts of the body. Embryonic respiration is often described in alchemical language, and the frequently sexual images of the "marriage" of Yin and Yang have led to interpretations (denied by some commentators) of *nei tan* treatises in terms of sexual practices.

The aim of *nei tan* was to reverse the normal vital processes leading to old age and death. Instead of externalizing vital energies, especially sexual energies, the Yin and Yang forces of the body are made to unite in a hierogamy (holy marriage) and to create the "elixir of life," the "golden flower," the "mysterious embryo," the "infant"; *i.e.*, the new immortal being that eventually leaves the mortal body through the occiput (the top of the skull) and ascends to Heaven.

Use of charms and talismans

Magical techniques. Charms and talismans were used in all sects. Magical diagrams written on paper were burned and swallowed for healing. In order to penetrate into the mountains in search of medicinal herbs and divine revelations, magic formulas and gestures as well as talismans and mirrors were needed for protection against demons and for communication with immortals. The *fu* ("talisman") was written on bamboo, metal, or paper and broken into halves for each partner of a contract between the adept and his human or divine master.

Communal religious Taoism. Practice of the above techniques of immortality was restricted to the affluent and educated few. In the rural communities of the two Han dynasties (206 BC–AD 220) emerged the popular Taoist religion.

The beliefs and institutions of the popular Taoist sects took form when Buddhism was beginning to spread in China, but scholars now tend to minimize any causal connection. The pantheon, ritual, and moral codes, as well as the theocratic organization of the early Taoist Church, bear much closer similarity to the Imperial and Confucian ideologies than to Buddhism. The new foreign religion, however, was at first mistaken for another form of Taoism because Lao-tzu was supposed to have preached to the "barbarians" after his legendary departure from China to the West. On this theory of the "conversion of the barbarians" (*hua hu*) focussed most of the subsequent polemics between the two religions. Buddhism soon became a formidable rival, conscious of its originality; and the Taoists endeavoured to limit its influence by polemics and by extensive (though more formal than essential) borrowings, imitating its institutions (*e.g.*, monasticism and celibacy) and the style of its scriptures. Folk Taoism came to be deeply influenced by the Buddhist beliefs in reincarnation and punishment in hells.

The rise and influence of Taoist sects

Religious communities. Popular Taoist communities can be traced back to the T'ai-p'ing ("Great Peace") movement and the "Way of the Celestial Masters" (*Tien shih Tao*, or "Way of the Five Pecks of Rice," *Wu-tou-mi Tao*), which propagated popular Taoist cults and formed village communities in the 2nd century AD. The legendary first celestial master, Chang Tao-ling, was a magician and faith healer. His descendants evolved into a hierarchy of priests ("libationers," *chi-chiu*) who assumed temporal as well as religious authority in their parishes. They were exorcists who treated the sick by ritual and prayer, using charmed water and talismans. Because disease was believed to be caused by sin, the sick were sent to "pure [or calm] houses" in order to repent and to recite the *Tao-te Ching*, which was explained to them in simple terms of physiology and moral conduct. They also were supposed to make amends by doing public works, such as road building. Taoist temples played an important social role in the villages throughout Chinese history as focuses of community organization.

Morality and the system of merits. Insistence on moral teachings and on seasonal collective ceremonies was a new accent. The moral code was based on the *Tao-te Ching*, of which the T'ai Shang Lao-chün (that is, Lao-tzu deified) became the divine revealer. Moral conduct was rewarded with health and long life; immorality caused sickness, premature death, and, according to later texts, suffering in hell. Numerous deities came to be the "directors of destiny" (*ssu ming*), record keepers of man's virtues and vices. A complex heavenly bureaucracy received annual reports on the deeds of the family or the individual from the god of the hearth (Tsao Chün) at New Year or from the malevolent "three worms" residing in the human body. A proliferating literature of merits and demerits, including schedules of points for good and bad deeds (the most famous being the *T'ai shang kan-ying p'ien*, or "Tract on the Most Exalted One on

Actions and Retributions," probably from the 11th century), probably shaped the minds of the masses more than any other religious texts in China.

Incipient messianic and millenarian ideals

Theocracy and messianism. The sect of the Celestial Masters developed into a lasting Taoist church. Other, dissident, movements shaped the idea of the Taoist saint as ruler into an alternative to the Imperial heavenly mandate. "The Way of the Great Peace" (T'ai-p'ing Tao) was the first known Chinese movement to attempt to fuse a religious organization with an Imperial type of administration into one messianic empire. Nostalgia for this ideal was a recurrent feature of Chinese peasant rebellions. Political distress or, more often, economic distress again and again sparked a latent messianism and millennialism. Charms played an important role in these movements: as badges of office, as protection against demons, and as the sign by which the expected divine ruler (the deified Lao-tzu or his emissaries) would distinguish the virtuous "chosen people" (*chung min*) from the villains and reward them with official appointments in the heavenly bureaucracy that would come down to Earth. The founders of the T'ang (618–907) and Ming (1368–1644) dynasties apparently utilized Taoist prophecies and magic to secure popular support. Dissident movements continued to be animated by the Taoist vision of a perfect government; some of their ideas live on in syncretistic religions and secret societies.

Sacred texts and revelation. All sects had an esoteric character; the Taoist priesthood could be entered only through an initiation that culminated in the transmission of the sect's sacred scriptures. Sacred texts served to procure communication with the gods and to ward off demons. There were manuals for immortality practices (arts that had to be learned orally from a master), lists of names of deities used in meditation and liturgy, and divine revelations. The mere perusal of a text without initiation into its esoteric meaning was worthless if not sacrilegious or fatal. The more important a text, the more carefully it was hidden from the profane and the more rarely revealed and transmitted.

Divine revelations were transmissions of a teaching the gods themselves had received from higher deities. Only the highest, the Yüan Shih T'ien Tsun ("Celestial Worthy of the Original Beginning"), could face the power and understand the meaning of the primordial revelation, the first and purest emanation of undifferentiated unity. In diminished and adapted form he passed it on until it reached the lower gods, who can reveal some of it to worthy adepts. In all ages adepts have gone to the mountains in search of texts hidden by the immortals in grottoes accessible only to the initiated. Many of the most important scriptures were revealed in trance. The Mao Shan sect is famous for its mediumistic seances, to which is owed the *Shang-ch'ing* (another name for the Mao Shan sect) writings, such as the *Huang-t'ing Ching* ("Classic of the Yellow Court"). The Mao Shan revelations were written with a brush by mediums versed in calligraphy. The immortals appearing at such seances used to identify themselves, relate their biographies, and specify their heavenly rank. Their hagiographies (biographies of saints) compiled from the records of medium cults at different periods became a prolific and confused lore that gradually constituted a pantheon. Thus, mediumism is one of the reasons for the unsystematic and contradictory character of the Taoist pantheon.

Preparatory conditions for the reception of revelations

The pantheon and the Taoist canon. The Chinese gods do not reign over the world. They act as administrators and control the harmonious functioning of the universe of which they themselves are a part. The Taoist pantheon is thus constructed as a vast other-worldly bureaucratic administration. Many names of gods are really titles of functions (*e.g.*, "Director of Destiny"), the holders of which are deified human beings who change their functions as they advance in grade. Another category of gods, introduced by the Ling Pao, or "Sacred Treasure," sect (5th century), are the celestial worthies (T'sien Tsun), personifications of aspects of the Tao. The foremost of these highest and remotest gods are the Taoist triad Yüan Shih T'ien Tsun, his disciple T'ai Shang Tao-chün ("Most Exalted Lord of the Tao"), and T'ai Shang Lao-chün ("Most Exalted Lord Lao"), who is identified with Lao-tzu. There are as many celestial worthies as there are aspects of the Tao that can become manifest to the believer; some texts enumerate hundreds. The Taoist master T'ao Hung-ching (flourished 5th–6th centuries), among others, attempted to codify the sprawling multiplicity in his *Chart of the Ranks and Functions of the Gods* (*Shen ling wei yeh t'u*).

The concept of a canon (accepted texts) of all Taoist scriptures goes back to the Taoist ritualist and reformer Lu Hsiu-ching (AD 406–477), and it bears witness to a cohesiveness existing among different schools of Taoism. Lu Hsiu-ching mentioned 1,228 scriptures—of which 1,090 were known in the world and 138 were "still in heaven," that is, not yet revealed—and divided them into three *tung*, in imitation of the early Buddhist canon. The root meaning of the word *tung* is "cave," but it also means "to communicate, to understand the mysteries." The term thus evokes the mysterious mountain grottoes, real or imaginary, harbouring hidden revelations and the dealings with the sacred powers that can initiate one into their meaning.

Priesthood and monasticism. The "libationers" of the sect of the Celestial Masters were a married, hereditary priesthood. They lived in the community they served and performed exorcisms and increasingly complex rituals. The present Chang family of the Celestial Masters claims descent from Chang Tao-ling, founder of the sect during Han times (since Sung times, it has been called "Way of Orthodox Unity," Cheng-i Tao). The libationers' title (*chi-chiu*) is still used in the Cheng-i sect of Fukien and Taiwan, pointing to an unbroken tradition. Organized religious Taoism never had a permanent central authority. There was no institutional cohesion among the sects.

Taoist monasticism as a reaction to Buddhism

Taoist monasticism is a later element but no less important than the priesthood. The Mao Shan sect formed religious communities of families in phalanstery-like, communal establishments called *kuan*. This way of life was created essentially in order to facilitate everyday observance of Taoist meditation, liturgy, hygiene, and other matters. It was also a reaction to the Buddhists' establishment of monasteries, with which the Mao Shan sect seems to have had close relations. The monk Sung Wen-ming (flourished 6th century) may have been the first to advocate celibacy. This was encouraged by two imperial edicts (972 and 1009) forbidding married monks to live in the tax-exempt monasteries. Since the Yüan dynasty (1279–1368), the monasticism of the Ch'üan Chen ("Perfect Realization") sect has been a conspicuous feature of the Taoist tradition. The founder of this sect, Wang Che (flourished 12th century), advocated the syncretism of the three religions and placed emphasis on meditation as a means to return to one's original nature and thus to prolong life. Monasteries were not subject to any higher jurisdiction except for their recognition by the government. The Ch'üan Chen sect is usually called the Northern School of Taoism.

Liturgical practices and ceremonies. The most important, as well as one of the oldest, liturgies is the *chiao*, a "pure" ritual (in contrast to rites for the dead) that renews the community through communication with the gods. A *chiao* of three, five, or seven days is still celebrated in Taiwanese villages, usually once every three years, with a more extensive ritual that is celebrated once in a generation and is combined with a restoration of the village temple. There are other *chiao* rites for the ordination of priests, for the birthdays of gods, for warding off calamities, and for other occasions. All are celebrated on two levels: a village feast outside the temple and an esoteric liturgy (in Classical Chinese) inside the closed temple. The inner *chiao* ritual's communication with the highest gods (the T'ien Tsun triad) is modelled on the ritual of Imperial court audiences with the Son of Heaven.

Ceremonial objects

An incense burner, the central object in the temple, is essential to every Taoist rite. A mystical incense burner is "lighted" inside the body of the main priest, whose meditation transforms him into a mediator with the divine



Taoist priest burning incense in honour of two minor deities.
Dmitri Kessel, *The World's Great Religions* © Time Inc.

who makes possible the efficacy of the *chiao* rite. Incense also plays a role in the ecstatic visit to Heaven performed inside a microcosmic sacred area marked by five buckets of rice. Several written documents are used, among them a "memorial to the throne" (*chang*), which is burnt ("sent off to Heaven") to announce the purpose and report the performance of the liturgy. Other important ritual objects are swords, lamps, flags, masts, charmed water, musical instruments, masks, vestments, statues, picture scrolls, and sacred texts.

In medieval China, among the collective rites called *chai*, was the "Mud and Soot [T'u T'an] *chai*," a frenzied public rite for the atonement of sins. In modern times, the term *chai* is reserved for funeral rites for individuals. Aside from the *chiao* and the *chai*, which belong simultaneously to both religions, it is difficult to distinguish between the festivals of folk religion and Taoism.

Communal folk Taoism (shen chiao). Popular, or folk, religion is not a separate religious tradition but the wholly unorganized undercurrent of Chinese religious culture from the earliest times, shared by all strata of society. The Chinese have no single name for it; it may be called the religion of the gods, or spirits (*shen chiao*). The deities of the popular pantheon come from all traditions. What the deities have in common is that in *shen chiao* they are all gods intimately involved in everyday life as givers of blessings or bringers of calamities. Every object or activity of daily life has its presiding spirit that has to be consulted and feasted or appeased and driven off, especially at all special occasions in the life of the family or the community. The person primarily involved in the practice of *shen chiao* in modern times is the *fa shih* (magician). For the orthodox Taoist priests the *shen chiao* rites are the "little rites"; the *chiao* rituals, the exclusive function of the Taoist priests, are the "great rites." Both kinds of priests—the orthodox and the magicians—operate on different occasions in the same temples and are consulted for the family rites of burial, birth, marriage, house construction, and business affairs.

Major exorcism rites (*e.g.*, purification of haunted houses and treatment of the sick or mentally deranged) are performed by the orthodox Taoist priests, who, being ordained into the ranks of the *shen*, have power over the demons with whom they are on an equal footing. The *fa shih* priest's specific function is the manipulation of possessed mediums (specially gifted lay persons). The medium puts himself into a trance in which he becomes the mouthpiece of a deity (or a deceased relative) giving medical, personal, or business advice that is interpreted by the *fa shih*. Professional mediums attached to a temple or a private cult lacerate themselves in trances. This is considered to be a vicarious atonement for the community during the great feasts. A different form of mediumistic communication among lay people is automatic writing, either with a brush on paper or with a stick on sand.

The functions of the folk priest (or magician)

Contemporary religious Taoism. Religious Taoism lives on wherever its basis, traditional Chinese culture, survives. The best example is in present-day Taiwan, where the religion has had, since about 1960, a veritable renaissance—manifested in the restoration and construction of costly temples and in the frequency of lavish festivals. The Taiwanese orthodox Taoist priests preserve a vast liturgical heritage and still receive their rank in the Taoist church from the Celestial Master. The 63rd Celestial Master, Chang En-pu, fled to Taiwan in 1949. After his death in 1970 he was succeeded by the 64th Celestial Master, Chang Yüan-hsien. On a smaller scale, similar active Taoist communities exist in the Chinese quarters of Pinang (Malaysia; Penang), Singapore, and Bangkok (Thailand).

Almost nothing is known about the fate of Taoism on the Chinese mainland. The Po Yün Kuan ("White Cloud Monastery") in Peking and some other temples elsewhere were open and inhabited by monks until the mid-1960s. In 1958 it was officially claimed that 30,000 Taoist priests were still active in different parts of the country. At the end of 1971, soldiers were seen at the Po Yün Kuan, apparently quartered there.

Not only on the mainland but throughout the Chinese culture area, the surviving Taoist traditions can be expected to lose their traditional place in society as their gods and practices are increasingly challenged by the advance of Western education and ways of life.

Other forms of Taoist thought and influence. *Secret societies.* Politically dissident messianic movements have existed and developed separately from the established Taoist church from the very beginning (2nd century AD). Their leaders were priest-shamans, similar to the modern *fa shih* priests of folk Taoism. Their followers were the semiliterate or illiterate classes socially below the tradition of orthodox Taoism, and their organization was similar to that of the syncretistic religions and of modern secret societies. Although the secret societies have had no organizational contact with the Taoist tradition for centuries, their religious beliefs, practices, and symbols contain some Taoist elements, such as initiation rites, worship of Taoist deities, mediumism, and the use of charms and amulets for invulnerability. These influences reached them either directly or through popular religion.

Relationship of secret societies to Taoism

The syncretistic religions (*hui-men tao*), such as the I Kuan Tao, or "Way of Pervading Unity," and at least 40 other named organizations that have no organizational ties with religious Taoism, have to be distinguished from the secret societies. In their practice of mediumism and in many of their beliefs they participate in a popular tradition nourished from Taoist sources.

Taoist aspects of traditional Chinese medicine. Taoist physiological techniques have, in themselves, no devotional character. They have the same preoccupations as physicians: to preserve health and to prolong physical life. Medicine developed independently from about the 1st century AD, but many Taoist faith healers and hygienists added to medical knowledge.

The earliest surviving medical book, the *Huang Ti nei Ching*, or "The Yellow Emperor's Esoteric Classic" (3rd century BC?), presents itself as the teachings of a legendary Celestial Master addressed to the Yellow Emperor.

Experiments with minerals, plants, and animal substances, inspired to some extent by Taoist dietetics and by the search for the elixir of life, resulted in the 52 chapters of pharmacopoeia called *Pen-ts'ao kang-mu*, or "Great Pharmacopoeia" (16th century).

Taoism in Japan. There was never an attempt to implant a Taoist religion officially in Japan, but a random choice of Taoist beliefs and customs have, at various ages, been adopted and transformed at the Japanese

court, in the temples, and among the people. Records from the early 7th century contain traces of Taoism, which was appreciated chiefly for its magical claims. The "masters of Yin and Yang" (*ommyō-ji*), a caste of diviners learned in the *I Ching*, Chinese astrology, and occult sciences who assumed importance at court in the Heian period (8th–12th centuries), probably were responsible for the introduction of Taoist practices, such as the Keng-shen (Japanese Kōshin) vigil and the observance of directional taboos (*katatagae*). In the 8th century, disputations were held at court over Buddhism and the philosophy of Lao-tzu and Chuang-tzu. The *Pao-p'u-tzu* was known, and Kōbō Daishi, the founder of Shingon Buddhism, reported (in 797) on Taoist physiological practices and beliefs in immortals. Buddhist (Shingon and Tendai) ascetics, wandering healers, and mountain hermits known as *yamabushi* probably came closest to Taoism in their techniques for prolonging life (abstinence from grains, etc.) and their magical arts (exorcisms, sword dance) and objects (mirrors, charms), which must have reached them through the Tantric elements in Shingon. Taoist mysticism lives on in that it has influenced the two Chinese Zen schools of Lin-chi (Rinzai) and Ts'ao-tung (Sōtō), introduced in the 12th and 13th centuries and still active in Japan. Popular Taoist moral tracts were printed and widely diffused in the Tokugawa period. Modern Japanese scholarship on Taoism (Dōkyō) ranks very high in the world.

Introduction of Taoist practices to Japan

TAOISM AND OTHER RELIGIONS

Confucianism and Buddhism. Confucianism is concerned with human society and the social responsibilities of its members; Taoism emphasizes nature and what is natural and spontaneous in man. The two traditions, "within society" and "beyond society," balance and complement each other. This classic definition is generally correct concerning orthodox Han Confucianism; it neglects some aspects of Confucian thought, such as the speculations on the *I Ching*, that is considered to be among the Confucian Classics and the prophetic occult (*ch'an-wei*) commentaries to the classics. As far as Taoism is concerned, this definition neglects the social thought of the Taoist philosophers and the political aspects of Taoist religion. Chinese Buddhism has been viewed not as a Sinified Indian religion but as flowers on the tree of Chinese religions that blossomed under Indian stimulus and that basically maintained their Chinese character. Indian Buddhist thought was understood in China, but the more accurate and the less Sinified the translations became in the T'ang dynasty, the less appeal they had for the Chinese mind and the less stimulus they provided for spiritual creativity.

Relationship of Taoist and Buddhist concepts

The idea of Buddhist Mādhyamika (Middle View) philosophy that there is no difference between worldly existence and otherworldly salvation; *i.e.*, that *saṃsara* (the round of mortality) and Nirvāṇa (the state of salvation) coincide, led to the concept of the inborn Buddha-nature and of salvation not as an effort to become a Buddha after many lives but as an illuminating realization—here and now—of the Buddha-nature ever-present within. At this point Chinese Buddhism met with the Taoist ideal of "non-activity" and of the union with the Tao. The result of this process of assimilation was Chinese Ch'an (Zen) Buddhism, a revival of Taoist mysticism in new form.

Ch'an Buddhism deeply influenced Neo-Confucianism, the renaissance of Confucian philosophy in Sung times (960–1279), which in Chinese is called "Learning of the Tao" (Tao Hsüeh). In this movement Confucianism acquired a universal dimension beyond a concern for society. Neo-Confucian thought often seems as Taoist as Neo-Taoist philosophy and literature seem Confucian. What is called Neo-Taoism was a world view accepted by disillusioned officials who had to cope with political chaos, whereas Neo-Confucianism was the attempt by state-supported intellectuals to make the official ideology more attractive by incorporating Taoist (and Buddhist) thought. Viewed from their share in both traditions, it appears almost an historical accident that the one movement was named Taoism, the other Confucianism.

As early as the T'ang dynasty, there are traces of the syncretism of the "Three Religions" (San Chiao), which became a popular movement in Sung and Ming China. A mixture of Confucian ethics, the Taoist system of merits, and the Buddhist concept of reincarnation produced such "books on goodness" (*shan-shu*) as the *Kan ying p'ien* ("Tract on Actions and Retributions"). The school of the "three Religions" was rejected by most Confucians and Buddhists but received wide support in Taoist circles. Many Taoist masters of those periods transmitted *nei tan* and other techniques of inner cultivation to their disciples while at the same time preaching the moralism of the "Three Religions" to outsiders.

Other Asian religions. The affinities of Taoism with other Asian religions are numerous. If one distinguishes between universal religions of salvation, such as Buddhism and Islām, and the older, more culture-bound, such as Japanese Shintō and Hinduism, Taoism undoubtedly belongs to the second category.

The fact that no record of Shintō antedates the introduction of Chinese script makes it difficult to distinguish between Taoist affinities and influences on Shintō features, such as the cult of holy mountains, the representation of the human soul as a bird, bird dances, the representation of the world of the dead as a paradisiac country of immortality, and the concept of the vital force (*tama*, in objects as well as in man). Like Taoism, Shintō is the religion of the village community.

Western mysticism and religions. The similarity of mysticism in all religions points to the fact that there is only one Inner Way, the experience of which is expressed differently in the respective cultural and religious environments. Lao-tzu's notion of "the One," which is not only primordial unity but the oneness underlying all phenomena, the point in which all contraries are reconciled, was spoken of by such Western mystics as Plotinus, a 3rd-century-AD Greek philosopher, and Nicholas of Cusa, a 15th-century French philosopher.

Similarities between Taoism and Western mystical thought

Taoism, as all other forms of Eastern mysticism, distinguishes itself from Western mysticism by its conscious techniques of mind and body designed to induce trance and to give access to mystical experience. These disciplines of learning to "sit in forgetfulness" are akin to Plotinus' concern to "be deaf to the sounds of the senses and keep the soul's faculty of apprehension one-pointed" and to the 16th-century Spanish mystic Teresa of Avila's state where "the soul is fully awake as regards God, but wholly asleep as regards things of this world and in respect of herself." Lao-tzu's strangely sober and abstract descriptions of ecstatic union with the Tao have been compared to the medieval German mystic Meister Eckeharts' "still desert of the Godhead" and his pupil Heinrich Suso's union of the essence of the soul with "the essence of Nothingness." One instance of Western physiological techniques is the Hesychasts, a sect of Greek Orthodox mystics on Mt. Athos in the 14th century who used respiratory practices and concentration on internal organs to prepare for the mental "Jesus prayer."

TAOISM AND CHINESE CULTURE

Taoist contribution to Chinese science. Scholars have stated that the Taoists initiated the sciences of chemistry, mineralogy, botany, zoology, and pharmaceutics in East Asia. This interest in science is considered a reflection of the Taoist emphasis on direct observation and experience of the nature of things, as opposed to Confucian reliance on the authority of tradition. Chuang-tzu declared that tradition tells what was good for a bygone age but not what is good for the present.

The Taoist secret of efficacy is to follow the nature of things; this does not imply scientific experimentation but rather a sensitivity and skill obtained by "minute concentration on the Tao running through natural objects of all kinds." This knowledge and skill cannot be handed down but is that which the men of old took with them when they died (*Chuang-tzu*). The image for it is the skill of the artisan admired by the Taoists in their numerous parables on wheelwrights, meatcutters, sword makers, carvers, animal tamers, and musicians.

Though extolling the intuitive comprehension and skillful handling of matter, the Taoists did not observe nature in the Western sense and rejected technology out of their aversion to the artificial. Any new idea or discovery in China was phrased as "what the old masters really meant." This ideology of rediscovery makes it hard to study the evolution of scientific thought. Some progress over the ages (for example, in alchemy) can be seen, but the Taoist contribution to Chinese science might be smaller than it has been assumed.

Taoist influence on literature and art. The Taoist conception of the craftsman's skill is already on the threshold of art. Like the meatcutter and the woodworker in *Chuang-tzu*, the artist has to feel his subject from the inside and see it with the eyes of the spirit in order to grasp the vital breath of what he is painting or putting into tunes, words, or calligraphy. As the first of the six principles presiding over painting, the painter and critic Hsieh Ho (*c.* AD 500) names "consonance with the vital breath" (*ch'i-yün*), which means that the artist should attune himself to the vital rhythm of his subject in order to be able to express the life and movement manifest in nature. If the *ch'i* in the subject is missed, the spark of Tao is absent from the work of art. A similar concept applies to Chinese poetry, one of the most characteristic themes of which is harmony with nature. Creative art in China is less compartmentalized than in the West. One man was often a painter, calligrapher, and poet. Underlying all artistic creation, even by Confucian artists, is the concept of "spontaneity" that was cherished especially by the Neo-Taoists (among them the greatest Chinese calligrapher, Wang Hsi-chih, flourished 4th century, who was a member of the Celestial Masters sect). Chinese art criticism and art philosophy are animated by Taoist views of the artist and of the world he expresses.

The Taoist view that the artist must see his subject with the eyes of the spirit

Both Confucianism and Buddhism have a limited understanding of art and value it only as an expression of basically nonartistic—ethical or religious—concerns. The Taoist pervasion of the great tradition dominated by Confucianism grew stronger the more the Confucian gentleman embraced all human endeavours, including the arts, for ends that were not purely moral. Through the practice of poetry, calligraphy, and painting, the Taoist understanding of nature and creativity became a necessary part of Confucian individual cultivation, especially with the rise of Neo-Confucianism in Sung times.

Distinctive religious elements of Taoism. The essential quest of Taoism is how to gain, preserve, and increase the vital force through a realization of harmony. On the basis of the Chinese concept of a concentric, well-ordered, and single universe, Confucianism built its wisdom of the order and functioning of human society in accordance with the Whole; and Taoism expressed the individual's as well as the group's need of harmony with the perceptible and the imperceptible Great Order (Tao) in all variations of magic, religion, and mysticism. To live in harmony with the universal mechanism without intervening (*wu-wei*) anywhere is to preserve one's energy.

The essential quest of Taoism

This preoccupation with the vital force is one reason for that lack of speculative thought distinctive of Taoism. Taoist mystics are poets, priests are technicians of the occult, and thinkers are specialists in alchemy, ritual, or pharmacognosy. There are no theologians in the Western sense and few philosophers in the Buddhist sense. This accounts for the Taoist weakness in competing with Confucians and Buddhists in debate. This also accounts for the fact that among the highly developed religions, Taoism is the one with the least rationalization, in which the communication of man with the sacred appears in its most immediate and unreflective form.

BIBLIOGRAPHY

General works: MAX KALTENMARK, *Lao tseu et le taoïsme* (1965; Eng. trans., 1969), a good general introduction to Taoist philosophy and religion; HOLMES WELCH, *The Parting of the Way: Lao Tzu and the Taoist Movement* (1957), a most readable interpretation of the *Tao-te Ching* and first coherent account of the Taoist movement; HENRI MASPERO, *Le Taoïsme et les religions chinoises,* rev. ed. (1971), a classic pioneer work on religious Taoism.

Philosophy: ARTHUR WALEY, *The Way and Its Power* (1934), a classic translation of the *Tao-te Ching* preceded by a good introduction on its place in Chinese thought; BURTON WATSON, *The Complete Works of Chuang Tzu* (1968), a very readable translation of the most important and most difficult text of Taoist mysticism; A.C. GRAHAM, *The Book of Lieh-tzu* (1961), a good translation and introduction to Taoist mysticism; MARCEL GRANET, *La Pensée chinoise* (1934), the classic work on the basic systems of classification in Chinese thought—difficult but highly enlightening; FUNG YU-LAN, *A History of Chinese Philosophy,* 2nd ed., 2 vol. (Eng. trans. 1952–53), a classic work with sections on Taoist thought; JOSEPH NEEDHAM, *Science and Civilization in China,* vol. 2, *History of Scientific Thought* (1956), a section on Taoist contribution to scientific thought (most thorough documentation and bibliography).

Religion: MAX KALTENMARK, *Le Lie-sien tchouan* (1953), a translation of the earliest extant Taoist hagiography with much information on mythology hidden in long notes; *History of Religions,* vol. 9, no. 2 and 3 (1969–70), issues devoted to the reports from the First International Conference on Taoist Studies, papers on the conference discussions, on Chuang-tzu, on Neo-Taoism, on Taoist antecedents in Buddhist thought, and on Taoist messianism; ANNA K. SEIDEL, *La Divinisation de Lao tseu dans le Taoïsme des Han* (1969), a study of the beginnings of the popular Taoist religion; KRISTOFER SCHIPPER, *L'Empereur Wou des Han dans la légende taoïste* (1965), a translation of a Taoist hagiographic novel and study of its ritual background in the Mao Shan sect; JAMES R. WARE, *Alchemy, Medicine and Religion in the China of A.D. 320: The Nei P'ien of Ko Hung* (1966), an imperfect but only complete translation of probably the most important text of religious Taoism; ARTHUR WALEY, *The Travels of an Alchemist* (1931), a translation of the contemporary account of the journey of the Taoist priest Ch'ang-ch'un from China to the Hindu Kush at the summons of Genghis Khan, with a good historical introduction; NATHAN SIVIN, *Chinese Alchemy: Preliminary Studies* (1968), an annotated translation of a treatise by a 6th-century alchemist—a most scholarly introduction on the study of Chinese alchemy; MICHEL SOYMIE, "Bibliographie du Taoïsme—études dans les langues occidentales," in *Études Taoïstes (Dôkyô kenkyû),* vol. 3–4 (1968–71), the first and only comprehensive bibliography of Western-language studies on Taoism and Chinese folk religion.

(A.K.Se.)

Taoism, History of

Taoism, with Buddhism and Confucianism one of the major spiritual "ways" of China, is said to be founded, according to tradition, by the sage Lao-tzu in the 6th century BC, and for more than 2,000 years it has influenced Chinese thought in every sphere—in religion, politics, the social system, and the arts and sciences. Further, it has provided the impetus for various changes in the historical development of China.

TAOISM IN CHINESE HISTORY

"Taoism" is commonly used to designate two distinct aspects of Chinese thought and belief. First, Taoism is viewed as a philosophy or form of mysticism, of which the first evidence is contained in texts of the 4th and 3rd centuries BC. In this regard it has continued throughout Chinese history to be a spiritual goal or diversion of the leisured and the learned class. Secondly, it is viewed as a religion that is highly organized, with a hereditary priesthood (of which the earliest original documents date from the 2nd century AD). As such, it began as a popular movement and soon extended its influence to all ranks and classes of the Chinese population. Though these two views of Taoism appear to be quite distinct, they in fact have many points of interconnection; the same person might participate in the communal ceremonial aspects of organized religious Taoism (Tao-chiao) as well as involve himself in solitary mystical abstraction. Many conceptual and linguistic links also exist between the classics of Taoist speculation and the liturgies and contemplative tracts of the organized religion.

The two common views about Taoism

Although philosophical Taoism (Tao-chia) has attracted much attention among scholars, investigation of the Taoist religion (Tao-chiao) is only just beginning. Western scholars generally have placed emphasis on certain practices or crafts that they suppose have been par-

ticularly "Taoist," notably the quest for physical immortality, breath control, techniques of sexual union, herbalism, dietetics, and alchemy. In such a view, though, as in the question of doctrine in general, there is some ambiguity as to what is specifically Taoist and what is simply Chinese. An adequate investigation and description of Taoism, for more than 2,000 years a crucial component of Chinese spiritual and cultural history, cannot be undertaken on a purely intellectual or doctrinal or, especially, technological basis.

Taoism has at all times been a social phenomenon, and this is reflected throughout Taoist literature. Religious Taoism has from the beginning had a distinctive organization—a hierarchy of priesthood and believers, linked by sworn bonds in spiritual and economic union. Developing and changing over the centuries, certain residues of this organizational structure still survive in Taoism in the 20th century, especially in Taiwan and Southeast Asia. It is precisely these institutional and economic aspects of the religion that are in most need of profound and detailed study by modern scholars.

Religious Taoism was quite distinct from Buddhism, traditionally supposed to have been its constant opponent and competitor for official patronage. Far more fundamental and historically important, however, has been Taoism's basic opposition to those beliefs and practices comprising the worship of what Taoists termed the "gods of the profane" (*su shen*); *i.e.*, the whole body of popular religious tradition and folklore. The Taoist master, with his hereditary lineage, his institutional affiliations, and his complex rituals with their liturgies in the Chinese literary language, considered himself to be far apart from both the exorcist of the folk religion, whose task was to cast out demons, and the medium, who was chosen for his visionary or hypnotic aptitude and who performed his operations in the colloquial language, expressing his spiritual and therapeutic efficacy through ecstatic and frenzied performances. The activities of the exorcists and mediums seemed to Taoist adepts to partake more of the demonic than of the pure and impersonal Tao (the "Way"); indeed, for the Taoists, the whole proliferating pantheon of Chinese popular gods were, at best, auxiliary spirits, or, at worst, dangerous demons, leading their deluded worshippers to insanity and perdition. At this point, Taoism converged with both Buddhism and the secular ("Confucian") state administration in their desires to rechannel or suppress unauthorized worship (*yin-ssu*, "abusive cults"). Unregulated religious expression among the populace was considered as dangerous to the commonwealth as it was to the individual worshipper who had become addicted to unrestrained spiritual enthusiasm.

Differences between Taoist masters and the priesthood of folk religion

Religious Taoism, then, may be considered as China's own indigenous higher religion; and its history thus requires close critical study in its dynamic interaction with the other elements constituting the spiritual life of China (*e.g.*, Buddhism, Confucianism, and folk religion). Despite the fundamental distinctiveness of Taoism's special character, it has both influenced and been influenced by other traditions with which it has so long been in contact.

Future studies of Chinese religion should therefore take adequate account of the Taoists' own view of their beliefs and practices, instead of considering Taoism exclusively from outside views, especially as Taoism appears in generally hostile Buddhist or Confucian documents. The basic textual source materials for such study are available in the great collection of Taoist texts, the Taoist canon (*Tao tsang*). Scholarly work on the canon has dealt primarily with texts of Taoism's formative period (*c.* 2nd to 6th centuries AD), and historical accounts must as a result devote a disproportionate amount of space to the first five centuries of the movement. Documentary sources, however, can give but a partial picture of the religion. Traditional Taoist practices may still be observed in Taiwan and the Chinese communities of Southeast Asia; and in these areas more recording and study must be carried out before this tradition accommodates itself to the pressures of modernization (see also TAOISM; CHINESE RELIGION; TAOIST LITERATURE; and ALCHEMY).



A Confucian (left), a Buddhist (centre), and a Taoist, a painting of a tranquil discussion that exemplifies the peaceful coexistence of the "three ways" in China. The figures show ideal types, each representative of his faith; the Confucian is dressed as a scholar, the Buddhist as a monk, and the Taoist, in his garment of leaves, as a forest ascetic.
By courtesy of the Chinese Classic Art Publishing House, Peking

THE ORIGIN AND EARLY DEVELOPMENT OF TAOISM

Philosophers and mystics. *The interpretation of the role of Lao-tzu.* Behind all forms of Taoism stands the figure of Lao-tzu, who has been viewed as a supposedly historical sage, an often reincarnated personage of legendary history, or a god—*i.e.*, the personification of the primal creative life breath of the universe. Lao-tzu as an historical figure is the reputed author of the book that bears his name, the *Lao-tzu*, or *Tao-te Ching* ("Classic of the Way of Power"). At once a manual of the art of government by inaction (*wu-wei*) and a unique spiritual testament, this brief work has long exercised a dominant influence on all Chinese thought.

The first mention of Lao-tzu is found in another early classic of Taoist speculation, the *Chuang-tzu* (4th–3rd centuries BC). In this work Lao-tzu is described as being one of Chuang-tzu's own teachers, and the same book contains many of the Master's (Lao-tzu's) discourses, generally introduced by the questions of a disciple. The *Chuang-tzu* also presents seven versions of a meeting of Lao-tzu and Confucius. Lao-tzu is portrayed as the elder and his Taoist teachings confound his celebrated interlocutor. The *Chuang-tzu* also gives the only account of Lao-tzu's death. Thus in this early source, Lao-tzu appears as a senior contemporary of Confucius (6th–5th centuries BC) and a renowned Taoist master, a curator of the archives at the court of the Chou dynasty (*c.* 1122–221 BC) and, finally, a mere mortal.

The teachings of Lao-tzu and Chuang-tzu

The first consistent biographical account of Lao-tzu is found in the "Historical Records" (*Shih-chi*)—China's first universal history (2nd century BC)—of Ssu-ma Ch'ien. This concise résumé has served as the classical source on the philosopher's life. Lao-tzu's family name was Li, his given name Erh; and he occupied the post of archivist at the Chou court. He is said to have instructed Confucius on points of ceremony. Observing the decline of the Chou dynasty, Lao-tzu left the court and headed west. At the request of Yin Hsi, the guardian of the frontier pass, he wrote his treatise on the Tao, some 5,000 characters, in two scrolls. He then left China behind, and what became of him is not known. The historian quotes variant accounts, including one that attributed to Lao-tzu an exceptional longevity; the narrative terminates with the genealogy of eight generations of Lao-tzu's supposed descendants. With passing references in other early texts, this constitutes the body of information on the life of the sage as of the 2nd century BC; it is presumably legendary. Modern scholarship has little to add, and the book, the *Tao-te Ching*, doubtfully credited to Lao-tzu, stands

alone, with all its attractions and enigmas, as the fundamental text of both philosophical and religious Taoism (see also LAO-TZU).

The interpretation and role of Chuang-tzu. Pseudo-historical knowledge of the sage Chuang-tzu is even less well defined. Most of Ssu-ma Ch'ien's brief portrait of the man is transparently drawn from anecdotes in the *Chuang-tzu* itself and as such has no necessary basis in fact. The book that bears his name, the *Chuang-tzu,* however, is valuable as a monument of Chinese literature and because it contains considerable documentary material, describing numerous speculative trends and spiritual practices of the Warring States period (481–221 BC). Breath control, meditation, gymnastics, and dietetics are all mentioned, as are spiritual beings ("perfected ones," *chen-jen*), who have transcended the human condition. The work describes spirit journeys that seem to adapt and refine the concrete practices of ecstatic religion, and it gives accounts of masters and disciples that prefigure the spiritual relationships current within the later Taoist religion. The earliest Taoist thinkers were the most articulate members of a much larger class of philosophers and mystics. The "Taoist school" (Tao-chia) only dates from the early Imperial period, at which time it first becomes possible to discuss in factual detail the history of a Taoist movement (see also CHUANG-TZU).

Taoism in the Ch'in and Han periods (221 BC–AD 220) of the Chinese empire. *Esoteric traditions of eastern China.* The textual remains of Taoism during the Warring States period were all presumably produced in connection with official patronage; similarly, developments in Taoist thought and practice during the early Imperial age principally have to be studied from the vantage point of the court. At the Imperial court, representatives of different local traditions met as competitors for official favour, and the court consequently served as the principal meeting place for the exchange of ideas. The historians who recorded the progress of these varying intellectual and religious currents were themselves court officials and often were active participants in the movements they describe. The emperors, anxious to consolidate and expand their power, were a natural focus for wonder-workers and specialists in esoteric arts.

Taoism's influence at the Imperial court

A series of such wonder-workers from the eastern seaboard visited the courts of the Ch'in and early Han. They told of islands in the ocean, peopled by immortal beings —which the *Chuang-tzu* had described—and so convincing were their accounts that sizable expeditions were fitted out and sent in search of them. The easterners brought the cults of their own region to the capital, recommending and supervising the worship of astral divinities who would assure the emperor's health and longevity. One of their number, Li Shao-chün, bestowed on the Han emperor Wu Ti counsels that are a résumé of the spiritual preoccupations of the time. The emperor was to perform sacrifices to the furnace (*tsao*), which would enable him to summon spiritual beings. They in turn would permit him to change cinnabar powder (mercuric sulfide) into gold, from which vessels were to be made, out of which he would eat and drink. This would increase his span of life and permit him to behold the immortals *(hsien)* who dwell on the Isles of P'eng-lai, in the midst of the sea. Here, for the first time, alchemy joins the complex of activities that were supposed to contribute to the prolongation of life.

The Huang-Lao tradition. Also originating in the eastern coastal region (Shantung), alongside these same thaumaturgic (wonder-working) tendencies, was the learned tradition of the Huang–Lao masters, devotees of the legendary "Yellow Emperor" (Huang Ti) and Lao-tzu. The information on the life of Lao-tzu transmitted by Ssu-ma Ch'ien probably derives directly from their teaching. They venerated Lao-tzu as a sage whose instructions, contained in his cryptic book, describe the perfect art of government. The Yellow Emperor, with whose reign Ssu-ma Ch'ien's universal history opens, was depicted as a ruler of the Golden Age who achieved his success because he applied his teachers' precepts to government. The Yellow Emperor also was celebrated as the patron of technology; and the classic works of many arcane arts, including alchemy, medicine, sexual techniques, cooking, and dietetics, were all placed under his aegis. Unlike Lao-tzu, the Yellow Emperor is always the disciple, an unremitting seeker of knowledge, and the Huang–Lao masters' ideal of the perfect ruler who values good teachers.

From the court of the King of Ch'i (in present-day Shantung Province) where they were already expounding the *Lao-tzu* in the 3rd century BC, the teachings of the Huang–Lao masters soon spread throughout learned and official circles in the capital. Many early Han statesmen became their disciples and attempted to practice government by inaction *(wu-wei);* among them were also scholars who cultivated esoteric arts. Although their doctrine lost its direct political relevance during the reign of the emperor Hsiao Wu Ti (reigned 141–87 BC), their ensemble of teachings concerning both ideal government and practices for prolonging life continued to evoke considerable interest and is perhaps the earliest truly Taoist movement of which there is clear historical evidence.

Revolutionary messianism. Among the less welcome visitors at the Han court had been a certain Kan Chung-k'o. At the end of the 1st century BC he presented to the emperor a "Classic of the Great Peace" (*T'ai-p'ing Ching*) that he claimed had been revealed to him by a spirit, who had come to him with the order to renew the Han dynasty. His temerity cost him his life, but the prophetic note of dynastic renewal became stronger during the interregnum of Wang Mang (AD 8–23); and other works—bearing the same title—continued to appear. At this time, promotors of a primitivistic and utopian T'ai-p'ing (Great Peace) ideology continued to support the Imperial Liu (Han) family, claiming that they would be restored to power through the aid of the Li clan. A century and a half later, however, as the power of the Later Han dynasty (AD 23–220) declined, the populace no longer hoped for a renewal of Han rule.

T'ai-p'ing messianism

The great Yellow Turban Rebellion broke out in the east in AD 184. Its leader, Chang Chüeh, declared that the "blue heaven" was to be replaced by a "yellow heaven"; and his followers wore yellow turbans in token of this expectation. Worshipping a "Huang-lao Chün," the movement gained a vast number of adherents throughout eastern China. Though they were eventually defeated by the Imperial forces, the tendency towards messianic revolt continued to manifest itself at frequent intervals. A great many charismatic leaders came from the Li family, and certain of them claimed to be the god Lao-tzu returned to earth; a sage of western China, Li Hung, who had actually lived during the 1st century BC, became the favourite recurrent figure of later would-be messiahs. Such revolutionary religious movements, which included Taoist ideological elements remained a persistent feature of medieval Chinese history. The last recorded Li Hung was executed in 1112. These sporadic popular manifestations of revolutionary messianism, though, did not represent the activities of the formal Taoist organization and must be distinguished from the organized religious Taoism that also appeared at the end of the Later Han period.

DEVELOPMENT OF THE TAOIST RELIGION FROM THE 2ND TO THE 6TH CENTURIES

The Way of the Celestial Masters. *The development of organized Taoism.* The Taoist organization that was to endure first developed in the mountains of the province of Szechwan. There, a certain Chang Tao-ling, in AD 142, is said to have received a revelation from T'ai-shang Lao-chün ("Lord Lao the Most High"). The deified Lao-tzu bestowed on Chang his "orthodox and sole doctrine of the authority of the covenant" (*cheng-i meng-wei fa*), meant as a definitive replacement for the religious practices of the people, which are described as having lapsed into demonism and degeneracy.

The new dispensation at first was probably intended as a substitute for the effete rule of the Han central administration. Chang is said in time to have ascended on high and to have received the title of *t'ien-shih* ("celestial master") and by the latter part of the 2nd century, under the leadership of his descendants, the T'ien-shih Tao

(Way of the Celestial Masters) constituted an independent religio-political organization with authority throughout the region, a "Taocracy" (rule of Tao), in which temporal and spiritual powers converged. For ceremonial and administrative purposes, the realm was divided into 24 (later 28 and 36) units, or parishes (*chih*). The focal point of each was the oratory, or "chamber of purity" (*ching-shih*), which served as the centre for communication with the powers on high. Here the *chi-chiu* ("libationer"), the priestly functionary of the nuclear community, officiated. Each household contributed a tax of five pecks of rice to the administration, whence came the other common name of the movement, the Way of the Five Pecks of Rice (Wu-tou-mi Tao).

Ritual activities of the libationer

The ritual activities of the libationer seem principally to have been directed towards the cure of disease by prescribed ceremonial means. Believed to be a punishment for evil deeds, whether committed by the sufferer himself or by an ancestor, illness was in fact a sentence pronounced by the San Kuan (Three Officials), judges and custodians of the dead. The sentence was carried out by the spectral hordes of the Six Heavens (Liu T'ien), a posthumous dwelling place of all unhallowed mortals. Against such judicial severity, only formal appeal to higher authority might avail. Using the rising flame and smoke of the incense burner in the centre of the oratory to transmit the message borne by spirits exteriorized from within his own body, the libationer submitted petitions (*chang*) to the appropriate bureau of the three Taoist heavens (*san t'ien*). The Taoist canon contains long lists of the "officials and generals" (*kuan chiang*), each specializing in a different sort of complaint, who would respectively pronounce on the appeal and marshal the celestial forces against the offending demons.

The officiant came to dispose of a large selection of bureaucratic stock drafts: memorials, plaints, and appeals, all of which were modelled on secular administrative usage. Also effective were written talismans (*fu*); drawn by the libationer, these would be burned and the ashes, mixed with water, swallowed by the demons' victim. The libationer also functioned as a moral preceptor, instructing the faithful in the sect's own highly allegorical interpretation of the *Lao-tzu*, which they considered to be the revealed work of Lord Lao the Most High. Their fundamental concern with right actions and good works as being most in the spirit of the Tao and consequently ensuring immunity from disease is also shown by their construction of way stations in which provisions and shelter were placed for the convenience and use of travellers, as well as in the numerous injunctions to charity and forebearance recorded in the written codes of the movement.

Communal ceremonies. Both the nuclear communities and the "Taocratic" realm as a whole were bound together by a ritual cycle, of which only fragmentary indications remain. Among the most important ceremonial occasions were the communal feasts (*ch'u*) offered at certain specific times throughout the year (during the first, seventh, and 10th months) as well as on other important occasions, such as initiation into the hierarchy, advancement in rank or function, or the consecration of an oratory. These feasts were of varying degrees of elaborateness, depending on the circumstances. The common essential element, however, was the sharing of certain foods, in prescribed quantities, among masters and disciples. This was envisaged as a communion with the Tao, at once attesting the close compact with the celestial powers enjoyed by the members of the parish and reinforcing their own sense of cohesion as a group.

The Union of Breaths rite

Much more notorious was the rite known as the Union of Breaths (Ho Ch'i), a communal sexual ritual said to have been celebrated at each new moon. Later Buddhist sources described this as a riotous orgy of outrageous and disgusting license. Several cryptic manuals of instruction for the priest in charge of these proceedings are preserved in the canon; and they depict, however, scenarios of a highly stylized erotic choreography of cosmic significance. Like the communal feasts, these rites might be interpreted as a concentrated and idealized adaptation of older, more diffuse agrarian religious customs. This suggests a pattern of the integration of local practices that has remained characteristic of Taoism throughout its history.

Official recognition of the Taoist organization. In AD 215, the celestial master Chang Lu, grandson of Chang Tao-ling, submitted to the authority of the Han general Ts'ao Ts'ao, who six years later founded the Wei dynasty in the north. This resulted in official recognition of the sect by the dynasty; the celestial masters in turn expressed their spiritual approbation of the Wei's mandate to replace the Han. Under these conditions a formal definition of the relations of organized Taoism to the secular powers developed. In contrast to the popular messianic movements, Lao-tzu's manifestation to Chang Tao-ling was considered to be definitive; the god was not incarnate in them but rather designated Chang and his successors as his representatives on earth. Under a worthy dynasty, which governed by virtue of the Tao, the role of the celestial masters was that of acting as intermediaries for celestial confirmation and support. Only when a responsible ruler was lacking were the celestial masters to take over the temporal guidance of the people and hold the supreme power in trust for a new incumbent. Abetted by this flexible ideology of compromise, the sect made constant progress at the courts of the Wei and Western Chin dynasties until, by the end of the 3rd century, it counted among its adherents many of the most powerful families in North China.

Buddhist and Taoist interaction. The first mention of Buddhism in China (AD 65) occurs in a Taoist context, at the court of a member of the Imperial family known for his devotion to the doctrines of Huang-Lao. The Indian religion was at first regarded as a foreign variety of Taoism; the particular Buddhist texts chosen to be translated during the Han period reveal the Taoist preoccupation of the earliest converts with rules of conduct and techniques of meditation. Early translators employed Taoist expressions as equivalents for Buddhist technical terms. Thus, the Buddha, in achieving enlightenment (*bodhi*) was described as having "obtained the Tao"; the Buddhist saints (*arhat*) become perfected immortals (*chen-jen*); and "non-action" (*wu-wei*) was used to render Nirvāṇa (the Buddhist state of bliss). A joint sacrifice to Lao-tzu and the Buddha was performed by the Han emperor in AD 166. During this period occurred the first reference to the notion that Lao-tzu, after vanishing into the west, became the Buddha. This theory enjoyed a long and varied history. It claimed that Buddhism was a debased form of Taoism, designed by Lao-tzu as a curb on the violent natures and vicious habits of the "western barbarians," and as such was entirely unsuitable for Chinese consumption. A variant theory even suggested that, by imposing celibacy on Buddhist monks, Lao-tzu intended the foreigners' extinction. In approximately AD 300, the Taoist scholar Wang Fou composed a "Classic of the Conversion of the Barbarians" (*Hua hu Ching*), which was altered and expanded in subsequent centuries to encompass new developments in the continuing debate. Although there is no evidence that the earliest Taoist organization, literature, or ceremonies were in any way indebted to Buddhism, by the 4th century there was a distinct Buddhist influence upon the literary form of Taoist scriptures and the philosophical expression of the most eminent Taoist masters.

Taoist interpretations of Buddhism

The process of interaction, however, was a mutual one, Taoism participating in the widening of thought because of the influence of a foreign religion and Buddhism undergoing a partial "Taoicization" as part of its adaptation to Chinese conditions. The Buddhist contribution is particularly noticeable in the developing conceptions of the afterlife; Buddhist ideas of purgatory had a most striking effect not only upon Taoism but especially on Chinese popular religion. On a more profound level the ultimate synthesis of Taoism and Buddhism was realized in the Ch'an (Japanese Zen) tradition (from the 7th century on), into which the paradoxes of the ancient Taoist mystics were integrated. Likewise, the goal of illumination in a single lifetime, rather than at the end of an indefinite

succession of future existences, was analogous to the religious Taoist's objective of immortality as the culmination of his present life.

The Southern tradition. The political partition of China into three parts following the collapse of the Han dynasty in AD 220, the so-called period of the Three Kingdoms, had its spiritual counterpart in certain well-defined regional religious differences. Against the independent dynasties in the north and west stood the empire of Wu, south of the Yangtze River.

Developments in alchemical and other traditions. A region exposed comparatively lately to Chinese influence, this southeastern area had long been famous for its aboriginal sorcerers and dancing mediums. In the course of Chinese colonization, separate learned spiritual traditions developed alongside the ecstatic practices of the populace. To the court of the emperors of Wu came savants and wonder-workers representing a variety of traditions that were to acquire lasting influence.

Among these personages was a certain Ko Hsüan (3rd century AD), who was said to have been initiated into an ancient alchemical tradition. His great-nephew Ko Hung in the next century became one of the most celebrated writers on the various technical means for attaining immortality. In his major work, the *Pao-p'u-tzu* ("He Who Holds to Simplicity"), Ko Hung expounded the alchemical formulas received and transmitted by Ko Hsüan. In so doing, he took care to distinguish the divinely inspired "gold elixir" (*chin-tan*), or "liquefied gold" (*chin-i*)—*i.e.*, preparations of true edible, or potable, gold, the consumption of which leads to immortality (aurifaction)—from the mere counterfeiting of the precious substance, with intention to deceive (aurifiction). These alchemical methods have been designated as belonging to the T'ai-ch'ing (Great Purity) tradition, from the name of the heaven of the Immortals to which the elixirs were said to elevate their consumer. The chapters of alchemy in the *Pao-p'u-tzu* are among the earliest documents to describe the art in detail.

The goals and methods of alchemy

Ko Hung enumerated an extensive selection of material substances and practical operations to which he attributed varying degrees of relative efficacy in the prolongation of life. Dietetics (grain and alcohol avoidance), ingestion of solar, lunar, and astral exhalations and their cycling within the body, gymnastics, and conservation of vital fluids through proper sexual techniques were all necessary and fundamental. The usefulness of written talismans and the performance of good works were also not denied. Above all, it was essential that all disease be eliminated from the body before undertaking more positive, specialized practices for achieving immortality. Herbs and plants were useful not only against disease, but in many cases (particularly in that of mushrooms) their use resulted in definite lengthening of life. For a definitive transformation into an immortal (*hsien*), with all the powers and prerogatives that implied, however, an alchemical elixir must be compounded and consumed. Ko Hung admitted, however, that he himself had never succeeded in making one. After a strenuous life in civil and military service, in the course of which he managed to write voluminously on many subjects, this great eclectic scholar is said to have undertaken a long journey to China's colonial dominions in Vietnam in quest of the pure cinnabar found there. He stopped at Lo-fou Shan, near Canton, however, where he died.

The *Pao-p'u-tzu* was nearly finished in 317, when Lo-yang, capital of the Western Chin dynasty, fell to the Hsiung-nu. This event set off a considerable emigration to the unsubdued region south of the Yangtze River. The Imperial household was followed in its flight by numerous high-ranking dependents and their spiritual ministers. During this period the Way of the Celestial Masters, established at the court of Lo-yang since the early 3rd century, apparently first penetrated in force to the Southeast. While the secular, military menace remained in the North, and factional struggles raged among the emigrants, the Way of the Celestial Masters waged unremitting war against the indigenous sects and cults of demons of the Southeast. Many of the old established families, settled in the region since the end of the Han dynasty, turned away from local traditions to become members of the Taoist faith of their new political superiors. At first these converts were content to entrust the direction of their spiritual lives to the libationers of the movement, though these religious specialists were generally men of lower social standing than themselves. Among the second and third generation of converts from the old aristocracy of Wu, however, new and original impulses, which were to have most profound effects upon the development of Taoism as a whole, began to occur.

The Mao Shan Revelations. The most brilliant synthesis of the Way of the Celestial Masters with the indigenous traditions of the Southeast occurred in the 4th century AD in a family closely related to Ko Hung. Hsü Mi, an official at the Imperial court, and his youngest son, Hsü Hui, were the principal beneficiaries of an extensive new Taoist revelation. A visionary in the Hsüs' service, Yang Hsi, was honoured with the visits of a group of perfected immortals (*chen-jen*) from the heaven of Shang-ch'ing (Supreme Purity), an improvement on the T'ai-ch'ing heaven and the ordinary immortals (*hsien*) that peopled it. In the course of his visions, which lasted from AD 364 to 370, Yang received a whole new scriptural and hagiographic literature, in addition to much practical information from the "perfected" (*chen*) on how it was to be understood and employed. Like the Ko family, the Hsüs belonged to the old aristocracy of Wu, who had been displaced from prominence by the arrival of the great families of the North, to whose Taoist beliefs they had been converted. The perfected assured them that the present unjust order was soon to end and that the rule of men on Earth was to be replaced by a universal Taoist imperium. The present (*i.e.*, the 4th century) was a time of trials, given over to the reign of the demonic Six Heavens, and marked by war, disease, and the worship of false gods. The sole mission of the demonic forces, however, was to cleanse the earth of evildoers, a task that would be completed by an overwhelming cataclysm of fire and flood. At that time the Good would take refuge deep in the earth, in the luminous caverns of the perfected beneath such sacred mountains as Mao Shan (in Kiangsu Province), which was the immediate focus of spiritual interest for the Hsüs. There they would complete the study of immortality already begun in their lifetimes, so as to be ready for the descent from heaven of the new universal ruler, Lord Li Hung, the "sage who is to come" (*hou-sheng*). This was prophesied for the year 392. Yang and the Hsüs would get high office in the heaven of Shang-ch'ing and rule over a newly constituted earth below, peopled by the elect (*chung-min*).

Synthesis of the Way of Celestial Masters with the Southeast traditions

Yang Hsi's prodigious genius gave great consistency and consummate literary form to his comprehensive synthesis of many spiritual traditions. Popular messianism was adapted to provide an encompassing framework and temporal cogency. Yang and his patrons, however, were also thoroughly familiar with Buddhist thought. In addition to integrating Buddhist concepts into their Taoist system, the perfected also dictated a "Taoicized" version of large portions of an early Buddhist compilation, the *Sūtra in Forty-two Sections* (*Ssu-shih-erh chang Ching*). Buddhist notions of predestination and reincarnation were subtly blended with native Chinese beliefs in hereditary character traits and the clan as a single unit involving mutual responsibility on the part of all its members, living and dead. Furthermore, building upon the Way of the Celestial Masters, the Mao Shan revelations envisaged some reform of the practices of the parent sect. Its sexual rites in particular were stigmatized as inferior practices, more conducive to perdition than to salvation. In place of this, a spiritualized union with a celestial partner was apparently realized by Yang Hsi and promised to his patrons. Other rituals of the Celestial Masters were allowed to continue in use among the Mao Shan adepts but were relegated to a subordinate position. Thus, the movement did not reject but rather incorporated and transcended the older tradition.

Though the perfected inveighed against the popular cults, even elements of these were absorbed and trans-

formed. There is some evidence that, before Yang's inspired writings, Lord Mao himself, the august perfected immortal who gave his name to the mountain, was no more than a local minor god worshipped by an exorcistic priestess in the shadow of Mao Shan. Among the more learned traditions, alchemy received particular attention, being adopted for the first time into the context of organized religious Taoism. The perfected revealed the highly elaborate formulas of several of the elixirs that served them as food and drink. For all their extravagance, they were intended as real chemical preparations and described as being deadly poisonous to mortals. By preparing and ingesting one of them, the younger Hsü probably willingly ended his earthly existence in order to take up the post that had been offered him in the unseen world and to make ready for the coming of the new era.

Absorption of elements from local popular cults

The Ling Pao scriptures and liturgies. Another member of the Ko family was responsible for the second great Taoist scriptural tradition. Ko Ch'ao-fu began composing the *Ling pao Ching* ("Classic of the Sacred Jewel") *c.* AD 397. He claimed that they had been first revealed to his own ancestor, the famous Ko Hsüan, early in the 3rd century. In these works the Tao is personified in a series of "celestial worthies" (*t'ien-tsun*), its primordial and uncreated manifestations. These in turn were worshipped by means of a group of liturgies, which, during the 5th century, became supreme in Taoist practice, completely absorbing the older, simpler rites of the Way of the Celestial Masters. As each celestial worthy represented a different aspect of the Tao, so each ceremony of worship had a particular purpose, which it attempted to realize by distinct means. The rites as a whole were called *chai* ("retreat"), from the preliminary abstinence obligatory on all participants. They lasted a day and a night or for a fixed period of three, five, or seven days; the number of persons taking part was also specified, centring on a sacerdotal unit of six officiants. One's own salvation was inseparable from that of his ancestors; the Huang-lu chai (Retreat of the Yellow Register) was directed towards the salvation of the dead. Chin-lu chai (Retreat of the Golden Register), on the other hand, was intended to promote auspicious influences on the living. The T'u-t'an chai (Mud and Soot Retreat, or Retreat of Misery) was a ceremony of collective contrition, with the purpose of fending off disease, the punishment of sin, by prior confession; in Chinese civil law, confession resulted in an automatic reduction or suspension of sentence. These and other rituals were accomplished for the most part in the open, within a specially delimited sacred area, or altar (*t'an*), the outdoor complement of the oratory. The chanted liturgy, innumerable lamps, and clouds of billowing incense combined to produce in the participants a cathartic experience that assured these ceremonies a central place in all subsequent Taoist practices.

The great Southern masters. Though Taoism never became the exclusive state religion in the South, its most eminent representatives founded powerful organizations that received considerable official support. Lu Hsiu-ching in the 5th century epitomized the Ling Pao tradition, the liturgies of which he codified. His establishment at the great Buddho-Taoist centre, Lu Shan (in Kiangsi Province), carried out ceremonies and provided auspicious portents in favour of the Liu-Sung dynasty (420–479), in whose rulers Taoists complacently agreed to recognize the fulfillment of the old messianic prophesies and the legitimate continuation of the Han dynasty. Lu was frequently invited to the capital (present-day Nanking), where the Ch'ung-hsü Kuan (Abbey) was founded for him and served as the focal point of the Ling Pao movement.

The influence of Lu Hsiu-ching and T'ao Hung-ching

Like Lu, who was a member of the old aristocracy of Wu, T'ao Hung-ching of the 5th and 6th century enjoyed even greater renown as the most eminent Taoist master of his time. He spent years in searching out the manuscript legacy of Yang Hsi and the Hsüs, and in 492 retired to Mao Shan, where he edited and annotated the revealed texts and attempted to re-create their practices in their original setting. T'ao's fame as a poet, calligrapher, and natural philosopher has persisted throughout Chinese history; he is perhaps best known as the founder of critical pharmacology. T'ao was an intimate friend of the great Liang emperor Wu Ti (of the 6th century), and his Mao Shan establishment was able to survive the proscription of all other Taoist sects in 504. Though whole Taoist families lived under T'ao's spiritual rule at Mao Shan, he himself stressed the need for celibacy and full-time commitment to the work of the Tao. In his state-sponsored Chu-yang Kuan, T'ao appears to have effected a working synthesis of the public rites of the Ling Pao liturgies with the private and individual practices enjoined in the Mao Shan revelations. This dual practice was to remain a feature of all subsequent Taoist sects. T'ao's primary interest, however, was in the scriptures of the perfected of Shang-ch'ing; and this is reflected in the revelations vouchsafed by these same spiritual agents to a 19-year-old disciple of T'ao's, Chou Tzu-liang, in 515–516. These revelations show a pronounced Buddhist influence, and T'ao was himself reputed to be a master of Buddhist as well as Taoist doctrine. His writings evidence a complete familiarity with Buddhist literature, and it is reported that both Buddhist monks and Taoist priests officiated at his burial rites.

State Taoism in the North. Under the foreign rulers of North China, independent developments likewise were in progress. In 415, one K'ou Ch'ien-chih received a revelation from Lao-chün himself. According to this new dispensation, K'ou was designated celestial master and ordered to undertake a total reformation of Taoism. Not only were all popular messianic movements claiming to represent Lao-chün unsparingly condemned but K'ou's mission was particularly aimed at the elimination of abuses from the Way of the Celestial Masters itself. Sexual rites and the taxes contributed to the support of the priesthood were the principal targets of the god's denunciations; "What have such matters to do with the pure Tao?" he irately demanded. The proposed reform was far more radical than that forseen in the Mao Shan revelations of the Southeast, and K'ou was given concrete temporal power of a sort that the Hsüs had not envisaged. Political and economic factors favoured the acceptance of his message at court; Emperor T'ai Wu Ti (5th century) of the Northern Wei dynasty put K'ou in charge of religious affairs within his dominions and proclaimed Taoism the official religion of the empire. The Emperor considered himself to reign as the terrestrial deputy of the deified Lao-tzu, as is indicated by the name of one of the periods of his reign: T'ai-p'ing Chen-chün (Perfect Lord of the Great Peace). The dominant position of Taoism under the Northern Wei, however, apparently did not long survive K'ou Ch'ien-chih's death in 448.

New revelations in the North

TAOISM UNDER THE T'ANG, SUNG, AND LATER DYNASTIES

Taoism under the T'ang dynasty (618–907). China's reunification under the T'ang marked the beginning of Taoism's most spectacular success. The dynasty's founder, Li Yüan, claimed to be descended from the Lao-tzu; as his power increased, even the influential Mao Shan Taoists came to accept him as the long-deferred fulfillment of messianic prophecy. This notion was built into the dynasty's state ideology, and the emperor was commonly referred to as the "sage" (*sheng*). Prospective candidates for the civil service were examined in either the Ling Pao "Classic of Salvation" (*Tu-jen Ching*) or the Mao Shan "Classic of the Yellow Court" (*Huang-t'ing Ching*). Under a series of celebrated patriarchs, the Mao Shan organization dominated the religious life of the age. One of the greatest of the line, Ssu-ma Ch'eng-chen, initiated innumerable government officials and eminent men of letters and served as spiritual master to emperors. The personnel of the Mao Shan revelations even entered into the formal framework of state religion. When Ssu-ma Ch'eng-chen pointed out that the sacred peaks of the Imperial cult were in reality under the superintendence of the perfected of Shang-ch'ing, officially sponsored shrines were erected to them there; and their propitiation was incorporated into the traditional rites.

The period of Taoism's greatest success

The wide diffusion of Taoism throughout the vast T'ang

empire is reflected by the sizable proportion of Taoist texts discovered in the walled-up caves at Tun-huang (in Kansu Province). This town in the far west of China was the gateway to Central Asia; and here Taoists came into contact not only with Buddhists of many different doctrinal persuasions but also with Nestorian Christians (members of a church independent of the patriarch of Constantinople) and Manichaeans (Iranian religious dualists). Copies of the *Lao-tzu* were sent to the King of Tibet, and the book was translated into Sanskrit at the request of the ruler of Kashmir. It also reached Japan in the 7th century, as did texts of religious Taoism; reports of Taoism's dominance on the continent may still be read in the diaries of Japanese Buddhist pilgrims. The geographic extension of the religion at this time was also represented, in the legendary sphere, by the systematic elaboration of its sacred mountains and the traditions attaching to each of them. They are described by the great hagiographer, Tu Kuang-t'ing, at the end of the T'ang dynasty. In addition to the great "cavern-heavens" (*tung-t'ien*), 10 holy mountains known to the original Mao Shan revelations, he lists 36 lesser cavern heavens and 72 sanctuaries (*fu-ti*). Situated throughout the length and breadth of the empire, they are fitting spiritual guideposts across the dominions of the T'ang, which saw itself as an essentially Taocratic realm.

Taoism under the Sung and Yüan dynasties. *Internal developments.* The Sung (960–1279) and Yüan (1279–1368) periods witnessed a great religious effervescence, stimulated in part, under the Sung, by the menace of foreign invasion and, during the Yüan, by Tantric (esoteric, or occultic) Buddhism that was in vogue among the new Mongol rulers of China. During the preceding centuries the Way of the Celestial Masters, centred at Lung-hu Shan (Dragon-Tiger Mountain, Kiangsi), had been eclipsed by the prestige of Mao Shan. At the end of the Northern Sung period, the 30th celestial master, Chang Chi-hsien, was four times summoned to court by the Sung emperor Hui Tsung, who hoped for spiritual support for his threatened reign. Chang Chi-hsien was credited with a renovation of the ancient sect, thereafter called the "New Orthodox Unity" (Hsin Cheng-i), and with the introduction of the influential rites of the "five thunders" (*wu-lei*) into Taoist liturgy.

The founding of new Taoist sects

After the retreat of the Sung government south of the Yangtze River (1126), a number of new Taoist sects were founded in the occupied North and soon attained impressive dimensions. Among them were: the T'ai-i (Supreme Unity) sect, founded *c.* 1140 by Hsiao Pao-chen; the Chen-ta Tao (Perfect and Great Tao) sect of Liu Te-jen (1142); and the Ch'üan-chen (Perfect Realization) sect, founded in 1163 by Wang Ch'ung-yang (Wang Che). This last sect came to the favourable attention of the Mongols, who had taken over in the North, and its second patriarch, Ch'iu Ch'ang-ch'un, was invited into Central Asia to preach to Genghis Khan. The sect enjoyed great popularity, and its establishments of celibate monks continued to be active into the 20th century, with the famous White Cloud Monastery (Po-yün Kuan) at Peking as headquarters. In the South, Mao Shan continued to prosper, while the Ko-tsao sect flourished at the mountain of that name, in Kiangsi Province. This was said to be the spot where the 3rd century immortal, Ko Hsüan, had ascended to heaven; the sect looked to him as its founder, and it transmitted the Ling Pao scriptures, which he was believed to have been the first to receive.

Alchemical developments. While learned specialists continued to refine alchemical theory, the period witnessed increasing interest in internal alchemy (*nei tan*), in which the language of the laboratory was used to describe operations realized within the body. This, in a sense, was nothing new. Alchemical metaphors had very early been applied to physiology; Ko Hung, for example, called semen the "Yin elixir." By Sung times, however, the systematic interiorization and sublimation of alchemy had become so widespread that all earlier texts of operative, external alchemy (*wai tan*) were henceforth supposed to have really been written about *nei tan*, and the attempt to compound a tangible chemical elixir was thought to have been no more than a hoax. Liturgy also provided its own sublimation of the older art: the *lien-tu* ("salvation by smelting") funeral service was developed at this time, in which an "elixir of immortality" was compounded of written talismans and offered to the deceased.

Syncretism. With such prestigious examples as Ch'an Buddhism (emphasizing intuitive meditation) and Neo-Confucianism (emphasizing knowledge and reason) before them, Taoists did not long delay in constructing interesting syntheses of their own and other beliefs. Confucianism now joined Buddhism as a fertile source of inspiration. The revelations of Hsü Sun, supposed to have lived in the 4th century AD, to one Ho Chen-kung in 1131 inspired the "Pure and Luminous Way of Loyalty and Filial Obedience" (Ching-ming Chung-hsiao Tao). This sect preached the Confucian cardinal virtues as being essential for salvation, and consequently won a considerable following in conservative intellectual and official circles. Another highly popular syncretistic movement of Taoist origin was that of the Three Religions (San Chiao). Its composite moral teachings are represented by popular tracts, the so-called "books on goodness" (*shan-shu*), which have been in extremely wide circulation since the Ming dynasty (1368–1644).

TAOISM IN MODERN TIMES

Taoism on Taiwan

The principal refuge of Taoism in the 20th century is on Taiwan. Its establishment on the island is doubtless contemporary with the great emigration from the opposite mainland province of Fukien in the 17th and 18th centuries. The religion, however, has received new impetus since the 63rd celestial master, Chang En-pu, took refuge there in 1949. On Taiwan, Taoism may still be observed in its traditional setting, distinct from the manifestations of popular religion that surround it. Hereditary Taoist priests (Taiwanese *sai-kong*), called "blackheads" (*wu t'ou*) from their headgear, are clearly set off from the exorcists (*fa-shih*) or "redheads" (*hung-t'ou*) of the ecstatic cults. Their lengthy rites are still held, now known under the term *chiao* ("offering"), rather than the medieval *chai* ("retreat"). The liturgy chanted, in expanded Sung form, still embodies elements that can be traced back to Chang Tao-ling's sect. The religion has enjoyed a renaissance since the 1960s, with great activity being carried on in temple building and restoration. The most significant event in the past several centuries of Taoist history, however, probably is the ordination (in 1964) in Taiwan of a Dutch scholar, K.M. Schipper, as a Taoist priest. His systematic, first-hand researches into Taoist practices may very well revolutionize scholarly knowledge of the religion, which will thus acquire an unforeseen historical extension, in the West and into the future.

BIBLIOGRAPHY

Taoist speculation and mysticism: ARTHUR WALEY, *The Way and Its Power: A Study of the Tao tê ching and Its Place in Chinese Thought* (1934), a classic translation and study of the *Lao-Tzu*; *Three Ways of Thought in Ancient China* (1939), the *Chuang-tzu* studied in relation to the thought of its time; JOSEPH NEEDHAM, *Science and Civilisation in China*, vol. 2, *History of Scientific Ideas* (1956), a highly lucid exposé of early Taoist speculation and later Taoist technology; FUNG YU-LAN, *A History of Chinese Philosophy*, 2nd ed., 2 vol. (Eng. trans. 1952), a standard reference work on the classical period of Chinese thought.

The Taoist religion: HENRI MASPERO, *Le Taoïsme et les religions chinoises*, rev. ed. (1971), contains all his writings on Chinese religion, including fundamental studies of Taoism and popular mythology; ANNA K. SEIDEL, *La Divinisation de Lao tseu dans le Taoïsme des Han* (1969), the best detailed study of the formation of the Taoist system in the early Imperial period; EDOUARD CHAVANNES, *Le Jet des dragons* (1919), a classic study and translation of a Taoist liturgy dating from the end of the T'ang dynasty; ARTHUR WALEY, *The Travels of an Alchemist* (1931), the account of the Central Asian journey of the second Patriarch of the Ch'üan-chen sect; J.J.M. DE GROOT, *Les Fêtes annuellement célébrées à Émoui (Amoy)* (1886) and *The Religious System of China*, vol. 6 (1910, reprinted 1964), both valuable pioneering descriptions of the Taoist priesthood and popular exorcists in Fukien at end of the 19th century.

(M.St.)

Taoist Literature

Taoist literature includes the sacred texts and speculative literature of the Chinese religion and philosophy known as Taoism.

TAOIST LITERATURE AND CHINESE SOCIETY

The two classes of Taoist texts

The literature of Taoism may conveniently be divided into two classes, which may be called philosophical (or speculative) Taoism and esoteric (or religious) Taoism; they correspond largely to two distinct historical periods. The basic texts of the first group date from the 4th to 2nd centuries BC and are speculative in tone. The Taoism of this period—the Warring States period—was one among many contending schools of thought. Like other schools of the time, its treatises were probably composed with the aim of convincing a ruler. Since all ancient Chinese philosophies were essentially theories of government, these works were probably in the public domain from the outset. But documents of the period are few, and in most cases scholars know nothing of the authors of this literature or of the milieu in which they worked.

With the political unification of China at the end of the 3rd century BC, the old schools blended into the complex fabric of thought of the early Han Empire. It is Han thought, likewise, that served as a point of departure for the second group of texts, which may properly be called Taoist—the esoteric Taoism beginning in the 2nd century AD. These writings represent the literature of a succession of organized religious movements and were originally intended for the limited circle of their members. The circumstances of their transmission were strictly regulated, an oath of secrecy being the bond that united disciple and master. The texts of the first, or philosophical, group, like other speculative treatises, had come into general circulation among men of culture and learning, and works of that period have survived in a variety of secular editions. The texts of the esoteric tradition, however, were from the first the property of restricted social groups, which included comparatively extensive missionary movements as well as narrow circles of masters and disciples. When comprehensive collection of texts began in the 5th century, sectarian considerations still played an important role. The successors to these early collections were the immense Taoist canons (*Tao Tsang*); the present-day canon, printed in the 15th century, contains nearly 1,500 titles, among which are many documents from the first centuries of esoteric Taoism. It is to this single repository that scholars must turn in order to study the literature of the Taoist religion.

Of these two distinct phenomena, the earlier, philosophical Taoism has until now received far more scholarly attention. This is not surprising because the classics of speculative Taoism include some of the masterpieces of wit and imagination in all Chinese literature, and an enticing aura of mystery shrouds the circumstances of their composition. It is paradoxical that so much less is generally known of esoteric Taoist literature, though potential sources of information are far more numerous. It has by far the longer and broader creative history, for while speculative Taoism remained an upper class phenomenon, the esoteric tradition extended to all levels of society and found expression in a much wider variety of literary genres. Not only is it represented by many more datable texts, conveniently accessible in the canon, but also the outside supporting material is both more extensive and more varied and includes the dynastic histories, inscriptions, and the collected literary works of individual authors. It is thus possible to study many key works of religious Taoism in close relation to the society that produced and employed them. It is in this almost unexplored field that the most exciting and significant research remains to be done. There is increasing evidence of the pervasive influence of esoteric Taoism on every aspect of Chinese life and thought. As the literature of China's indigenous organized religion, Taoist texts represent the expression of all that is most intrinsically Chinese. The continuing study of the Taoist Canon in its wider historical context can radically alter many established conceptions of both Chinese religion and Chinese society (see also CHINESE RELIGION; CHINESE PHILOSOPHY; and CHINA, HISTORY OF).

TEXTS OF SPECULATIVE TAOISM

The Tao-te Ching. The *Tao-te Ching* (an Eng. trans., *Book of the Way and Its Power*), or *Lao-tzu*, occupies a position of importance in Chinese thought out of all proportion to its size. Very brief, its 81 sections contain only about 5,000 characters in all, from which it derives still another of its titles, *Lao Tzu's Five Thousand Words*. The text itself appears in equal measure to express a profound quietism and determined views on government. It is consequently between the extremes of meditative introspection and political application that its many and widely divergent interpreters have veered.

The book shares certain constants of classical Chinese thought but clothes them in an imagery of its own. The sacred aura surrounding kingship is here rationalized and expressed as "inaction" (*wu-wei*), demanding of the sovereign no more than right cosmological orientation at the centre of an obedient universe. Survivals of archaic notions concerning the compelling effect of renunciation —which the Confucians sanctified as ritual "deference" (*jang*)—are echoed in the recommendation to "hold to the role of the female," with an eye to the ultimate mastery that comes of passivity.

The meaning of the Tao

It is more particularly in the function attributed to the Tao, or Way, that this little tract stands apart. The term Tao was employed by all schools of thought. The universe has its Tao; there is a Tao of the sovereign, his royal mode of being, while the Tao of man comprises continuity through procreation. Each of the schools, too, had its own Tao, its way or doctrine. But in the *Tao-te Ching*, the ultimate unity of the universal Tao itself is being proposed as a *social* ideal. It is this idealistic peculiarity that seems to justify later historians and bibliographers in their assignment of the term Taoist to the *Tao-te Ching* and its successors.

From a literary point of view, the *Tao-te Ching* is distinguished for its highly compressed style. Unlike the dialectic or anecdotal composition of other contemporary treatises, it articulates its cryptic subject matter in short, concise statements. More than half of these are in rhyme, and close parallelism recurs throughout the text. No proper name occurs anywhere. This singular work is traditionally ascribed to a contemporary of Confucius (6th–5th century BC) named Lao-tzu. Current scholarly opinion prefers to regard it as a compilation, which reached its present form only in the 3rd century BC. Rather than the work of a single author, it is now considered to be an anthology of sayings drawn from various currents of thought. Some would even deny that it reflects a distinctive point of view. Although its historical enigmas are apparently insoluble, there is abundant testimony to the vast influence exercised by the book since the earliest times and in surprisingly varied social contexts. Among the classics of speculative Taoism, it alone holds the distinction of having become a scripture of the esoteric Taoist movements, which developed their own interpretations of its ambiguities and transmitted it as a sacred text (see also LAO-TZU).

Chuang-tzu. In turning from the solemn profundities of *Tao-te Ching* to the other great classic of Taoist speculation, the *Chuang-tzu*, a different world is entered. Here there is none of the *Tao-te Ching*'s studied density. Instead, this large book opens with a sprightly fable, illustrating the incomprehension of small wildfowl of the majestic splendour of a gigantic bird. Other such parables demonstrate the relativity of all values: the sliding scales of size, utility, beauty, and perfection. There is a colloquy between the Lord of the Yellow River and the God of the Eastern Ocean, in which the complacent self-satisfaction of the lesser spirit is shaken by his unexpected meeting with inconceivable vastness. Humble artisans are depicted, who, through the perfect mastery of their craft, exemplify for their social superiors the art of mastering life. Life and death are equated, and the dying are seen to welcome their approaching transformation as a fusion

with the Tao. A succession of acquiescent cripples exclaims in rapture on the strange forms in which it has pleased heaven to shape them. Those involved in state ritual are brought onstage only to be mocked, and the propositions of contemporary logic-choppers are drawn into the unending whirl of paradox, spun out to their conclusions, and so abolished. Such are a few aspects of this wild kaleidoscope of unconventional thought, a landmark in Chinese literature. Its concluding chapter is a systematic account of the pre-eminent thinkers of the time, and the note of mock despair on which it closes typifies the *Chuang-tzu's* position regarding the more formal, straitlaced ideologies that it parodies.

The Taoist Immortals

Among the strange figures that people the pages of *Chuang-tzu* are a very special class of spiritualized being. Dwelling far apart from the turbulent world of men, dining on air and sipping the dew, they share none of the anxieties of ordinary folk and have the smooth, untroubled faces of children. These "supreme men," or "perfect men," are immune to the effects of the elements, untouched by heat and cold. They possess the power of flight and are described as mounting upward with a fluttering (*hsien*) motion. Their effortless existence was the ultimate in autonomy, the natural spontaneity that *Chuang-tzu* ceaselessly applauds. These striking portraits may have been intended to be allegorical, but whatever their original meaning, these Immortals (*hsien*), as they came to be called, were to become the centre of great interest. Purely literary descriptions of their freedom, their breathtaking mobility, and their agelessness were construed as practical objectives by later generations. By a variety of practices, men attempted to attain these qualities in their own persons, and in time *Chuang-tzu's* unfettered paragons of liberty were to see themselves classified according to kind and degree in a hierarchy of the heavenly hosts (see also CHUANG-TZU).

Taoist thought in eclectic texts. In the welter of compilations that represent the opinions of the late Warring States and early imperial times (*c.* 300–100 BC), sayings of Taoist thinkers or passages from Taoist texts are often drawn upon. Even the Confucian *Lun-yü* (*Analects*) did not escape this genial contamination; among the sober sayings of the Sage are scattered a number of Taoist tales, in which crusty recluses or inspired madmen make fun of Confucius, precisely as happens in *Chuang-tzu*. Curiously, the ontology of *Tao-te Ching* proved particularly attractive to the proponents of the cynical "realist" or "legalist" theories of authoritarian statecraft. The *Han-fei-tzu* (3rd century BC) cites long extracts from the *Tao-te Ching*, which it interprets in its own Draconian manner, while the contemporary *Kuan-tzu* quotes other, unspecified Taoist writings and turns them to like account. The pervasiveness of Taoist thought is revealed in two great encyclopaedic miscellanies: the *Ch'un Ch'iu* ("Spring and Autumn" Annals), produced at the court of the King of Ch'in (*c.* 240 BC), and still more, in the *Huai-nan-tzu*, from the circle of Liu An, the Han dynasty Prince of Huai-nan (died 122 BC). The *Lieh-tzu* (compiled 3rd century AD from older materials) is in a rather special position; it shares 18 anecdotes with the *Chuang-tzu* and takes as its title the name of a mysterious figure mentioned in that work. But the last of its eight chapters expounds the doctrine of a not noticeably Taoist figure, the arch-hedonist Yang Chu. There is no lack of later collections, inevitably ascribed to the most archaic authors, which cite and expand upon one or another of the old Taoist classics. *Wen-tzu*, *Kuei-ku-tzu*, and *Ho-kuan-tzu* are all among such repositories of secondhand wisdom (see also CONFUCIAN TEXTS, CLASSICAL).

Commentaries on the *Tao-te Ching*

The scholiasts. The most famous of the many commentaries on *Tao-te Ching* was written by Wang Pi (AD 226–249). He is regarded as a founder of the school of Dark Learning (*hsüan-hsüeh*), a highly conservative philosophical movement that enjoyed a certain vogue among the cultured elite of the 3rd and 4th centuries. The *Chuang-tzu* was not long afterward annotated by Kuo Hsiang (died 312), in whose work the fundamental Confucian bias is even more prominent. The writings of these men have in recent years sometimes been called "Neo-Taoism," but nothing could be more misleading. Their primary aim was to harmonize *Tao-te Ching* and *Chuang-tzu* with their own conception of a practical life devoted to affairs of state. As administrators confronted with the challenge of Taoist thought, they preferred not to take its message at face value. Interpretative commentaries continued to be written on the classics of speculative Taoism in which the aid of the most diverse philosophies was called upon, not excluding Buddhism. Like the work of the 3rd and 4th century scholiasts, these represent the ideas of a tiny minority, the members of the scholar-official class. Though excursions into ever more refined scholasticism continued to be a diversion for them, the real creative vitality of Taoism was to be found elsewhere.

THE LITERATURE OF TAOIST ESOTERISM

Early hagiography of the Immortals. *Lives of the Immortals.* By the Han period, the careers of those free spirits described in *Chuang-tzu* were the subject of universal interest. The earliest systematic collection of biographical notices on these legendary figures is the *Lives of the Immortals* (*Lieh-hsien chuan*) of the early 2nd century AD. Such collections were a genre of the time. Brief sketches were provided for 72 figures: the same symbolic number as was found in contemporary collections of the "Lives" of the disciples of Confucius, eminent scholar-officials, and famous women. Thus Immortals came to be classified as yet another category in the highly stylized gallery of ancient worthies. Each notice is followed by a short hymn of praise. This was the standard form of inscriptions on stone; its employment in hagiographic literature may have influenced the later development of the chantefable in alternating passages of prose and verse. The text appears to reflect a growing number of local cults dedicated to individual Immortals, while the many plants mentioned suggest the extent of the use of herbal compounds as a means to transcendence.

Inscriptions. These literary notices are supplemented by epigraphic evidence, inscriptions on stone or bronze. The simplest of these are bronze mirrors depicting the plumed figures of airborne Immortals and bearing short rhyming texts of a general nature. Longer and more explicit are the texts of inscriptions on stone: tablets dedicated to the cult of a particular Immortal. They open with their subject's vital statistics, list his latter-day manifestations, and commemorate offerings made in his honour. But all of this is only by way of preface to the core of the inscription, in which his merits are celebrated in verse. Such was the eloquent votive tablet erected in honour of Wang Tzu-ch'iao, a perennial favourite among the Immortals (AD 165). Another, dedicated to Lao-tzu in the same year, describes the supposed author of the *Tao-te Ching* as a god, to whom worship had been paid by the then reigning emperor.

Lao-tzu as a god

Texts on the cult of Lao-tzu. One of the most complex and interesting phenomena in Chinese religious history is Lao-tzu's advancement from sage to god. A scroll found in the walled-up desert library at Tun-huang, the *Book of the Transformations of Lao Tzu* (*Lao-tzu Pien-hua Ching*), shows him in cosmic perspective, omnipresent and omnipotent, the origin of all life. His human manifestations are listed, followed by his successive roles in legendary history, as the sage counsellor of emperors. Next, five of his more recent appearances are mentioned, dated AD 132–155, and localized in west China, where a temple is said to have been dedicated to him in 185. Then the god speaks, to describe his own powers. He recommends to his votaries the recitation of "my book in 5,000 words" (the *Tao-te Ching*) and enjoins a meditation on his own divine attributes as they appear within the adept's body. Finally, he calls upon the faithful to join him, now, when he is about to strike at the tottering rule of the Han dynasty. Evidently the product of a messianic group in west China at the end of the 2nd century, this valuable fragment of only 95 lines is written in a strangely disfigured Chinese, in part a reflection of its popular milieu. But it still shows more clearly than many longer and better preserved texts the essential cohesion of the several aspects of esoteric Taoism: hagiography, recitation of

scriptures, and visionary meditation, all of which are here given additional temporal unity by the messianic context.

The most ancient doctrinal literature. *The Book of the Great Peace.* Just as hagiography was conceived in terms of the thought of its time, the earliest texts of esoteric Taoism that explicitly treat of doctrinal matters also fully share the holistic view of the universe (*i.e.*, stressing the organic relations between all things), which was the heritage of all traditional Chinese thinkers. But Taoists looked to an accelerated union with the universal Tao in their own lifetime, striving inwardly for its realization within their bodies and socially for its advent in their communities of initiates. One of the earliest known of these organizations was the Way of the Great Peace (*T'ai-p'ing tao*), which attracted a huge following in northeast China during the latter half of the 2nd century AD. There is an enormous *Book of the Great Peace* (*T'ai-p'ing Ching*) that may in part represent the teachings of this sect. Its style is archaic-seeming, and the frequent difficulties of interpretation may reflect northeastern regionalisms. It seems clear, too, that the work was composed at a social level rather lower than that of the highly literate metropolitan classes. Its contents are too difficult, diverse, and prolix to be easily summarized. It is written in the form of a colloquy between a "celestial master" (*t'ien-shih*) and his disciples, called "perfected-ones" (*chen-jen*). The dialogue form provides a dramatic framework in which the more-than-mortal preceptor can make his points with great effectiveness. He interrogates his anxious pupils, expressing studied astonishment when they fail to comprehend basic issues, which he then proceeds to expound. This didactic exchange is in the best tradition of the old philosophical schools, but the subject is now the correspondence and interaction of the mortal plane with the invisible world, and the attainment of transcendence.

Texts of the Celestial Masters. The protagonist of the *Book of the Great Peace* is a celestial master. When another important religious movement began in China's far west at about the same time as the group in the northeast arose, in the second half of the 2nd century AD, the same title was given to its founder, Chang Tao-ling. It is with this "Way of the Celestial Masters" (*T'ien-shih tao*) that the history of organized religious Taoism may be said to begin, in that there has been an unbroken continuity to the present day, as the movement soon spread to all of China. A few surviving works appear to belong to the organization's early years. Its central preoccupation has always been liturgical, the formalized communication with the celestial hierarchy. Liturgical fragments are preserved in 4th-century texts and presumably derive from much older practice. Liturgy was not static; it continued to grow, and liturgical texts fill many volumes of the Taoist Canon. Its full development is represented at the present time in the rites of the Taoist masters on Taiwan. The intoned liturgy has at all times been the communal expression of faith, furnishing a rhythmical support to all personal spiritual aspirations, and a backdrop for much of Taoism's subsequent literature. Still another text discovered at Tun-huang reveals a further aspect of the sect's practice. The *Hsiang-erh Commentary on the Lao Tzu*, composed by one of the first celestial masters, shows the way in which the faithful were taught to interpret and apply the *Tao-te Ching*. Lao-tzu, as Lord Lao the Most High (*T'ai-shang Lao-chün*), is seen as the concrete manifestation of the universally diffused Tao; his book is envisaged as a revelation from which clear rules of conduct are to be derived. This sectarian interpretation, then, was a predominantly moral one, and it is known that the *Tao-te Ching* continued to be transmitted as a sacred text to beginners in the "Way of the Celestial Masters."

The Book of the Yellow Court. Very different in form is the *Book of the Yellow Court* (*Huang-t'ing Ching*), also among the earliest and most influential of Taoist scriptures. It is written in verse of seven syllables to the line, rhyming in successive lines: a rather inept prosody, reminiscent of the doggerel verse on Han mirrors, as well as certain secular didactic stanzas from the 2nd century AD. The work is imperfectly understood, but what it sets forth in often puzzling detail is a procedure of systematic communication with the internal counterpart of the celestial hierarchy. The book's patron is, once again, Lao-tzu. As indicated above, he was as much at home within the bodies of his worshippers as in the heavens. The mnemonic verses were presumably to be chanted as the adept turned his concentrated gaze within: it is in fact an inward-looking liturgy (see also TAOISM; TAOISM, HISTORY OF).

Pao-p'u-tzu. It would appear that esoteric Taoism became more significant among the cultured elite in the 4th century AD. With this upward move in society, works of conscious literary art appear. The most celebrated and elaborate treatise of the period is Ko Hung's *Pao-p'u-tzu* ("He Who Holds to Simplicity"), the author's pseudonym. This voluminous collection contains 70 extended essays, of which the first 20, the "Inner Chapters," deal with the arts of immortality. In the debate on the existence of Immortals that had been waged in literate circles sporadically since Han times, Ko Hung was the most valiant champion of the Immortals' claims. But he opposed the notion that they belonged to a radically different order of beings. Rather, he stressed that all known Immortals had themselves once been men and so their state must be attainable by men today. The essentials are a good teacher and tireless perseverance. The argument is presented in the most refined but fluent prose and, oddly enough, in the dialectic form perfected by the Han skeptics. A question is put forward or an objection proposed by an imaginary opponent; this is then decisively dealt with by "Pao-p'u-tzu," who draws upon a great store of anecdotes to illustrate his own nonskeptical beliefs. His topics cover the whole range of the macrobiotic preoccupations of his day: breath control, herbalism, dietetics, sexual techniques, and alchemy. To mineral elixirs he accorded a place above all other means of attaining everlasting life, and the two long essays on the subject in *Pao-p'u-tzu* are among the earliest and most detailed documents in the history of alchemy. The 70th chapter of this remarkably rich work is Ko Hung's own autobiography, as fascinating as the rest of his book. But *Pao-p'u-tzu* is in large part a polemic exercise, intended either as a systematic refutation of unbelief or as encouragement to the adept whose enthusiasm was on the wane. For all his charm, however, the contradictory opinions that the author in his enthusiasm lets slip raise doubts as to his real understanding of the beliefs he so lustily defended. But there was soon to be more convincing literary testimony to the truly creative response that esoteric Taoism was awakening among members of the governing class. This dates from some 40 years later, and it began among Ko Hung's own kinsmen.

Techniques for achieving immortality

The Mao Shan revelations. Between 364 and 370 a young man named Yang Hsi experienced a series of visions in which he was visited by some dozen of the Perfected (*chen*), who were now formally classified as Immortals of higher rank than mere *hsien*. They presented him with a number of celestial texts, some of which were intended as more accurate versions of scriptures already in circulation among men. Among these was a new *Book of the Yellow Court*, in rather smoother verse than the older one and a good deal more intelligible. The Perfected also bestowed on Yang the texts of several of their own biographies: long narratives containing descriptions of the stages and methods by which they had achieved perfection. Their midnight interviews with the visionary were highly dramatic occasions, and Yang transcribed with care the allusive poems and ornate prose essays they dictated to him, as well as their explanations of the revealed scriptures. Most of these communications were destined for Yang Hsi's patron, Hsü Mi, a court official, and his son, Hsü Hui, who lived in retirement at Mao Shan, a chain of hills southeast of Nanking. The Hsüs recopied the revelations and also kept clear records of their own spiritual practices and dreams. So impressive were these extensive materials, both as examples of fine literary style and as tokens of divine favour toward the Hsü family, that much was preserved. Toward the end of the 5th century, T'ao Hung-ching, the greatest Taoist

Visions of Yang Hsi

master of his day and a relative of the Hsüs, collected all the surviving original manuscripts of the scriptures, lives, and visionary *dictées* (writings from dictation). He edited these *dictées*, with a commentary of his own, as *Declarations of the Perfected* (*Chen Kao*). The technical matter from the remainder of the revealed texts was included in his *Secret Instructions for Ascent to Perfection* (*Teng-chen yin-chüeh*), with an elaborate commentary treating all aspects of spiritual practice. This immense collection of dated documents, the most complete dossier of any early Taoist movement, is invaluable for the study of the period. The Mao Shan establishment that T'ao founded became the most powerful of Taoist organizations, and its literature had tremendous influence. Though he is not yet mentioned in histories of Chinese literature, the visionary Yang Hsi may well have been the most truly inspired poet of his time. The revealed documents show that he and the Hsüs expected the imminent descent of a saviour and the universal rule of the Tao, as an end to the turbulence and sufferings of the age. In a series of texts of great expressiveness and force, the Perfected told of the approaching end of the old order; they also set forth the spiritual discipline that would assure survival through the cosmic upheavals ushering in the new era and would guarantee exalted rank in the kingdom of the elect. This revelation was the highest literary achievement of esoteric Taoism's formative period, and its impact was considerable. There were numerous imitations of the scriptures, while many ostensibly independent works in the Taoist Canon are no more than extracts from *Chen Kao*.

The Ling-pao movement. The chief influence on liturgical tradition, however, came into being slightly after the Mao Shan revelations. At the very end of the 4th century, a member of Ko Hung's own clan composed the Ling-pao ("Transcendent Jewel") scriptures, which brought upon the scene a series of Celestial Reverences (*t'ien-tsun*). The supreme and primordial character of these figures may owe something to the stately succession of Buddhas, from remotest times into the incalculable future, found in the Mahāyāna (the "Greater Vehicle," a form of Buddhism predominant in East Asia) scriptures. But these hierarchs are described as emanations of the Tao, and their liturgies came to occupy a prominent place in communal Taoist worship from the 5th century on. The question of possible Buddhist influence upon the liturgical portion of this movement, too, remains open, and the whole subject requires much more research.

Subsequent developments. The influential Mao Shan and Ling-pao movements began in south China, at a time when the north was under foreign domination. Their literature soon spread northward, where independent developments had taken place. The need for a comprehensive collection of information on all the schools resulted in the first great Taoist encyclopaedia (*c.* 570). Like other such compilations in China, it was made up of extracts from sundry books, classified by subject matter. Similar reference works continued to be compiled. The most important is the *Seven Slips from the Bookbag of the Clouds* (*Yün-chi ch'i-ch'ien*) (*c.* 1022), made just after the first printing of the Taoist Canon in about 1016. It is a canon in miniature and contains many important works in their entirety. Hagiography continued to flourish. In addition to many local and sectarian compilations, there were huge general collections, containing the lives of both legendary and historical figures, such as the immense *Comprehensive Mirror of the Immortals* (*Chen-hsien t'ung-chien;* early 12th century). Sectarian historiography also developed; of particular interest are the extensive monographs devoted to the great mountain centres of Taoism. The *Treatise on Mao Shan* (*Mao Shan Chih*) (1329) is among the most monumental. It includes lives of the saints and patriarchs, notes on topography and history, and a valuable selection from 1,000 years of literary testimony and inscriptions on the mountain and its Taoism. The new Taoist movements, which took north China by storm in the 12th and 13th centuries, also furnish their own very copious literature: biographies of their masters and collections of their sayings. Among them is the famous account of the travels (1220–24) of a patriarch of the Ch'üan-chen sect into Central Asia in response to the summons of Genghis Khan. Short moral tracts for missionary purposes were yet another popular genre, and, finally, there are innumerable inscriptions from all periods that provide important data on Taoist establishments and their patrons over the centuries.

Taoist reference works

AESTHETICS AND INFLUENCE OF TAOIST LITERATURE

Taoist imagery. Taoist literature manifests such richness and variety that scholars tend naturally to seek the symbolic modes of expression which served as points of unity within its historical diversity. No image is more fundamental to all phases of Taoism than that of the child. *Tao-te Ching* praises the infant's closeness to the Tao in its freedom from outside impressions, and *Chuang-tzu* describes the spiritual beings nurtured on primal substances, air and dew, as having the faces of children. Thus many of the spirits, both indwelling and celestial, in the esoteric system are described as resembling newborn babes, while the Immortals who appear in visions, though hundreds of years old, are at most adolescent in appearance. Other persistent images are those of mountain and cavern. Present in the older texts, they are carried over, with particular connotations, into the later works. The mountain as a meeting place of heaven and earth, gods and men, and master and disciple (as already in *Chuang-tzu*), takes on a vast downward extension. Beneath the mountains are the great "Cavern-heavens" (*tung-t'ien*) of esoteric Taoism, staffed by a numerous immortal hierarchy. Thus, for example, while Mao Shan is only some 400 metres (1,300 feet) high to the gaze of the profane, the initiate knows that its luminous grottoes plunge thousands of metres into the earth. And *light* is everywhere in Taoist revelation: spirits and paradises alike gleam with brilliance unknown in the world of men.

The image of the child

Influence on secular literature. Already during the Warring States period and the early Han, Taoism had made its appearance in the works of the other schools. Both direct quotations and patent imitations were frequent, and citations from *Tao-te Ching* and *Chuang-tzu* abound throughout later Chinese literature, as do reminiscences of both their style and their content. Esoteric Taoist writings, too, held great fascination for men of letters. Their response might vary from a mere mention of the most celebrated Immortals to whole works inspired directly by specific Taoist texts and practices. Many a poet recorded his search, real or metaphorical, for Immortals or transcendent herbs, or described his attempts at compounding an elixir. A certain number of technical terms became touchstones of poetic diction. The revealed literature of Mao Shan came to have the greatest effect on secular writings. As works of great literary refinement, the *Lives* of the Perfected directly inspired a very famous tale, the *Intimate Life of Emperor Wu of Han* (*Han Wu Ti nei-chuan*; late 6th century), which in highly polished terms describes the visit to the emperor of a goddess, the Queen Mother of the West. This work, in turn, made a decisive contribution to the development of T'ang romantic fiction. Literary accounts of fantastic marvels also drew heavily on the wonders of Mao Shan hagiography and topography. The Mao Shan influence on T'ang poetry was no less important. Precise references to the literature of the sect abound in the poems of the time, while many of the greatest poets, such as Li Po, were formally initiated into the Mao Shan organization. As awareness of these influences increases, scholars are faced with the intriguing question of the possible religious origins of whole genres of Chinese literature (see also LITERATURE, EAST ASIAN).

Influence of Mao Shan literature

Influence on the arts. A number of early Chinese books of spiritual interest claim to have been inspired by pictures seen on the walls of local temples. A similar tradition attaches to the *Lives of the Immortals*, which is said to derive from a pictorial work called *Portraits of the Immortals*. As has been noted, the Immortals were

depicted on Han mirrors. Other illustrative materials were in close relation to the earliest esoteric Taoist literature. Graphic guides existed from early times to aid in the identification of sacred minerals and plants, particularly mushrooms. A later specimen of such a work is to be found in the Taoist Canon. This practical aspect of Taoist influence resulted in the exceptionally high technical level of botanical and mineralogical drawing that China soon attained. In calligraphy, too, Taoists soon set the highest standard. One of the greatest of all calligraphers, Wang Hsi-chih (303–361), was an adherent of the Way of the Celestial Master, and one of his most renowned works was a transcription of the *Book of the Yellow Court*. The efficacity of talismans, in particular, depended on the precision of the strokes from which they were created. Figure painting was another field in which Taoists excelled. China's celebrated painter Ku K'ai-chih, a practicing Taoist, left an essay containing directions for painting a scene in the life of the first Celestial Master, Chang Tao-ling. Many works on Taoist themes, famous in their time but now lost, have been attributed to other great early masters. Of these, some may have been painted for use in ritual, and religious paintings of the Taoist pantheon are still produced today. The Taoist scriptures, with their instructions for visualization of the spiritual hierarchy, including details of apparel and accoutrements, are ready-made painter's manuals. Finally, the language of speculative Taoism was pressed into service as the basic vocabulary of Chinese aesthetics. Consequently, many secular artists attempted to express their own conceptions of the "natural spontaneity" of *Chuang-tzu*, and *Lao-tzu*'s "spirit of the valley." Here Taoism found still wider imaginative extension, and the efforts of these painters are embodied in those magnificent landscapes that have come to be thought of as most characteristically Chinese.

BIBLIOGRAPHY

Speculative Taoism: M. KALTENMARK, *Lao Tzu et le taoïsme* (1965; Eng. trans., 1969), a general introduction stressing speculative Taoism; A. WALEY (ed. and trans.), *The Way and Its Power* (1934), a classic interpretation and translation of the *Lao Tzu*; B.D. WATSON, *Chuang-Tzu: Basic Writings* (1968), a readable complete translation.

Esoteric Taoism: H. MASPERO, *Le Taoïsme* (1950), the pioneering description of esoteric Taoism, still not entirely superseded; M. KALTENMARK, *Le Lie-sien tchouan: biographies légendaires des immortels taoistes de l'antiquité* (1953), a richly documented translation of the first classic of Taoist hagiography; A.K. SEIDEL, *La Divinisation de Lao Tzeu dans le taoïsme des Han* (1969), the most important work on the beginnings of Taoism in Han thought—treats the oldest literature of the movement; K.M. SCHIPPER, *L'Empereur Wou des Han dans la légende taoiste* (1965), unique as a study of esoteric Taoist imaginative literature; A. WALEY, *The Poetry and Career of Li Po, 701–762* (1950), the life of a Taoist poet; *The Travels of an Alchemist* (1931), translation of the account of a Taoist's Central Asian journey.

(M.St.)

Tapestry

Historically almost any heavy material, handwoven, machine woven, or even embroidered, used to cover furniture, walls, or floors or for the decoration of clothing, has been called tapestry in popular usage. Since the 18th and 19th centuries, however, the technical definition of tapestry has been narrowed to include only heavy, reversible, patterned or figured handwoven textiles, usually in the form of hangings or upholstery fabric.

Tapestry traditionally has been a luxury art afforded only by the wealthy, and even in the 20th century large-scale handwoven tapestries are too expensive for those with moderate incomes.

Tapestry sets

Tapestries are usually designed as single panels or sets. A tapestry set is a group of individual panels related by subject, style, and workmanship and intended to be hung together. The number of pieces in a set varies according to the dimensions of the walls to be covered. The designing of sets was especially common in Europe from the Middle Ages to the 19th century. A 17th-century set, the "Life of Louis XIV," designed by the king's painter Charles Le Brun, included 14 tapestries and two supplementary panels. The number of pieces in 20th-century sets is considerably smaller. "Polynesia," designed by the modern French painter Henri Matisse, for example, has only two pieces, and "Mont-Saint-Michel," woven from a cartoon by the contemporary engraver and sculptor Henri-Georges Adam, is a triptych (three panels). Until the 19th century, tapestries were often ordered in Europe by the "room" rather than by the single panel. A "room" order included not only wall hangings but also tapestry weavings to upholster furniture, cover cushions, and make bed canopies and other items. Most Western tapestry, however, has been used as a type of movable monumental decoration for large architectural surfaces, though in the 18th century, tapestries were frequently encased in the woodwork.

In the West, tapestry traditionally has been a collective art combining the talents of the painter, or designer, with those of the weaver. The earliest European tapestries, those woven in the Middle Ages, were made by weavers who exercised much of their own ingenuity in following the cartoon, or artist's sketch for the design.

Though he followed the painter's directions and pattern fairly closely, the weaver did not hesitate to make departures from them and assert his own skills and artistic personality. In the Renaissance, tapestries increasingly became woven reproductions of paintings, and the weaver was no longer regarded as the painter's collaborator but became his imitator. In medieval France and Belgium, as well as now, a painter's work was always executed in tapestry through the intermediary of the weaver. Tapestry woven directly by the painter who created it remains an exception, almost exclusive to ladies' handiwork.

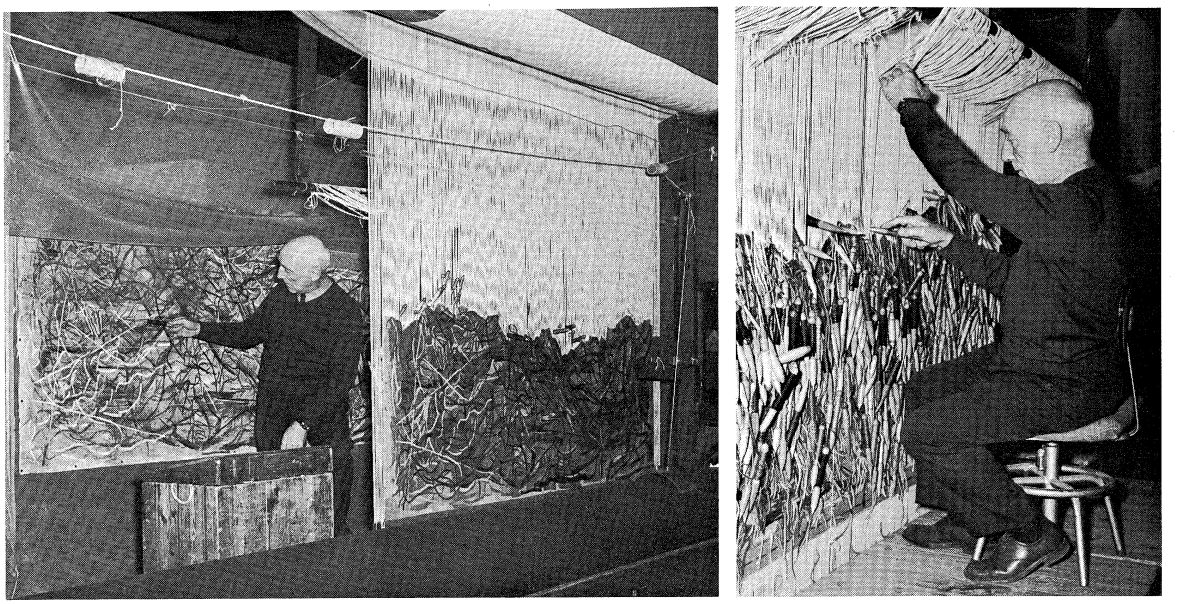
MATERIALS

Wool has been the material most widely used for making the warp, or the parallel series of threads that run lengthwise in the fabric of the tapestry. The width-running, weft, or filling threads, which are passed at right angles above and below the warp threads, thereby completely covering them, are also most commonly of wool. The advantages of wool have been its availability, workability, durability, and the fact that it can be easily dyed to obtain a wide range of colours. Wool has often been used with linen, silk, or cotton threads for the weft. These materials give variety and contrast of colour and texture and are better suited than wool to detail weaving or to creating delicate effects. In European tapestry, light-coloured silks were used to create pictorial effects of tonal gradation and spatial recession. The sheen of silk thread was often used for highlights or to give a luminous effect when contrasted to the dull and darkly coloured heavier woollen threads. In 18th-century European tapestries, silk was increasingly used, especially at the Beauvais factory in France, to achieve subtle tonal effects. Most of the Chinese and Japanese tapestries have both warp and weft threads of silk. Pure silk tapestries were also made in the Middle Ages by the Byzantines and in parts of the Near East. Wholly linen tapestries were made in ancient Egypt, while Copts, or Egyptian Christians, and medieval Europeans sometimes used linen for the warp. Cotton and wool were employed for pre-Columbian Peruvian tapestries as well as for some of the tapestries made in the Islāmic world during the Middle Ages. Since the 14th century, European weavers have used gold and silver weft threads along with wool and silk to obtain a sumptuous effect. These threads were made of plain or gilded silver threads wound in a spiral on a silk thread.

TECHNIQUES

Tapestry is first of all a technique. It differs from other forms of patterned weaving in that no weft threads are carried the full width of the fabric web, except by an occasional accident of design. Each unit of the pattern or the background is woven with a weft, or thread of the required colour, that is inserted back and forth only over the section where that colour appears in the design or cartoon. As in the weaving of plain cloth, the weft threads pass over and under the warp threads alternately and on the return go under where before it was over and

Weaving tapestry



Weaving on a high-warp loom.
(Left) Weaver indicating an area on the cartoon that corresponds to the portion of the tapestry (right foreground) he has woven. (Right) Weaver, working on the reverse side of the tapestry, pulls a set of warps forward and passes the bobbin behind it. (Photographs taken at the Gobelins factory, France.)
By courtesy of the Mobilier National, Paris; photograph, Visages de France

vice versa. Each passage is called a pick, and when completed the wefts are pushed tightly together by various devices (awl, reed, batten, comb, or serrated fingernails in Japan). The weft threads so outnumber the warps that they conceal them completely. The warps in a finished tapestry appear only as more or less marked parallel ridges in the texture, or grain of the fabric, according to their coarseness or fineness.

The thickness of the warp influences the thickness of the tapestry fabric. In Europe during the Middle Ages, the thickness of the wool tapestry fabric in such works as the 14th-century "Angers Apocalypse" tapestry was about five threads to the centimetre (ten to 12 to the inch). By the 16th century the tapestry grain had gradually become finer as tapestry more closely imitated painting. Known for the regularity and distinctness of its tapestries the royal French tapestry factory in Paris known as the Gobelins used six to seven threads per centimetre (15 to 18 per inch) in the 17th century and seven to eight (18 to 20) in the 18th century. Another royal factory of the French monarchy at Beauvais had as many as ten or even 16 threads per centimetre (25 to 40 threads per inch) in the 19th century. These excessively fine grains make the fabric very flat and regular, tending to imitate the canvas of a painting. The grain of 20th-century tapestry approximates that used in 14th- and 15th-century tapestry. The Gobelins factory, for instance, uses five or six threads per centimetre (12 or 15 per inch).

In many 20th-century tapestries a finer grain is contrasted with the effects of a heavier weave. The grain of silk tapestries, of course, is much finer than those made of wool. It is not uncommon for the silk tapestries of China to have as many as 24 warp threads per centimetre (about 60 per inch).

Where the weft margin of a colour area is straight and parallel to the warps, it forms a kind of slit, or *relais*, which may be treated in any of five different ways. First, it may simply be left open, as in Chinese silk tapestries, which are called *k'o-ssu* (cut silk) for that reason. Second, it may be left open on the loom but sewed up afterward, as in European tapestries from the 14th to the 17th centuries and also in some later types. Third, the weaver may dovetail his wefts, passing from one side and from the other in turn over a common warp. This may be either "comb" dovetailing—single wefts alternating—or "sawtooth" dovetailing—clusters first from one side, next from the other. Dovetailing has the double disadvantage of making the fabric heavier at that point and of blurring the outline. Persian weavers of the 16th century developed a successful variant in silk tapestry rugs whereby a black outline weft was dovetailed over two warps—one of each of the adjacent colour areas—effectively hiding the coloured wefts in the compacting of the weave and providing a strong clear image. The same device is found in pre-Columbian Peru.

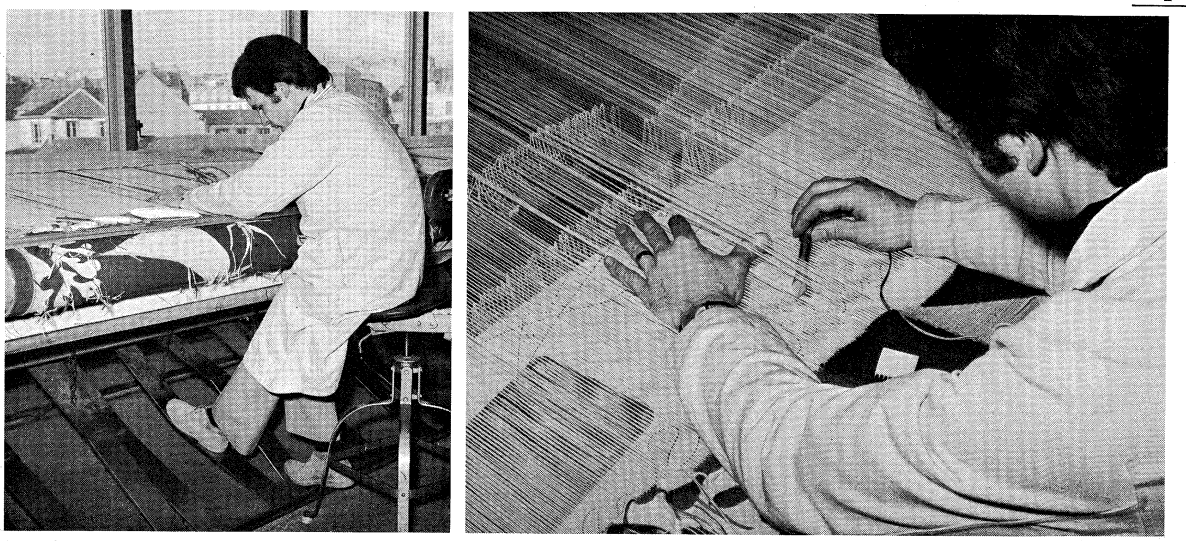
The fourth treatment—interlocking—was introduced in the Gobelins factory in the 18th century. Here wefts of juxtaposed colour segments are looped through each other between the two warps that mark, respectively, the margin of each colour. This gives a continuous surface of even weight and was prized by the French weavers because the resultant effect more closely approximated that of painting.

A curious variant of these weaving techniques is achieved when between every two rows of wefts there is a weft that runs the full width of the tapestry, thereby making the fabric solid. This technique, if strictly classified, would be called brocade weaving, but the principle is that of tapestry, with the cloth insert subordinate. Rarely used, the technique was employed in Japan in the 7th and 8th centuries, in eastern Persia in the 10th century, and in pre-Columbian Peru.

Twill technique of weaving

Instead of the plain-cloth method of weaving usually used in making tapestries, a twill technique can be used. In this type of weave the weft is floated over two or more warps, then under one or more warps, with this underpassage shifting always one to the right or left, thereby making a diagonal ribbing. As far as can be determined, this technique first appeared in medieval Persia and from the 17th century on was especially used in the Iranian provinces of Khorāsān and Kermān to make shawls of goat's hair or wool. It is also used to make the famed Kashmir shawls and, along with many other crafts, was probably introduced into Kashmir from Persia, in the 16th century. In contemporary European tapestries this technique, usually called eccentric weaving, occasionally has been used in making some of the experimental abstract hangings of the later 20th century.

European tapestry may be woven on either a vertical loom (high-warp, or *haute-lisse* in French) or a horizontal loom (low-warp, or *basse-lisse*). In early high-warp looms the warps were attached to a beam at the top, and groups of warp threads were weighted at the bottom. The weft was beaten up (*i.e.*, pushed) toward the top as the weaving progressed. Such looms are pictured on ancient Greek vases. In later high-warp looms the vertical frame



Weaving on a low-warp loom.
(Left) Weaver depresses the treadles with his feet, thereby controlling the space between warps through which the shuttle, held in his right hand, will be moved. (Right) Working on the reverse side of the tapestry with the cartoon positioned under the warps, the weaver passes the shuttle between the warps. (Photographs taken at the Beauvais factory, France.)
By courtesy of the Mobilier National, Paris; photograph, Visages de France

has heavy uprights holding a horizontal roller at top and bottom, on which the warps are stretched. Each warp passes through a loop of cord (the *lisses*), and the loops encircling the warps that correspond to uneven numbers are fastened to one slender cylinder; those to the even to another cylinder. Both cylinders are above the weaver but within reach so that he can pull forward with first one, then with the other set of warps (*i.e.*, form the shed) in order to pass his bobbin behind them. The bobbin (broche) is a short, pointed, slim cylinder of polished wood on which the weft yarn is wound.

The low-warp loom, on the other hand, has the rollers on the same level at table height so that the warps stretched between them are horizontal. To leave the weaver's hands free, the warps are attached to two slats, or poles, each of which is connected with a treadle so that the weaver's foot depresses the odd-numbered or even-numbered series of warps to form a passageway for the bobbin, called a shuttle on the low-warp loom. The cylinders in both instances serve to roll up the finished portion and unroll a further length of unwoven warps so that the section in process is always taut and in a convenient relation to the weaver. At both types of loom the weaver works from the back side, that is, he weaves the tapestry on the wrong side. He has, however, a hand mirror, which he puts through the unwoven warps holding it to reflect the right side of the portion in process. While the high-warp weaver can examine his finished work directly by walking round to the other side of his loom, the low-warp worker has to tilt up his frame.

Of the two techniques, low-warp is more commonly used. Of the great European tapestry works only one, Gobelins, has traditionally used high-warp looms. Several weavers can work simultaneously on either kind of loom. Depending on the complexity of the design and the grain or thickness of the tapestry texture, a 20th-century weaver at the Gobelins can produce three to seven square metres (about 32 to 75 square feet) a year.

The medieval cartoon

In Western tapestry the medieval cartoon, or preparatory drawing, was usually traced and coloured by a painter on a canvas the size of the tapestry to be woven. At the end of the 15th century the weaver probably wove directly from a model, such as a painting, and consequently copied not a diagrammatic pattern but the original finished work of the painter. At the beginning of the 17th century there arose a clear distinction between the model and the cartoon. The model was the original reference on which the cartoon was based. Cartoons were rapidly and freely used and were often copied.

More than one tapestry can be woven from a cartoon. At the Gobelins factory, for instance, the 17th-century "Indies" tapestry set was woven eight times, remade, and slightly altered by the late Baroque painter François Desportes (1661–1743); these cartoons were woven several more times during the 18th century.

The border of a cartoon was often redesigned every time it was commissioned, since each patron would have a different heraldic device or personal preference for ornamental motifs. Borders were frequently designed by an artist different from the one who conceived the cartoon for the central narrative or principal image. As an element of tapestry design, however, borders or frames were important only in European tapestry from the 16th to the 19th centuries. Medieval and 20th-century tapestries seldom use this device, which emphasizes the idea of the tapestry as a reproduction of or substitution for a painting.

Contemporary cartoons

A fully painted cartoon requires much of the painter's time and is tedious to make. In the 20th century, therefore, other solutions have been adopted. The cartoon may be a photographic enlargement of a fully painted model or, more simply, a numbered diagrammatic drawing. The latter type of cartoon was worked out by the famous French tapestry designer Jean Lurçat during World War II. In this method each number corresponds to a precise colour and each cartoonist has his own range of colours. The colours are not indicated in a photographic enlargement, but the weaver refers to a small colour model provided by the painter and from it makes a selection of wool samples.

The high-warp weaver has the full-size cartoon, which he follows as it hangs beside or behind him. The low-warp worker has the cartoon laid under the warps, so he follows it from immediately above. In both cases the main outlines are drawn with ink on the warps after they have been mounted, or attached to the loom. The design is executed, in all European work since the Middle Ages, at right angles to the loom, so that in the finished hanging the warps usually run horizontally rather than vertically as they ran on the loom. Though in certain pieces the warps run vertically, it is aesthetically advantageous for the tapestries to be executed horizontally, since the warp ribbing tends to create a texture more or less reinforced by linear shadows, which, if vertical, sever the design but if horizontal bind it into continuity. Practically, however, horizontal warps are disadvantageous, since the horizontal slits made in weaving will pull apart more rapidly than vertical slits because of the weight of the hanging.

PERIODS AND CENTRES OF ACTIVITY

Ancient Western world. Examples of tapestry weaving from the ancient world are so isolated and fragmentary as to make it uncertain either when or where the art

originated. The earliest known tapestry weaving was done in linen by the ancient Egyptians between 1483 and 1411 BC. Preserved by the dry desert climate of Egypt, three tapestry fragments were found in the tomb of Thutmose IV. Two of the fragments have cartouches of Egyptian pharaohs, and the third is a series of hieroglyphs. In the tomb of Tutankhamen (*c.* 1350 BC), a robe and glove woven by the tapestry technique have also been found.

Although no examples remain, writers of antiquity are unanimous in proclaiming the magnificence of Babylonian and Assyrian tapestries. Some scholars have speculated that the ancient Egyptians learned the art of tapestry from the ancient peoples of Mesopotamia. During that period when the few preserved Egyptian tapestry fragments were made, Mesopotamian ideas, techniques, and, perhaps, craftsmen were entering Egypt. These scholars conjecture that since tapestry weaving did not occur in quantity again in Egypt until the 4th century AD, it is likely that the craft was not indigenous.

Tapestry weaving continued to flourish in western Asia in the 1st millennium before Christ. Fragments of wool tapestries dating from the 4th or 3rd century BC have been found in graves in the U.S.S.R. near Kerch in the Crimean peninsula. The ornamental motifs of these fragments are of a widely diffused Hellenistic style that was especially prevalent in Syrian art at the time. Another fragment showing close Syrian connections is a piece of silk tapestry dating about 200 to 500 years later and found in China at Lou-lan in Sinkiang Uighur Autonomous Region. Other fragments have been found in Syria at the archaeological sites of Palmyra and Doura-Europus. If climatic conditions for textile preservation in the Near East had been more favourable, it might be possible to theorize that Syria was a great centre of tapestry weaving, especially at the start of the Christian Era.

Tapestry in ancient Greece and Rome

There are literary descriptions of the making of tapestry in ancient Greece and Rome. In the *Odyssey*, Homer (6th century BC?) describes Penelope working on a tapestry that was unravelled each night as she waited for Odysseus. The Roman poet Ovid (43 BC–AD 17) in the *Metamorphoses* describes the tapestry looms used by Minerva and Arachne in their mythological weaving contest. During the period of the empire the Romans apparently imported a considerable number of the tapestries used in their public buildings as well as in the homes of the wealthy. Since the Latin terms referring to tapestry and weaving are Greek in origin, it is generally supposed that the art of tapestry making was taught to the Romans by the Greeks.

Eastern Asia. Called *k'o-ssu* (cut silk), tapestry has long been produced in China, traditionally being made entirely of silk; Chinese tapestries are extremely fine in texture and light in weight. The weave is finished perfectly on both sides so that the tapestries are reversible. The warps are vertical in relation to the pattern, rather than horizontal as in European weaving. Sometimes the weaver uses metal threads to make his hangings more sumptuous or highlights the design by painting, although this is not considered a commendable expedient.

Many *k'o-ssu*, such as "T'ung Fung Stealing the Peaches of Longevity" (Ming dynasty, 1368–1644; Metropolitan Museum of Art, New York), imitated paintings and were mounted on scrolls or album leaves in the same manner as the pictures they copied. Tapestries to cover large wall surfaces, such as the *k'o-ssu* (2.20 by 1.75 metres; seven feet three inches by five feet nine inches) of "Feng-huang in a Rock Garden" (late Ming period, Metropolitan Museum of Art, New York), were usually brighter in colour, heavier in texture, and frequently woven with metal threads. Tapestry was also used to decorate furniture and clothing.

Early Chinese *k'o-ssu*

The earliest surviving examples of *k'o-ssu* date from the T'ang dynasty (AD 618–907). Eighth-century remains have been found in desert oases around Turfan in the Sinkiang Uighur Autonomous Region of China, and late T'ang fragments have been found in the Caves of the Thousand Buddhas (Ch'ien-fo-tung or Mo-kao-k'u) near the town of Tun-Huang in Kansu Province. It is thought that these weavings are probably not representative of the more fully developed *k'o-ssu* of the T'ang period because they show only simple repeating patterns of flowers, vines, ducks, lions, etc., and were found in relatively remote areas of Central Asia along the silk-trade route. In comparison is the more sophisticated 8th-century *k'o-ssu* that hangs in the Taima-dera, a temple near Nara, Japan. Based on the story of the T'ang dynasty priest Shan-tao, this four-square-metre (43-square-foot) weaving is the oldest known complete Chinese wall tapestry.



Chinese and Japanese tapestries.
(Left) "Horses," Japanese *tsuzure*, Edo period, late 18th century. In the Musée Historique des Tissus, Lyon, France. 70 × 65 cm.
(Right) "T'ung Fung Stealing the Peaches of Longevity," Chinese *k'o-ssu*, Ming Dynasty (1368–1644). In the Metropolitan Museum of Art, New York. 1.17 × 0.61 m.
By courtesy of (left) the Musee Historique des Tissus, Lyon, France, (right) the Metropolitan Museum of Art, New York, gift of Ellis G. Seymour, 1926

During the Sung dynasty (960–1279) the imperial family encouraged painting and patronized the art of tapestry. An important weaving centre was at Ting-chou in Hopeh Province. Under the Yüan dynasty (1279–1368) a government factory for weaving *k'o-ssu* was established at Hangchow (Lin-an) in Chekiang Province. Characterized by their rich ornamental designs, the Hangchow *k'o-ssu* were frequently woven with gold and silver thread. Examples of tapestry from the Ming period (1368–1644) are rare and exquisite. The *k'o-ssu* executed during the K'ang-hsi (1661–1722), or rule of the great Manchu emperor of China, Hsüan-yeh, are the finest tapestries produced during the Ch'ing dynasty (1644–1911). They are distinguished for their delicate colouring and the use of philosophical and religious themes. Later Ch'ing *k'o-ssu* has survived in great abundance and shows a decided artistic and technical decline. This is especially evident in the frequent use of painting to perfect design details in 19th-century *k'o-ssu*.

Japanese *tsuzure*

The tapestry technique travelled from China to Japan in the late 15th or early 16th century during the Muromachi (Ashikaga) period (1338–1573). Japanese tapestry called *tsuzure-nishiki* (polychrome tapestry) differs from the Chinese *k'o-ssu* in its more pronounced surface relief. This is achieved through the use of thick cotton weft threads covered with silk, gold, or silver thread.

Paralleling the great period of sumptuous brocade manufacturing, the production of *tsuzure* flourished during the Tokugawa (Edo) period (1603–1867), especially in the early 17th century and throughout the entire 18th century. These polychrome tapestries were primarily used to decorate garments and for wrapping gifts; on rare occasions they were also used as wall hangings. Although the tapestry industry declined in quality in the 19th century, it has been revitalized in the 20th century. Monumental wall hangings and theatre curtains are woven in the textile factories of Ōsaka and Kyōto by both traditional Japanese and European tapestry techniques.

The history of the art in Korea remains obscure. Rather coarse wool tapestry-woven rugs with stylized motifs, however, are still produced there.

Pre-Columbian Americas. The most skilled weaving in pre-Columbian America was achieved by the Andean



(Above) "St. Michael," detail from "Abraham and the Archangel Michael," Lower Saxony, mid-12th century. In Halberstadt Cathedral, East Germany. 1.10 × 10.26 m (whole tapestry).

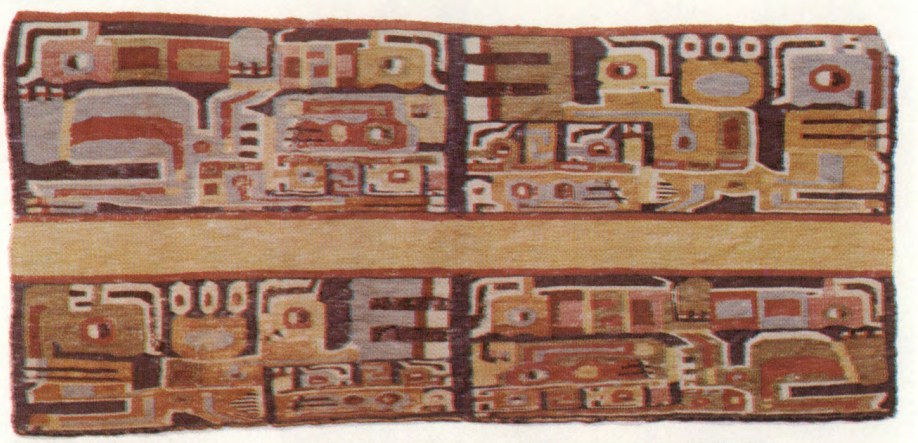


(Right) "Square with Nereids," Coptic, Egypt, 6th century. In the Louvre, Paris. 30 × 30 cm.



"Two Ducks," detail from the *k'o-ssy* panel "Feng-huang in a Rock Garden," Ming period, China (1368–1644). In the Metropolitan Museum of Art, New York City. 2.20 × 1.75 m (whole tapestry).

Early tapestries of East and West



Pre-Columbian fragment from the coast of Peru, Late Coastal Tiahuanaco period (1000–1300). In the pre-Columbian collection of Dumbarton Oaks, Washington, D.C. 34.4 × 16.5 cm.

Plate 1: By courtesy of (top left) the Institut fur Denkmalpflege, Halle, East Germany, (centre) the Metropolitan Museum of Art, New York, Seymour Fund, 1960, (bottom) Dumbarton Oaks, Washington, D.C.; photograph, (top right) Photo Hachette



Detail from "The Bear Hunt," one of the Devonshire hunting tapestries series, Brussels, second quarter of the 15th century. In the Victoria and Albert Museum, London. 4.90 × 10.15 m (whole tapestry).

Tapestries of the 14th and 15th centuries



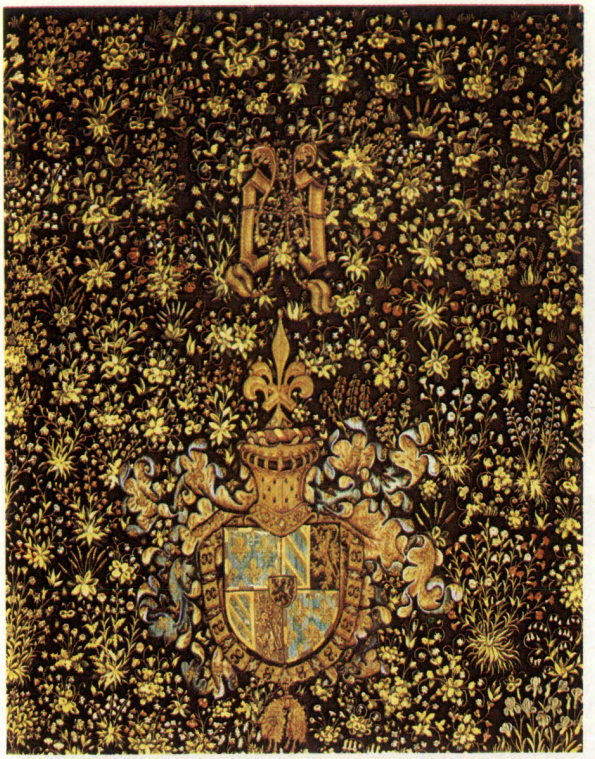


(Above) "St. Michael and the Dragon," from the set of seven tapestries in the "Angers Apocalypse," woven by Nicolas Bataille after the cartoons of Jean de Bandol, second half of the 14th century. In the Musée des Tapisseries, Angers, France. 2 m × 2 m 60 cm.

(Left) "The Annunciation," school of Arras, Flemish, early 15th century. In the Metropolitan Museum of Art, New York City. 3.46 × 2.90 m.

Plate 2: By courtesy of (bottom left) the Metropolitan Museum of Art, New York, gift of Harriet Barnes Pratt, 1949, in memory of her husband, Harold Irving Pratt; photographs, (top) J.R. Freeman & Co. Ltd., (bottom right) Photo Hachette

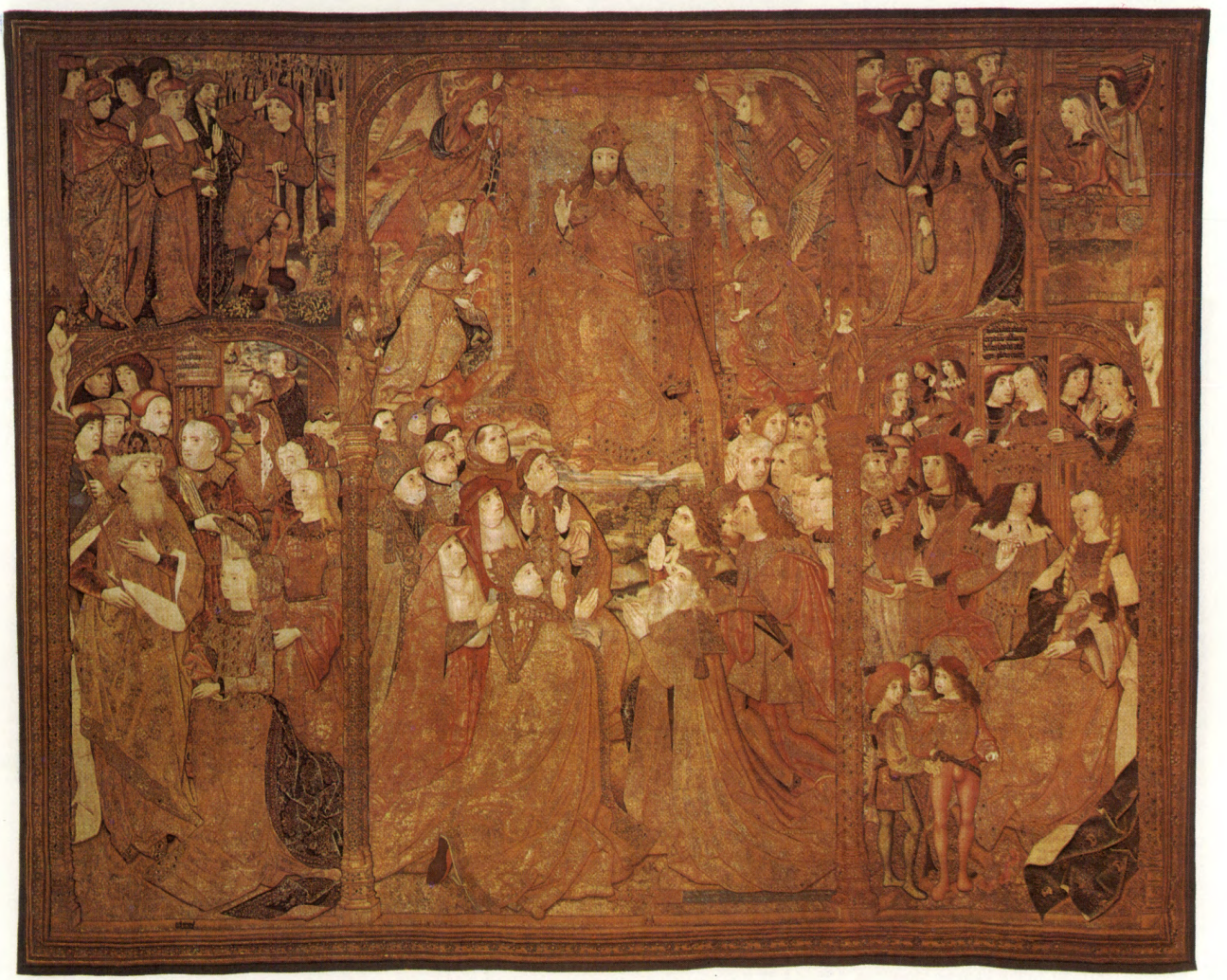
Plate 3: By courtesy of (top left) the Bernisches Historisches Museum, Switzerland, (bottom) the National Gallery of Art, Washington, D.C., Widener Collection; photograph, (top right) Lauros—Giraudon





(Above) "The Lady with the Unicorn," one of the six pieces of the tapestry, Loire workshop, late 15th century. In the Musée de Cluny, Paris. 3 × 3.10 m.

(Left) "Coat of Arms of Charles the Bold," Tournai, 1450–70. In the Bernisches Historisches Museum, Bern. 3.05 × 6.86 m (whole tapestry).



"The Triumph of Christ," known as the "Mazarin Tapestry," Brussels, *c.* 1500. In the National Gallery of Art, Washington, D.C. 3.30 × 3.90 m.



(Left) "Capture of Francis I," one of the panels of "The Battle of Pavia," Flemish workshop after cartoons by Bernard van Orley (1492?–1542). In the Museo e Galleria Nazionali di Capodimonte, Naples. 4.35 × 7.89 m.

(Below) Panel with grotesques woven by the workshop of N. Karcher after a cartoon by Bachiacca, *c.* 1550. In the Uffizi, Florence. 2.20 × 4 m.



Tapestries of the 16th, 17th, and 18th centuries



"The Miraculous Draught of Fishes," woven by Pierre d'Alost after a cartoon by Raphael, 1516–19. In the Vatican Museums. 4.95 × 4.39 m.



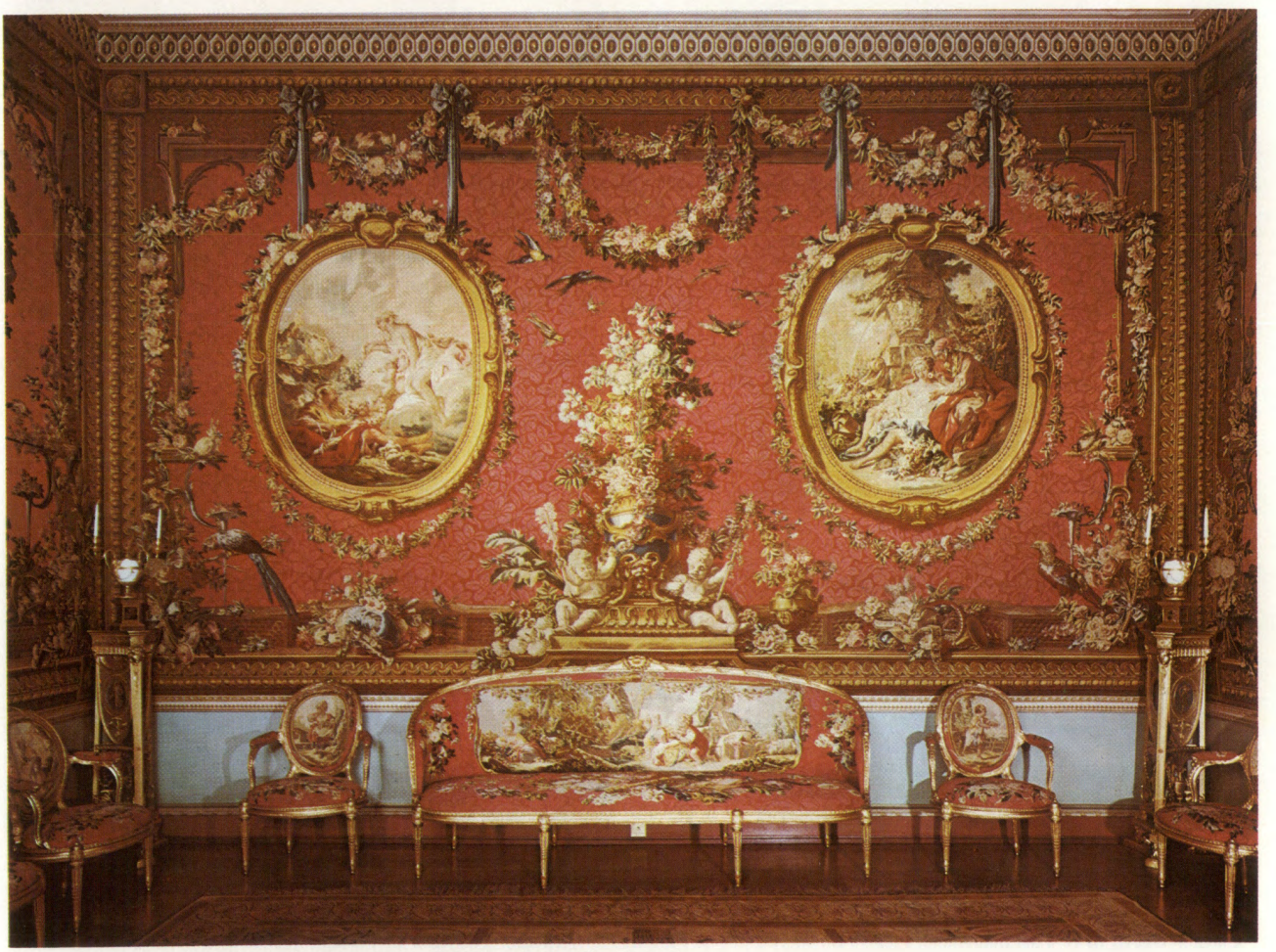
"Hercules and One of Diomedes' Stallions," from the set of four tapestries in "The Labours of Hercules," Oudenaarde, Belgium, second half of the 16th century. In the Louvre, Paris. 3.55 × 4.44 m.

Plate 4: (All but bottom right) SCALA, N.Y., (bottom right) Photo Hachette
Plate 5: By courtesy of (bottom) Osterley Park, London; photographs, (top left) SCALA, N.Y., (top right) J.E. Bulloz, (bottom) J.R. Freeman & Co. Ltd.





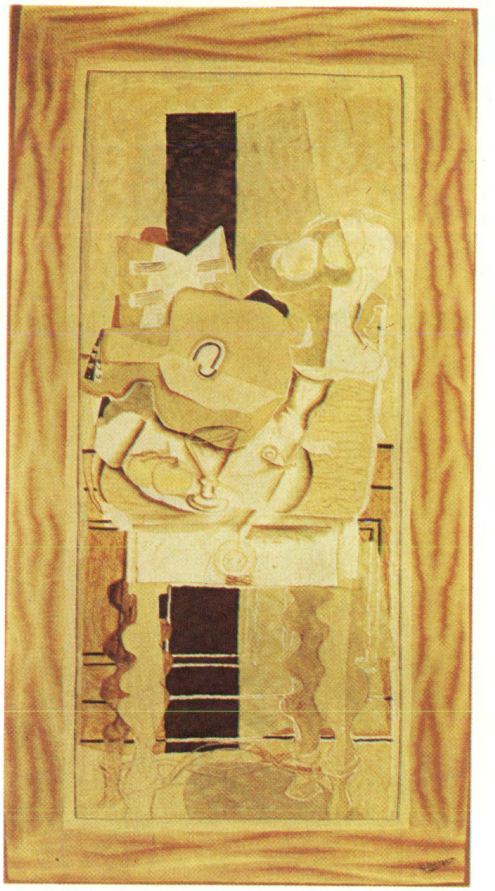
(Above) "The Triumph of Venus," one of four panels of "Marine Triumphs," workshop of Philippe Behagle, late 17th century. In the Banque de France, Paris. 4.35 × 2.60 m.
(Left) "El Cacharrero," woven by the Santa Bárbara Factory after a cartoon by Francisco de Goya, *c.* 1794. In the Palacio Nacional, Madrid. 3.75 × 2.60 m.



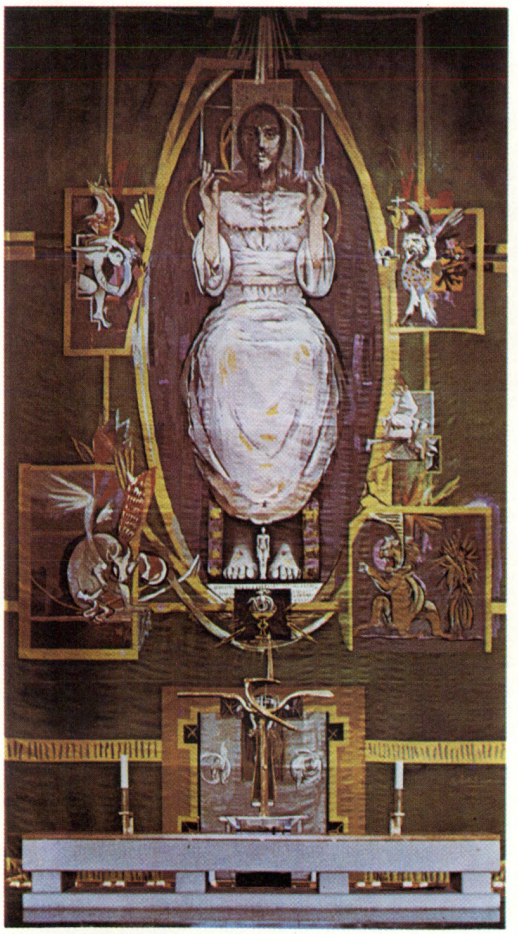
Wall, sofa, and chair tapestries *in situ* woven by the Gobelin workshop after François Boucher's "Loves of the Gods" and "Les Enfants Jardiniers," 1776. In Osterley Park, Middlesex. 3.67 × 6.25 m (wall tapestry).



"Scene from the Revelation of St. John," by Jean Lurçat, 1947. In Notre-Dame-de-Tout Grace, Plateau d'Assy, France. 5 × 18 m.

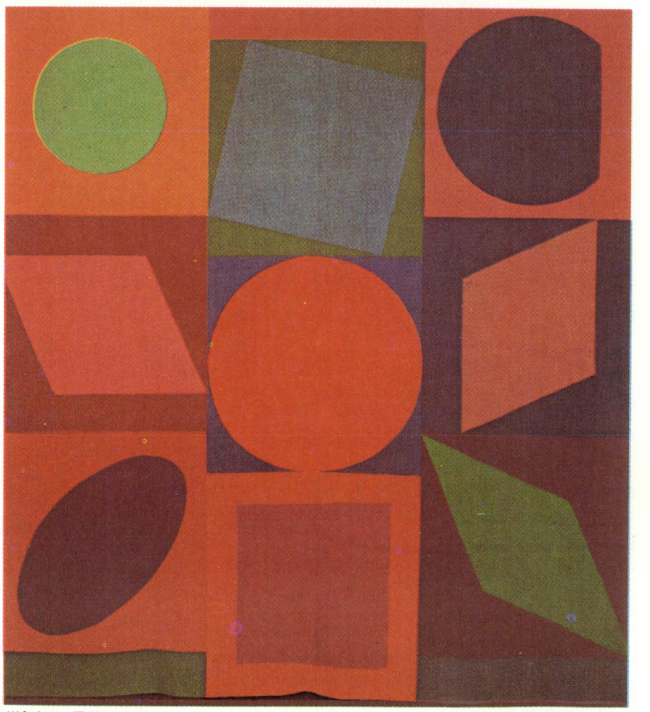


"Le Table et le pipe," by Georges Braque, 1932. In the Arts Club of Chicago. 2.11 × 1.12 m.



"Christ in Glory," by Graham Sutherland, 1962. In Coventry Cathedral, England. 23.94 × 12.05 m.

Tapestries of the 20th century



"Kalota T," woven by the workshop of Tabard after cartoon by Victor Vasarely, 1963. In the Galerie Denise René, Paris. 2.36 × 2.18 m.

Plate 6: By courtesy of (top right) the Arts Club, Chicago, (bottom right) Galerie Denise Rene, Paris; photographs, (top left) Dmitri Kessel, LIFE Magazine © Time Inc., (bottom left) Nicholas Servian—Woodmansterne Limited

Indian cultures of ancient Peru. The origins of tapestry weaving among these peoples are believed to date as early as the beginnings of the Christian Era. By the 6th and 7th centuries the technique of tapestry weaving was established, and a large number of pieces in this medium have survived, particularly from the 8th to the 12th centuries. Most of these tapestry weavings have been found in Peruvian coastal burial sites, where the dry desert climate prevented their deterioration. The dead were buried in clothes that display some of the most varied and skilled techniques of weaving and needlework ever current in any culture. Tapestry weaving was used principally to make garment decorations that were usually integral to the garment fabric. Narrow strips to ornament the edges of clothing were common, as were panels covering the entire surface of the *cuzma*, poncho-like Indian shirt. Fragments of tapestry wall hangings have also survived.

According to chronicles written by Spanish colonizers and scenes painted on ancient Peruvian pottery, weaving was generally done by women whose great manual skill made up for the simplicity of the looms, which are still used by Indian craftsmen. The workmanship was extremely fine. Certain tapestry fragments have been found with 60 to 100 weft threads per square centimetre (150 to 250 per square inch). The warps of the tapestries are of undyed cotton, being, therefore, either white or brown. The wefts are of wool from the llama, guanaco, alpaca, or vicuña, with cotton sometimes used to obtain bright white. The tapestries are usually polychrome, for the range of available colours made with natural dyes was large. Strong colour contrasts were preferred to the use of subtly graded tones of colours, especially in the Inca period (*c.* 13th to 16th century). Compositions tended to bold conventionalized designs often of human or animal figures and elaborate geometric patterns. Plant motives are comparatively rare.

After the Spanish conquest, looms from Spain were imported by the viceroyalty of Peru, and the weaving of tapestry was continued during the colonial period. The skilled Inca and later mestizo weavers evolved a curious blending of European influences and Indian traditions.

By courtesy of the Museum of Fine Arts, Boston



Colonial tapestry, late 16th–17th centuries. In the Museum of Fine Arts, Boston. 2.29 × 2.12 m.

Tapestry may also have been current in other developed pre-Columbian cultures of Central America and Mexico. Climatic conditions, however, have been destructive to textiles.

Middle Ages in Egypt and the Near East. Tapestry weaving was done by the Copts, or Egyptian Christians, from the 3rd to about the 12th century AD. Their tapestries are of great interest not only because of their artistic quality and technical skill but also because they are a bridge between the art of the ancient world and the art of the Middle Ages in western Europe. Fragments from the 5th to the 7th century are particularly numerous, and the largest number of examples have survived in the Egyptian cemetery sites of Akhmīm, Antinoë, and Saqqārah. As a result of a change in burial customs, perhaps attributable to Romanization and the widespread adoption of Christianity in Egypt, the ancient practice of mummification and its attendant ritual fell into disuse after the 4th century AD. The dead were subsequently buried in daily clothes or were wrapped in discarded wall hangings and tapestries. The clothing was ornamented with tapestry trimming, which was either woven into the fabric or attached to tunics and cloaks. Other burial furnishings included pillows and coverings. Tapestries were also used for the decoration of Christian churches, but few of these wall hangings have survived.

Coptic tapestry

By courtesy of (left) the Museum of Fine Arts, Boston, Ross Collection, (right) Yale University Art Gallery, the Hobart Moore Memorial Collection



Coptic and Persian tapestries.
(Left) Coptic Medallion tapestry with rabbit and grapes, Egyptian, 4th–7th centuries AD. In the Museum of Fine Arts, Boston. 21.5 × 21.0 cm. (Right) Fragment of Persian dovetailed tapestry with ibex, probably Sāsānian period (*c.* AD 224–651). In Yale University Art Gallery. 35.74 × 27.30 cm.

Coptic tapestries were woven with woollen wefts on linen warps, though a few with silk wefts have been preserved. Cotton wefts were occasionally used to obtain a brighter white. Primarily in the 7th century and perhaps also the 8th century, tapestry ornamentation was often supplemented by embroidery, as in border margins. In a special variant, which is not true tapestry, characteristic ornamental motifs such as meanders or other geometric repeats are executed with a free bobbin that follows the design without regard to consistency of weft direction.

Many of the early Coptic tapestries were done in a silhouette technique in which the motif or design was in a single dark colour, usually a tone of purple achieved by dying with madder and indigo, against a lighter background colour. After the 5th century, polychrome tapestries became increasingly common.

Many Coptic tapestry trimmings were woven with indigenous designs. Recurring motifs related to the ancient Egyptian funerary cult of Osiris and included the grape vine or ivy and the wine amphora. These motifs were considered appropriate to burial robes because of their relevance to revival in a life after death. Other favourite subjects were the hunter on horseback, boy-warriors, desert animals (especially the lion and the hare), creatures of mythology, dancing figures, and baskets of fruits and flowers. Christian subjects are as a rule late in date and are mostly figures of saints, standing or on horseback, against a red background. Depictions of biblical stories are rare. Some of the Coptic designs were copied, in a more or less distorted manner, from those woven into silk textiles imported from Syria.

Islāmic tapestry

After the invasion of Egypt by the Muslims in 640, the quality of Coptic tapestry began to deteriorate, although the industry continued to flourish by adapting itself to the tastes of the conquerors. During the Tūlūnid period (868–905) bands of tapestry trimming in wool or often in silk, occasionally with metal-thread enrichments, were

woven into white or dark green linen garments. In the Fātimid period (909–1171) silk tapestry weaving in golden yellow and scarlet became common. The motifs of the Islāmic period of Egyptian weaving were often interlacing geometric patterns frequently enclosing inscriptions or highly stylized small birds, animals, and flowers. Many of these inscriptions merely simulate writing, but many are legible. Giving religious phrases or the names and titles of rulers, they are in handsome angular Kufic scripts on earlier pieces and in cursive scripts later.

From the 6th to the 8th century AD, and doubtless from then on, striking wool tapestries were being made in Syria corresponding in style to the contemporary silk textiles with animals or birds in energetic heraldic stylization, framed in roundels, and almost always on a red ground. Later, from the 11th to the 13th century, highly distinctive silk- and gold-thread tapestries were produced in Syria incorporating pagan motifs from classical antiquity.

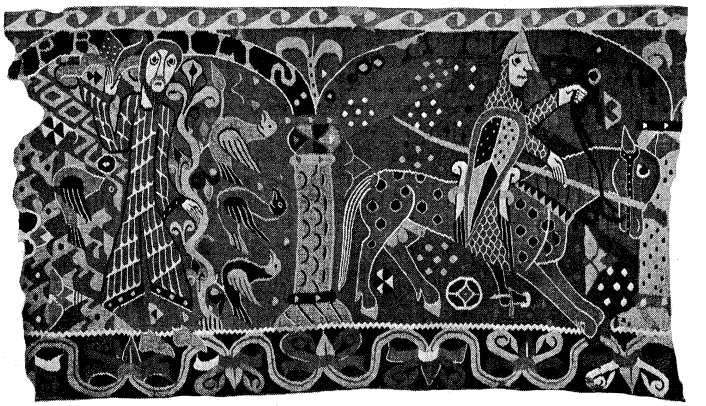
Fewer specimens of Persian tapestries have survived, but one notable fragment, now in the Moore Collection at Yale University, bears an ibex in the style of the Sāsānian period (*c.* AD 224–651). A single piece from the Seljuq period (11th century) established a continuation of the use of the tapestry technique, which reappears in the 16th century (intermediate examples apparently having all been destroyed) as the medium for rich silk- and metal-thread rugs, of which only three are known still to exist (also in the Moore Collection, New Haven, Connecticut), though others are illustrated in Persian miniatures. The modern descendants of these are kilims, or pileless carpets woven by the tapestry technique. Common to the entire Near East, these rugs are especially produced in the Caucasus and Asia Minor, as well as in parts of eastern Europe. Occasionally silk, they are more often wool with simple geometric patterns in bold colours.

Early Middle Ages in western Europe. Numerous documents dating from as early as the end of the 8th century describe tapestries with figurative ornamentation decorating churches and monasteries in western Europe, but no examples remain, and the ambiguity of the terms used to refer to these hangings makes it impossible to be certain of the technique employed. The 11th-century so-called Bayeux Tapestry depicting the Norman Conquest of England, for example, is not a woven tapestry at all but is a crewel-embroidered hanging.

Like the art of stained glass, western European tapestry flourished largely from the beginnings of the Gothic period in the 13th century to the 20th century. Few pre-Gothic tapestries have survived. Perhaps the oldest preserved wall tapestry woven in medieval Europe is the hanging for the choir of the church of St. Gereon at Cologne in Germany. This seven-colour wool tapestry is generally thought to have been made in Cologne in the early 11th century. The medallions with bulls and griffons locked in combat were probably adapted from Byzantine or Syrian silk textiles. The "Cloth of Saint Gereon" is thematically ornamental, but an early series of three tapestries woven in the Rhineland for the Halberstadt Cathedral were narrative. Dating from the late 12th and early 13th centuries, these wool and linen hangings are highly stylized and schematic in their representations of figures and space, with all forms being outlined. The "Tapestry of the Angels," showing scenes from the life of Abraham and St. Michael the Archangel, and the "Tapestry of the Apostles," showing Christ surrounded by his 12 disciples, were both intended to be hung over the cathedral's choir stalls and therefore are long and narrow. The third hanging called the "Tapestry of Charlemagne Among the Four Philosophers of Antiquity" is a vertical wall hanging related to works produced by the convent at Quedlinburg in the German Rhineland during the Romanesque period of the 12th and early 13th centuries.

Romanesque tapestry

Fragments of a tapestry with traces of human figures and trees reminiscent of hangings described in the Norse sagas were found in an early-9th-century burial ship excavated at Oseborg in Norway. One of the major works of Romanesque weaving is a more complete tapestry dating from around the end of the 12th or early 13th century that was made for the Norwegian church of Baldishol in the district of Hedemark. Originally a set of wool hangings on the 12 months of the year, only the panels of April and May have survived (Kunstindustrimuseet, Oslo). The pronounced stylization of the images relates these tapestries to those executed for Halberstadt Cathedral.



"April and May," fragment known as the "Baldishol Tapestry," after 1190. In the Kunstindustrimuseet, Oslo. 1.18 × 2.00 m.
By courtesy of the Kunstindustrimuseet, Oslo

14th century. In the 14th century the western European tradition of tapestry became firmly established. At that time the most sophisticated centres of production were in Paris and Flanders. Large numbers of tapestries are recorded in inventories. The more luxurious standards of living being adopted by the wealthy of the Gothic period extended the use of tapestries beyond the customary wall hangings to covers for furniture. Survivals of 14th-century workmanship, however, are rare, and the most important of these were produced by Parisian weavers. The outstanding example of their art is the famous Angers Apocalypse (Musée des Tapisseries, Angers, France), which was begun in 1377 for the Duke of Anjou by Nicolas Bataille (flourished *c.* 1363–1400). This monumental set originally included seven tapestries, each measuring approximately 16.5 feet in height by 80 feet in length (5.03 by 24.38 metres). Based on cartoons drawn by Jean de Bandol of Bruges (flourished 1368–81), the official painter to Charles V, king of France, only 67 of the original 105 scenes have survived. A slightly later series (*c.* 1385) possibly woven in the same Parisian workshop is the "Nine Heroes" (Metropolitan Museum of Art, The Cloisters, New York). This set is not a religious narrative but illustrates the chivalric text *Histoire des Neuf Preux* ("Story of the Nine Heroes") by the early-14th-century wandering minstrel, or jongleur, Jacques de Longuyon.

Paris tapestry

Flanders, particularly the city of Arras, was the other great centre of the tapestry industry in 14th-century Europe. The tapestry produced there had such an international reputation that terms for tapestry in Italian (*arrazzo*) and Spanish (*drap de raz*) and English (arras) were derived from the name of this Flemish city. Long a medieval centre of textile weaving, Arras became an important tapestry centre when the leading citizens decided to create a luxury industry to alleviate the economic crisis caused by a decline in the sale of Arras textiles due to the popularity of cloth from the Flemish region of Brabant.

15th century. The greatest tapestries of the 15th century were produced in the Flemish cities of Arras, Tournai, and Brussels. In the first half of the century it was Arras that particularly prospered under the patronage of the dukes of Burgundy. Duke Philip the Good (1396–1467) had a specially designed building erected in the city to allow for better conservation of his tapestry collection. Between 1423 and 1467 no fewer than 59 master tapestry weavers were working in Arras, but following the French siege of the city in 1477 under King Louis XI the industry declined. After approximately 1530 it was no longer active. While the importance of Arras waned, that of Tournai and eventually Brussels waxed—their tapestries be-

Arras tapestry

coming the most sought after in the late 15th century. Local identification marks did not become general until the 16th century, and continual intercourse between the various medieval centres of tapestry making, particularly Arras and Tournai, adds to the difficulty of determining where individual tapestries were made. Despite the prestige of Arras workmanship, it is ironic that only one set of tapestries dating from 1402 is inscribed with the actual name. Large fragments showing scenes from the lives of St. Piat and St. Eleutherius survive in the cathedral of Tournai, for which they were commissioned. The imagery of these tapestries, like that of most Gothic hangings, was closely related to the styles of painting current at the time. Other important examples of supposed Arras tapestries inspired by Franco-Flemish book miniatures or paintings on wood panels include the early-15th-century tapestry of "The Annunciation" (Metropolitan Museum of Art, New York), which was probably woven after a cartoon by Melchior Broederlam (active 1381–*c.* 1409), and the "Court Scenes" (Musée des Arts Décoratifs, Paris), related to the *Très Riches Heures* illuminated for the Duc de Berry by the brothers Limburg (active early 15th century).

Whether a tapestry is an Arras or not is usually determined by comparison with the "History of St. Piat and St. Eleuthère." One of the finest works so attributed is the early-14th-century fragment from the set in the Museo Civico at Padua, Italy, illustrating the *Geste of Jourdain de Blaye*, a medieval chivalric story adapted from the ancient Greco-Roman romance *Apollonius of Tyre*.

Tournai tapestry

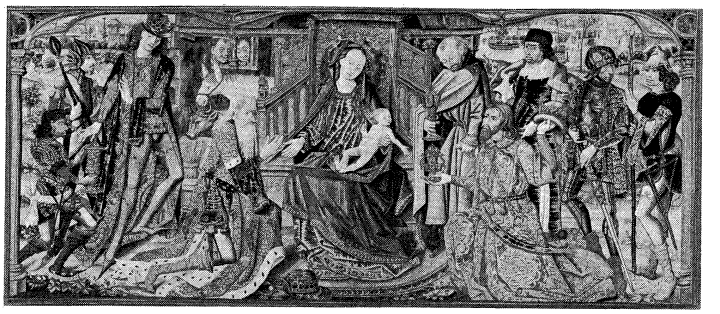
The craft, practiced since the end of the 13th century at Tournai, proved so prosperous that in 1398 a regulation concerning production was published. It is the oldest known ordinance regulating the craft of tapestry weaving. Among partially surviving tapestries ordered in the late 15th century by the Court of Burgundy were two sets produced by the weaver and tapestry merchant Pasquier Grenier (died 1493) for Philip the Good. One set, "The Story of Alexander" (Galleria Doria-Pamphili, Rome), was purchased in 1459, and the other, "The Knight of the Swan" (St. Katherine, Kraków, Poland, and Österreichisches Museum für Angewandte Kunst, Vienna), was bought in 1462.

Cited by many scholars as an example of mid-15th-century Tournai weaving under the influence of Arras are the four renowned tapestries of "The Hunts of the Dukes of Devonshire" (Victoria and Albert Museum, London). Typical of the developed late Gothic Tournai style are the compacted vertical compositions of "The Story of Strong King Clovis" (mid-15th century; Musée de l'Oeuvre Notre-Dame, Reims, France) and "The Story of Caesar" (*c.* 1465–70; Historisches Museum, Bern, Switzerland). Many of the attributed Tournai weavings are heavily outlined and have a solemnity that contrasts to the more fanciful nature of Arras weavings. A sense of monumentality is created by the immense size of many of these supposed Tournai weavings and by the way the vast surfaces are densely filled with superimposed imagery.

A producer of tapestry since the 14th century, in the 15th century Brussels vied with Arras and Tournai. By mid-century, Brussels was noted for its highly skilled reproductions of religious paintings by Flemish masters of late Gothic realism, such as in the tapestry of "The Adoration of the Magi" (1466–88; Cathedral of Sens, France). These panels were called "altarpiece tapestries" because they were usually intended for churches or private chapels, where they either were used as an altar cloth or antependium or were hung behind the altar as an altarpiece or fabric retable. In scale, altarpiece tapestries approximated the dimensions of the painting they copied and were, therefore, much smaller in size than the muralesque wall hangings of Arras and Tournai. Silk was commonly used to obtain the greater degree of naturalistic detail essential in reproducing a painting.

Tapis d'or and mille-fleurs tapestries

In the late 15th and early 16th centuries, Brussels also became famous for its production of *tapis d'or*, or "golden carpets," so called because of the profuse use of gold threads. Examples such as "The Glorification of Christ," popularly known as the "Mazarin Tapestry" (*c.* 1500; National Gallery of Art, Washington, D.C.), are characterized by their richness of effect.



"The Adoration of the Magi," Brussels altarpiece tapestry, 1466–88. In the Cathedral of Sens, France. 1.38 × 3.31 m.
Giraudon

Perhaps the best known late Gothic hangings were the fanciful tapestries usually referred to as mille-fleurs thousand flowers). A red or dark-blue ground strewn with flora and fauna sometimes serves as a setting for heraldic devices such as in the late-15th-century tapestry with the coat of arms of Philip the Good (Historisches Museum, Bern, Switzerland) or acts as a background for scenes of the chivalric aristocratic life during the late Middle Ages, such as in "The Hunt of the Unicorn" (Metropolitan Museum of Art, The Cloisters, New York) or "The Lady with the Unicorn" (Musée de Cluny, Paris). The origin of mille-fleurs tapestries is disputed, but it is thought that they were woven in the Flemish workshops of Brussels and Bruges and by itinerant weavers in the Loire Valley of France.

Itinerant Flemish and French weavers, setting up their looms in cities where there was temporary employment, carried tapestry weaving to Italy as early as the 15th century. Before the 16th century, however, most tapestries were bought in France and Flanders. Small workshops attached to the courts of various Italian nobles sporadically appeared for brief periods in Siena, Brescia, Todi, Perugia, Urbino, Mantua, Modena, Genoa, and Ferrara. The only one of importance was the Flemish-directed workshop of Ferrara, established around 1445 by the duke Lionello d'Este, who commissioned the famous Ferrarese early Renaissance painter Cosmè Tura (*c.* 1430–95) to make cartoons for his weavers.

Ferrara tapestry

16th century. Two new trends became apparent in the 16th century. The first, brought about by war and persecution in Flanders, resulted in the widespread diffusion of the Flemish art of tapestry weaving. Many Flemish artisans in the 16th century were forced to become refugees. Some grouped together to live the life of travelling craftsmen, while others attempted to re-establish their trade abroad. Flemish weavers were welcomed everywhere as carriers of a great tradition. Such itinerant masters established shops from England to Italy. The second important new trend emanated from Italy and reflected the superiority attached by the Italian Renaissance to the art of painting. The decisive step, which was to bring about the subordination of weaving to painting for more than 400 years in the art of tapestry, was taken when Pope Leo X commissioned the famed weaver Pieter van Aelst (flourished late 15th–early 16th centuries) of Brussels to make a series of tapestries illustrating the "Acts of the Apostles" from cartoons produced between 1514 and 1516 by Raphael (1483–1520). Little or no concession had been made to the tapestry medium for which the cartoons were intended, but the tapestries were a great success, and numerous copies of them were subsequently made.

The ascendancy of Brussels tapestry

The occupation of Arras by the French in the late 15th century and successive sieges of Tournai in the early 16th century contributed to the rise of Brussels as the leading tapestry centre of Flanders—a position it maintained until the 17th century. The patronage of the papacy and the imperial houses of Spain and Austria, along with other European royalty and the skill of its weavers, who were among the finest in Europe, combined to establish the international reputation of Brussels tapestry. The indus-

try was controlled by a monopoly of rich merchants. Tapestry making proved so prosperous in the period between 1510 and the outbreak of the Peasants War in 1568 that the industry had to be protected by regulations against frauds and forgeries. A number of communal ordinances followed one another in rapid succession, the most important being that of 1528, requiring each tapestry woven in Brussels to bear the mark of the city—a flat red shield flanked by two *B*'s standing for Brussels and the province of Brabant. The same imperial edict issued by Emperor Charles V also required manufacturers and merchants to use the signature or monogram of the master weaver or workshop.

It is the designs of the Flemish painter Bernard van Orley (1492?–1541) that are most characteristic of the Renaissance style of Brussels tapestry. Van Orley attempted to reconcile the traditions of late Gothic northern realism and the monumentality and idealism of Italian Renaissance art with the artistic potential of the tapestry medium. His earlier works, such as "The Legend of Our Lady of Le Sablon" (Musées Royaux d'Art et d'Histoire, Brussels) and "The Revelation of St. John" (1520–30; Patrimonio Nacional, Spain), still show compositional elements that link them to medieval Flemish art. Later, his work was influenced by the cartoons of Italian artists that were woven in Brussels workshops, such as Raphael's "Act of the Apostles" and the designs for "The Story of Scipio" and "Fructus Belli," executed by Raphael's disciple, the Mannerist painter and architect Giulio Romano (1499–1546). Van Orley adapted the Italians' preference for monumentality and their feeling for depth and sculptural modelling to Flemish tastes and traditions for genre and naturalistic detail in sets such as "The Battle of Pavia" (Museo e Gallerie Nazionali di Capodimonte, Naples), "The Story of Abraham" (Madrid and Vienna), "The Story of Tobias" (Vienna), and "The Hunts of the Emperor Maximilian I" (before 1528; Louvre, Paris). Among his followers in the first half of the 16th century were the Flemish painters Pieter Coecke van Aelst (1502–50), Jan Vermeyen (*c.* 1500–59), and Michel Coxcie (1499–1592). It was not only the cartoonists of Brussels who achieved international reputations but also the weavers of the early 15th century. Among the best known are Pieter van Aelst, Pieter and Willem Pannemaker, and Frans (active *c.* 1540–90) and Jacob Geubels (active *c.* 1580–1605).

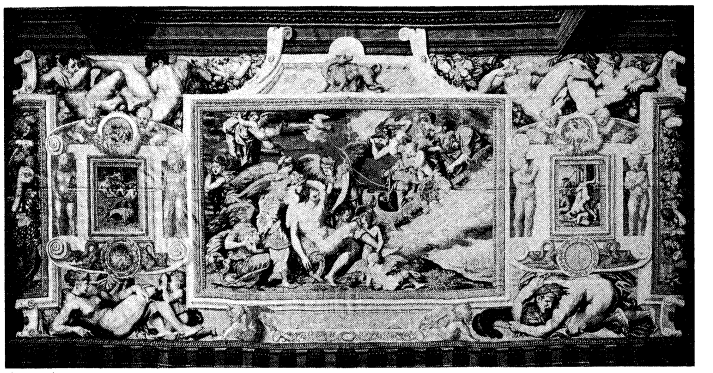
Other limited centres of tapestry making in 16th-century Flanders were Antwerp, Bruges, Enghien, Oudenaarde, Grammont, Alost, Lille, and Tournai. Perhaps the most distinctive type of tapestry produced in these cities was the *verdures* of Enghien and Oudenaarde. French tapestry weaving, after its eclipse in the 15th century when nomadic weavers seem to have been more active than established shops, owes much of its eventual prestige to an unusual degree of royal patronage. This resulted in the 17th century in the foundation of the Gobelins and Beauvais state factories, the names of which have now become household words. A prelude to this development was the factory established by Francis I in 1538 near Paris at the château of Fontainebleau to make tapestries for his royal residences. Staffed by Flemish weavers, the cartoons were largely furnished by two Italian Mannerist artists, Francesco Primaticcio (1504–70) and Rosso Fiorentino (1494–1540), who were court painters to the King.

French trompe-l'oeil tapestries

The six tapestries (Kunsthistorisches Museum, Vienna), based on their murals for the Galerie des Réformes in the château, are the first tapestries in which sculpture as well as painting is imitated in the highly illusionistic manner of a trompe-l'oeil (fool-the-eye) effect.

The Fontainebleau workshop, which was active for only 12 years, provided the springboard for subsequent developments in Paris, where in 1551 Henry II established and endowed with special privileges the Hôpital de la Trinité factory.

In the first third of the 16th century, Franco-Flemish weavers and small court workshops continued to supply the only indigenous Italian tapestry. Weaving was done in Genoa, Verona, Venice, Milan, and Mantua. The first internationally important Italian tapestry factory was established in 1536 in Ferrara by Duke Ercole II of the House of Este. The Arrazeria Medicea founded in 1546 in Florence by the Medici grand duke Cosimo I (1519–74) was the most important tapestry factory instituted in Italy during the 16th century and survived into the early 18th century. It was headed initially by the famous mid-15th-century Flemish weavers Nicolas Karcher and Jan van der Roost, both of whom had worked in the Ferrara workshop of Duke Ercole II.



"The Death of Adonis" (after a mural in the Galerie des Réformes, Fontainebleau), tapestry by Francesco Primaticcio, 1541–50. In the Kunsthistorisches Museum, Vienna. 3.30 × 6.40 m.
By courtesy of the Osterreichische Nationalbibliothek, Vienna

Cartoons were designed by such leading Mannerist artists of Florence as Jacopo Pontormo (1494–1556/57), Francesco Salviati (1510–63), Bronzino (1503–72), and Bachiacca (1494–1557), who designed the "Grotesques" (*c.* 1550; Uffizi, Florence), one of the most famous and influential tapestry sets produced by the Arrazeria Medicea.

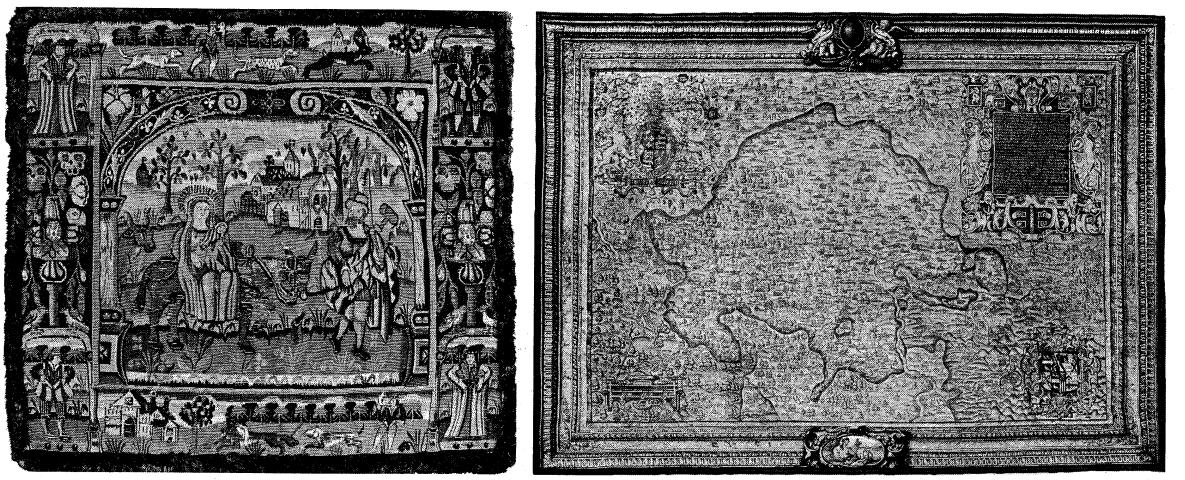
The major textile art in medieval England was embroidery. When woven tapestries were needed, they were imported from Flanders. Although occasional references to Arras weavers in England date from the 13th century and a few indigenous armorial tapestries have survived from the 15th century, it was only after the middle of the 16th century that the English organized tapestry works. The first important workshops were set up in Bercheston (Warwickshire) by a wealthy squire, William Sheldon (died 1570). They initially produced cushion covers and small hangings of heraldic and ornamental subjects. A later specialty of these shops was a series of topographic tapestries. Woven in 1588 from contemporary maps of the Midland counties, these tapestries featured bird's-eye views of hills, trees, and towns, surrounded, according to the custom of the period, by Flemish-styled borders of architectural and figural ornament. Many of the men who worked in these shops were Flemings who had fled the mid-16th-century religious persecutions in the Lowlands.

German court factories

Germany was one of the first regions to receive Flemish weavers fleeing religious persecution in the Lowlands. Their small workshops prospered in such cities as Cologne, Hamburg, Kassel, Leipzig, Torgau, Lüneburg, Frankenthal, and Stuttgart. Most of the works produced were in the Flemish style. In Switzerland, on the other hand, where tapestry making had flourished in the 14th and 15th centuries, the industry almost ceased to exist except around Basel and Lucerne.

17th and 18th centuries. It was due to the initiative of Henry IV, whose planning of his nation's economy emphasized the luxury production that has since been commercially important in France, that decisive steps were taken in establishing a French tapestry industry. In 1608 Henry gave official recognition to the French workshop (using the high-warp method) of Girard Laurent and Dubout by establishing them in the Louvre, and at the same time he encouraged the immigration of Flemish weavers practicing the low-warp method who would help Paris to compete with the flourishing industries of Brussels and Antwerp.

At the turn of the 16th–17th centuries, two Flemish weavers had been taken to France by government arrangement to establish low-warp looms in Paris: François de La Planche (or Franz van den Planken; 1573–1627)



English 16th- and 17th-century tapestries.
(Left) Tapestry cover with "The Flight into Egypt," *c.* 1600. In the Victoria and Albert Museum, London. 28.26 × 20.32 cm. (Right) Topographical tapestry depicting "Warwickshire," from late-16th-century designs by William Sheldon. In the Warwick Castle Museum, Warwick, England. 3.99 × 5.26 m.
By courtesy of (left) the Victoria and Albert Museum, London, (right) the County of Warwick Museum, England

and Marc de Comans (1563–before 1650). Satisfactory working conditions were found for them in the old Gobelins family dyeworks on the outskirts of the city, and so began the establishment commonly known by that name that has lasted ever since. One of its first ambitious productions was an allegorical invention lauding Catherine de Médicis under the guise of Artemisia. The cartoons for this set were chiefly by the French Mannerist painter Antoine Caron (*c.* 1515–93). The Baroque verve and vitality of the Flemish painter Peter Paul Rubens (1577–1640) and Simon Vouet (1590–1649) brought new life to French designs in the early 17th century.

De La Planche died in 1627 and was succeeded by his son, who broke with the Comans family and moved to the Faubourg Saint-Germain-des-Près, leaving the Comans at the Gobelins. Competition became bitter, but both continued to produce a considerable quantity, as well as good quality, until they were superseded in 1662 by the royal factory, which purchased the Gobelins works at its location.

Establishment of the Gobelins factory

The Gobelins was officially established in 1667, receiving the title Manufacture Royale des Meubles de la Couronne ("Royal Factory of Furnishings to the Crown"). Initially it included all the king's artisan corps (tapestry weavers, cabinetmakers, goldsmiths and silversmiths, etc.) that produced furnishings for the royal residences, especially the château of Versailles. Louis XIV's finance minister, Jean-Baptiste Colbert (1619–83), always alert to profitable opportunities, recruited skilled personnel not only from the de La Planche and Comans shops but also from the old Louvre enterprise and thus established a new tapestry works with both high- and low-warp looms. The Gobelins' first director was the painter Charles Le Brun (1619–90), who had managed the short-lived royal tapestry works established in 1658 by Colbert's predecessor, Nicolas Fouquet (1615–80), at his château of Vaux-le-Vicomte near Paris. Le Brun applied himself with prodigious energy to his new position and proved to have a special talent for the task of celebrating the glory of Louis XIV. Among the most important sets he designed were "The Elements," "The Seasons," "The Child Gardeners," "The Story of Alexander," and, above all, the "Life of Louis XIV" and the "Royal Residences" (most of these sets are in the possession of Mobilier National in Paris).

When Le Brun died, the painter Pierre Mignard (1612–95) became director. The draining of the royal treasury closed the Gobelins in 1694. The factory opened again in 1699, when a lighter spirit was introduced into tapestry design by the decorative inventions, especially grotesques, of Claude Audran III (1658–1734), who designed such sets as "The Grotesque Months" and "The Portières of the Gods."

Louis XV (1710–74), in his turn, was celebrated in a set of "Hunts" by the Rococo painter Jean-Baptiste Oudry (1686–1755). Oudry was director of the Gobelins from 1733 until his death in 1755, when he was succeeded by François Boucher (1703–70), the outstanding artist-director of the 18th century. Boucher and Charles-Antoine Coypel (1694–1752), a Rococo painter, designed many of the popular *alentours* tapestries, in which the central subject, presented as a painting bordered by a frame simulating gilded wood, is eclipsed by the rich use of ornamental devices surrounding it. Boucher's "Loves of the Gods" were also *alentours* and enjoyed a great success and popularity, especially among the English nobility. "The Story of Don Quixote" (Mobilier National, Paris) was designed by Coypel and woven nine times between 1714 and 1794.

Alentours tapestry

Oudry's sophistication and polished elegance posed new problems for the weavers. Now indeed it was necessary for them to learn to paint with a bobbin, and to this end hundreds of new dyes were perfected for both wool and silk, until about 10,000 hues were available, to effect almost imperceptible tonal modulations; and interlocking of the wefts was introduced to render the transitions practically invisible, while the finest textures practical were used.

The Gobelins succeeded in surviving the French Revolution. Napoleon as emperor, like Louis XIV, desired an art of apotheosis and ordered a set of tapestries (1809–15; Mobilier National, Paris) that were devoted to his reign. Paintings by such French Neoclassical painters as Jacques-Louis David (1748–1825), Carle Vernet (1758–1836), and Anne-Louis Girodet-Trioson (1767–1824) were woven into tapestries in the late 18th and early 19th centuries.

The Beauvais factory

Another major state-subsidized factory established in 1664 at Beauvais had been carried on by two Flemings, Louis Hinart for 20 years and Philippe Behagle for 27 more. It was administered in much the same way as the Gobelins. Beauvais, however, was a private enterprise with royal patronage intended to produce tapestries for the nobility and the rich bourgeoise, while Gobelins' work was only for the king.

Two types of decorative panels were particularly developed at Beauvais in the late 17th century, the architectural composition and the grotesque. The former, such as in the set of "Marine Triumphs" (1690; Banque de France, Paris), usually shows a complex fantasy architecture reminiscent of Baroque stage sets. In the latter, architectural tracery defines a complex of panels, framing a medley of festoons, scarves, vases, musical instruments, putti, masks, and comedy actors, such as in "The Rope Dancer and the Dromedary" (*c.* 1689; Mobilier National, Paris).

Both Oudry and Boucher designed for the Beauvais factory. The "Fables of La Fontaine," by Oudry, were among the most popular tapestries of the 18th century. In 1736 Boucher designed Italian genre scenes for the set "Village Festivities" and later in the "Second Chinese Set" did Chinese fantasies. He also designed various pastoral scenes with titillating overtones. The Beauvais factory became noted for tapestry to upholster furniture with and

By courtesy of the Mobilier National, Paris; photograph, Visages de France



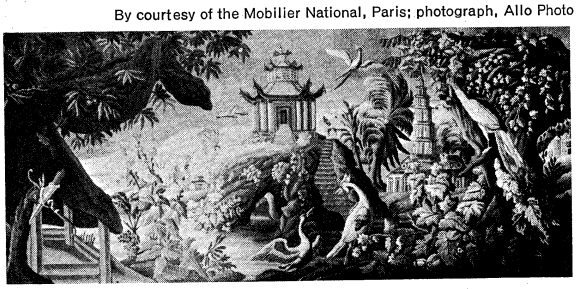
Fauteuil (armchair) covered with Beauvais silk tapestry, first quarter of the 19th century. In the Mobilier National, Paris.

panels for screens. These were usually floral designs and in the 19th century were especially fashionable in finely woven silk. By the end of the century, though technical standards were maintained, artistic deterioration set in.

The Aubusson–Felletin factory

Factories at the neighbouring old tapestry-making communities of Aubusson and Felletin, which had operated for a century and a half as modest private undertakings, were allowed to use the royal Aubusson mark as of 1665. From a small house industry, in which weavers independently produced inexpensive tapestries on their own low-warp looms for a bourgeois clientele, the tapestry makers soon produced hangings, upholstery fabrics, and carpets in Aubusson. The most effective tapestries are the chinoiseries, or genre fantasies set in China, a theme popular in Rococo art. Those designed by Jean Pillement

By courtesy of the Mobilier National, Paris; photograph, Allo Photo



18th-century French tapestry.
"The Pagoda," chinoiserie tapestry woven at Aubusson and designed by Jean Pillement, 18th century. In the Mobilier National, Paris. 2.85 × 5.84 m.

(1728–1808) are especially famous. Coarse and rather dull, the *verdures*, or "garden tapestries," which were the first Beauvais tapestries, were made in quantities. Aubusson architectural panels either imitate those of the Gobelins and Beauvais factories, often with more complex elements and the addition of animals, or depict a damasked wall hung with a painting or cluster of decorative objects and garlands. The factory was especially successful in its production of carpets with conventional geometric ornamental motifs or floral designs.

The dominant influence on the Brussels industry of the 17th century was the Antwerp painter Peter Paul Rubens, whose most famous set was the "Triumph of the Eucharist" (1627–28). Imitations and adaptations of his style were legion. Heavy and elaborate columns were often substituted for side borders. On a more modest scale are the tapestry versions of genre paintings by David Teniers the Younger (1610–90), in which the border frequently simulated the actual picture frame.

The first major tapestry factory to be established in Germany was founded in 1604 in Munich by Duke Maximilian of Bavaria. The designers and weavers were all Flemish. Although the factory closed after only 11 years of operation, the quality of its workmanship was outstanding. Following the loss of religious freedom in France when the Edict of Nantes was revoked in 1685, many French weavers, especially from the Aubusson factory, sought refuge from persecution in Germany as had the persecuted Flemish weavers of the 16th century. The workshop established in 1686 in Berlin by the great elector Frederick William of Brandenburg (1620–88) employed many of these displaced Aubusson weavers. It produced tapestries mainly for the palaces built by the Great Elector's son, King Frederick I of Prussia (1657–1713), after whose death the factory closed.

French designers and weavers continued to produce a large number of tapestries in the 18th century. Tapestry production was centred principally in Munich, Berlin, Würzburg, Dresden, Schwabach, and Erlangen.

Scandinavian folk tapestry

In Scandinavia tapestries for the Danish and Swedish royalty were woven in Copenhagen and Stockholm. The weavers and designers were usually French and Flemish. Norway and Sweden continued to produce folk tapestries. Of the nearly 1,300 registered Norwegian tapestries,

By courtesy of the Kunstindustrimuseet, Oslo



"The Feast of Herod," tapestry from Gudbrandsdal, Norway, 17th century. In the Kunstindustrimuseet, Oslo. 1.96 × 1.37 m.

approximately 1,250 originated in small rural communities. These tapestries were usually coarse in texture, stylized and schematic in design, and boldly coloured.

James I established in 1619 by royal charter a factory

of tapestry weaving at Mortlake near London. It was staffed by 50 Flemings. Philip de Maecht, a member of the famous late-16th- and 17th-century family of Dutch tapestry weavers, was brought from the de La Planche-Comans factory in Paris, where he had been the master weaver, to hold the same position at Mortlake. The royal factory flourished under the patronage of the Stuart monarchs James I and Charles I. Many of the early tapestries produced at Mortlake were modelled after hangings woven in Brussels. Rubens supplied cartoons and in 1623 suggested to Charles I the purchase of seven of the Raphael cartoons for the "Acts of the Apostles." A new set was woven from these cartoons at Mortlake and is preserved at the Mobilier National in Paris. The redesigned borders have been attributed to the renowned Flemish painter to the English court, Sir Anthony Van Dyck (1599–1641). Weathering the Puritan austerity of the Commonwealth period, the factory deteriorated under Charles II and closed in 1703.

From the late 17th century Francis Poyntz (died 1685) and his brothers had a studio in Soho, where a number of weavers originally employed in the royal factory produced a distinct style of tapestry based on Chinese and Indian lacquerwork.

Cardinal Francesco Barberini, the nephew of Pope Urban VIII, in 1633 established a tapestry factory in Rome. Even though it enjoyed papal patronage, it lasted only until 1679. Clement XI tried to establish another Roman tapestry works in 1710, which also failed. During the 18th century other small factories briefly existed in Turin and Naples. They were staffed mainly with weavers left unemployed by the closing of the Medici factory (Arrazeria Medicea) in Florence.

During the 15th and 16th centuries Franco-Flemish tapestries were imported in great quantities, and Flemish weavers were invited to Spain in order to repair and care for them. For a short time in the 17th century a factory, established by Philip IV (1605–65), operated at Pastrana near Madrid. It was not until Philip V (1683–1746) established the Real Fábrica de Tapices y Alfombras de Santa Barbara (Royal Factory of Tapestries and Rugs of St. Barbara) in 1720 at Madrid, however, that important tapestry was produced in Spain. Initially, the weavers and director were Flemings. The first tapestries made at Santa Barbara were woven from the cartoons of such Flemish Baroque painters as David Teniers the Younger (1610–90) and Philips Wouwerman (1619–68) or based on famous paintings by such Italian artists as Raphael (1483–1520) and Guido Reni (1575–1642). When the early Neoclassical painter Anton Raphael Mengs (1728–79) became director, the factory entered its most brilliant period of production. The Spanish painter Francisco Bayeu (1734–95) and his painter son-in-law Francisco Goya (1746–1828) were commissioned to make cartoons. From 1777 to 1790 Goya made 43 cartoons for the "Los Tapices" ("The Tapestries") series depicting Spanish daily life. The painted models for this are among the finest works of Goya's Rococo style.

The French destroyed the factory in 1808, but after the Napoleonic occupation, production was resumed until 1835. The tapestries produced during this period were largely copies of works woven in the 18th century.

A tapestry factory staffed by weavers from the Gobelins was established at St. Petersburg (now Leningrad) in 1716 by Tsar Peter the Great (1672–1725). Although tapestries were produced until 1859, production was often plagued with difficulties. The most striking designs were a set of grotesques (1733–1738) and a series of portraits, of which the ones of Catherine the Great (1729–96) are the most noteworthy.

19th century and 20th century. Most 19th-century tapestries reproduced paintings or previously woven designs. The influence of the Industrial Revolution was inescapable, of course, not only in tools, materials, and dyes but in the new middle class market and its demands. Machine-made tapestry, although an achievement in mechanical weaving, became a threat to the survival of the original handcraft.

The necessity for the revitalization and purification of the tapestry art was first recognized by the artists associated with the Arts and Crafts Movement in late-19th-century England. Decrying the loss of individual creativity, they revived the ideals of medieval craftsmanship in an attempt to counter the effects of industrialization on the decorative or applied arts. The leader and most important figure of the movement was the artist William Morris (1834–96), who established a tapestry factory at Merton Abbey in Surrey near London. For about 15 years he and his associates had been designing not only for looms but also for pictorial wall decorations and stained-glass windows. They were well prepared professionally, therefore, to design tapestries. Morris and the painter-illustrator Walter Crane (1845–1915) contributed cartoon sketches, but most Merton tapestries were designed by the Pre-Raphaelite painter Sir Edward Burne-Jones (1833–98).

Influence of the Arts and Crafts Movement

By courtesy of (left) the Victoria and Albert Museum, London, (right) the Museum fur Kunst und Gewerbe, Hamburg



Late-19th-century tapestries.
(Left) "Angeli Laudantes," tapestry designed by Edward Burne-Jones (workshop of William Morris, England), 1894. In the Victoria and Albert Museum, London. 2.25 × 2.00 m. (Right) "Swans," tapestry after Otto Eckmann (Scherrebek workshop, Germany), 1897. In the Museum für Kunst und Gewerbe, Hamburg. 2.35 × 7.65 m.

More venturesome than any of the Merton Abbey products were the tapestry designs made in the 1880s by the artist and architect Arthur Heygate Mackmurdo (1851–1942), who in 1882 founded the Century Guild, the first of many groups of artists-craftsmen-designers to follow the teachings of William Morris. This tradition, influenced by the tapestry revival in mid-20th-century France, has continued in Scotland. The most ambitious 20th-century tapestry designed by a British artist, Graham Sutherland's (1903–) enormous "Christ of the Apocalypse" (1962) for Coventry Cathedral, was, however, woven on Aubusson looms in France. This is the largest tapestry ever to have been made there (23.80 by 11.60 metres; 78 feet 1 inch by 38 feet 1 inch).

In Europe during the late 19th century there was a resurgence of tapestry based on folk traditions. This trend was already apparent in Norway shortly after 1890, when special efforts were made to base a modern tapestry art on native medieval weavings. The leaders were Gerhard Munthe (1849–1929), a well-known painter, and Frida Hansen (1855–1931), a weaver who studied the peasant craftsmanship of Norway and evolved an individual, light, and open weave. More recent developments in Scandinavia have been made in Sweden and Finland. Märta Måås-Fjetterström (1873–1941) is the best known Swedish tapestry artist, and her atelier continued to produce excellent works. In Finland a freer, more colourful art, more delicately scaled, has been practiced by many; among the best known are Martta Taipale, Laila Karttunen, and, for damask tapestry, Dora Jung (1906–).

Folk traditions in modern European tapestry

In Norway and to a lesser degree in Denmark similar work has been done. The church in the Scandinavian countries has been unusually receptive to this art. Traditional folk weaving has also been behind the revival of tapestry making in Czechoslovakia, Hungary, and Yugoslavia since World War I. Poland has produced especially original designs executed in a remarkably free technique. Following the tradition of heavy-grained native weaving, mid-20th-century Polish designer-weavers such as Magdalena Abakanowicz (1930–) and Wojciech Sadley (1932–) used unconventional materials such as jute, sisal, horsehair, raffia, etc., in abstract tapestries that emphasize the nature of the material, tactile stimulation, plasticity, or surface relief.

Germany, emulating Scandinavia, also began a revival of tapestry weaving around the turn of the century. In the state of Schleswig-Holstein a small tapestry industry was set up from 1896 to 1903 at Scherrebek, followed by similar enterprises at nearby Kiel and Meldorf. The most significant development, however, occurred at the design school of the Bauhaus, where tapestry was created during the 1920s and early 1930s. Abstract in composition, the Bauhaus designs were deeply rooted in the theory that the technology of the craft should be revealed in the work and in expressing the nature of the materials used, especially by the exploitation of heavy fibres as strong textural elements. Anni Albers, wife of the painter and Bauhaus instructor Josef Albers (1888–), became the chief practitioner of this kind of tapestry. Like most modern tapestry weavers, she also designed for the textile industry. After World War II, tapestry works were established in Munich and Nürnberg, and individual weavers worked throughout Germany and in Vienna. Among the Germans, unlike the French, stained glass, rather than tapestry, generated greater enthusiasm as a revived craft in the post-World War II period. A few individual designers worked on their own looms in the United States and Canada, where most large-scale tapestries continued to be imported from Europe. The Latin-American revival of indigenous folkcrafts aroused interest in tapestry making in Mexico and Panama. South American centres of tapestry art developed in Brazil, Chile, and Colombia.

Bauhaus tapestry

Modern tapestry design was hindered during the greater part of the 19th century in France by the academic administration of the state factories, although progressive artists began to be affected by the English Arts and Crafts movement in the late 1880s. The painters Paul Gauguin (1848–1903) and Émile Bernard (1868–1941) were among those who took an interest in tapestry weaving, though they did not actually do tapestry cartoons as did Aristide Maillol (1861–1944). It was not until after World War I that France initiated and led the 20th-century revitalization of tapestry as an art. Many of the great modern artists of the school of Paris—Pablo Picasso 1881–1973), Georges Braque (1882–1962), Henri Matisse (1869–1954), Fernand Léger (1881–1955), Georges Rouault (1871–1958), and Joan Miró (1893–), among others—permitted their works to be reproduced in 1932. It was done with extraordinary fidelity under the supervision of Marie Cuttoli, a Paris connoisseur and promoter of exceptional taste. The Aubusson factory, chosen for this important weaving, became once again a great centre for tapestry. The direct translation of painting into tapestry, however, left little scope for the weaver, and it is the trend begun simultaneously by Jean Lurçat (1892–1966) that may be said to have truly inaugurated the 20th-century tapestry renaissance. Although he began experimenting in 1916, Lurçat's art did not become definitive until the 1930s, when under the impact of Gothic tapestry, particularly the 14th-century "Angers Apocalypse," and in collaboration with François Tabard (1818–69), master weaver at Aubusson, he formulated the principles that were to make tapestry once again a joint creation between artist and weaver—an art in its own right. No longer merely an imitation painting, tapestry once again exploited the coarser texture and the bolder but more limited range of colours that characterized medieval hangings.

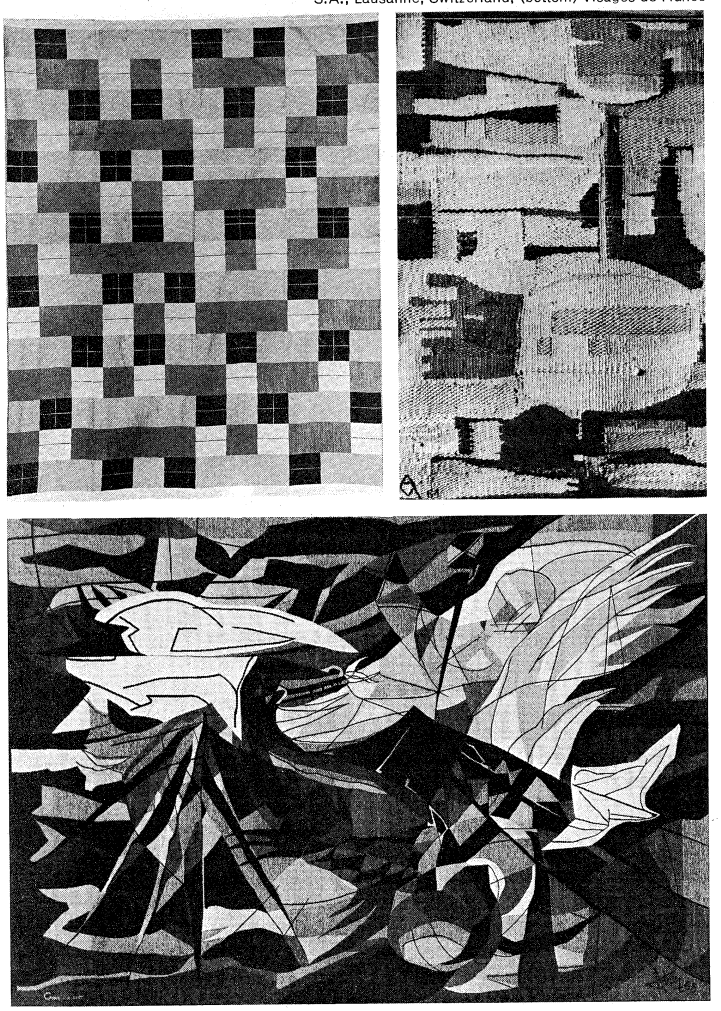
French revitalization of the tapestry art

In 1947 Lurçat founded the important Association des Peintures-Cartonniers de Tapisserie (Association of Cartoon Painters of Tapestry). Also active in this organization were the important French tapestry designers Marc Saint-Saëns (1903–) and Jean Picart Le Doux (1902–), who were Lurçat's foremost disciples. Lurçat was held in great esteem by Dom Robert (1907–), a Benedictine monk whose tapestries of poetic fantasy were largely inspired by Persian and medieval European manuscript illumination. Other major French designers of representational compositions were the artists Marcel Gromaire (1892–1971) and Henri Matisse and the architect Le Corbusier (1887–1965).

In the 1950s tapestry designs became increasingly abstract. Among the most notable pieces were those designed by the sculptor and printmaker Henri-Georges Adam (1904–1967). Using only black and white, his tapestries are monumental tonal abstractions that reflect his work as an engraver. The sculptor Jean Arp (1887–1966) and the painter Victor Vasarely (1908–) are other abstract designers of postwar tapestries.

After World War II the Belgians, influenced by the weaving activity in France during the 1930s, revived their tapestry industry. In 1945 the Forces Murales movement was organized in Tournai by a group of cartoon painters including Louis Deltour (1927–), Edmond Dubrunfaut, and Roger Somville (1923–), who became the leading designers of Belgian tapestries. This was followed in 1947 by the organization in Tournai of a collective tapestry workshop, Centre de Rénovation de la Tapisserie,

By courtesy of (top left) Anni Albers, (bottom) the Mobilier National, Paris; photograph, (top right) Edita S.A., Lausanne, Switzerland, (bottom) Visages de France



20th-century tapestries.
(Top left) Tapestry by Anni Albers, 1927. In the collection of the artist. 1.19 × 1.52 m. (Top right) "White Composition," tapestry after Magdalena Abakanowicz, 1962. In the collection of the artist. 1.05 × 1.80 m. (Bottom) "Mont Saint-Michel," after Henri-Georges Adam, 1965. In the Mobilier National, Paris. 4.02 × 5.65 m.

active until 1951. Small workshops continued to flourish in Belgium, especially in the cities of Tournai, Brussels, and Malines.

The renewed international interest in tapestry is clearly related to the austerity of modern architecture. Suitable settings for large-scale wall hangings are provided by the often vast expanses of bare wall surface in contemporary buildings. Le Corbusier not only used tapestries to decorate his architectural interiors but designed them. He frequently referred to tapestries as nomadic murals, recognizing their importance as movable and interchangeable decoration.

In 1962 the first international exhibition of tapestry was held at Lausanne in Switzerland, which after 1965 became an important biennial event. This exhibition clearly demonstrates the tremendous worldwide interest in the medium generated in the middle 20th century as well as indicating the immense variety of tapestry designs, materials, and techniques.

BIBLIOGRAPHY. M.J. GUIFFREY, E. MUNTZ, and A. PINCHART, *Histoire générale de la tapisserie*, 3 vol. (1880), French tapestries discussed by Guiffrey, Italian tapestries by Muntz, Flemish tapestries by Pinchart; J. GUIFFREY, *Histoire de la tapisserie*, (1878–85), one of the first books of synthesis on the history of tapestry from the Middle Ages to the mid-19th century; M. FENAILLE, *État général des tapisseries de la manufacture des Gobelins depuis son origine jusqu'à nos jours, 1600–1900*, 6 vol. (1903–23), a work of primary importance, presenting a detailed history of the Gobelins factory; J. BADIN, *La Manufacture de tapisseries de Beauvais depuis ses origines jusqu'à nos jours* (1909), a basic reference for the history of tapestry production at the Beauvais factory; MME SCHNEEBALG PEVELMAN, "Les Sources de l'histoire de la tapisserie bruxelloise et la tapisserie en tant que source," in *Annales de la Société royal d'Archéologi de Bruxelles* (1966), a refutation of Guiffrey's theory, that attributed mille-fleurs tapestries to the itinerant workshops of the Loire Valley; H. GOBEL, *Wandteppiche*, 6 vol. (1923–34; Eng. trans. of pt. 1, *Tapestries of the Lowlands*, 1924), a general worldwide treatment of tapestry, with numerous black and white illustrations, although many of the European medieval attributions have been questioned or rejected; G.L. HUNTER, *The Practical Book of Tapestries* (1925), precise and useful descriptions, with numerous reproductions; C.G. JANNEAU, *Évolution de la tapisserie* (1947), illustrations and technical information on collections of European tapestries, some of which have been subsequently disbanded with works relocated; J. CASSOU, M. DAMAIN, and R. MOUTARD-ULDRY, *La Tapisserie française et les peintres cartonniers* (1957), a well-illustrated book discussing the 20th-century revival of the French tapestry industry; D. HEINZ, *Europäische Wandteppiche*, vol. 1, *Von den Anfängen der Bildwirkerei bis zum Ende des 16. Jahrhunderts* (1963), a thorough treatment of tapestry up to the end of the 16th century, with a typological index, an extensive bibliography, and numerous illustrations; R.A. WEIGERT, *La Tapisserie et le tapis en France* (1964), a scholarly discussion of the history of French tapestry; P. VERLET *et al.*, *Le Grand livre de la tapisserie* (1965), a well-illustrated volume on Western tapestry from the Middle Ages to the 20th century; M. JARRY, *La tapisserie des origines à nos jours* (1968; Eng. trans., *World Tapestry*, 1969), a well-documented study of tapestry throughout the world, including an extensive bibliography and many black and white and colour illustrations; V. FOUGRE, *Tapisseries de notre temps* (1969), a short study of contemporary French tapestry, with an index of the tapestry artists and black and white and colour reproductions; R.A. D'HULST, *Flemish Tapestries* (1967), an elaborate study of Flemish tapestry from the Middle Ages to the Baroque periods, with many colour illustrations; R. KAUFMANN, *New American Tapestry* (1968), a well-illustrated text dealing with technique as well as with the works of leading American tapestry designers and weavers; G. SUTHERLAND, *Coventry Tapestry* (1964), an interesting account of the design, weaving, and installation of Sutherland's tapestry for Coventry cathedral; P. ACKERMAN, *Tapestry: The Mirror of Civilization* (1933, reprinted 1970), one of the classic works in English dealing with the historical development of European tapestry.

Collection catalogs: Important illustrated catalogs of tapestry collections include: A.S. CAVALLO, *Tapestries of Europe and Colonial Peru in the Museum of Fine Arts, Boston*, 2 vol. (1967); D. DU BON, *Tapestries from the Samuel H. Kress Collection at the Philadelphia Museum of Art* (1964); A.J. WACE and H.C. MARILLIER, *Marlborough Tapestries at Blenheim Palace* (1968); N.Y. BIRYUKOVA, *The Leningrad Hermitage Gothic and Renaissance Tapestries* (1966); M. CRICK KUNTZIGER, *Musées royaux d'Art et d'Histoire: Catalogues des tapisseries* (1956); A.M. ERKELENS, *Wandtapijten I. Late Goteik en vroege Renaissance, Rijksmuseum, Amsterdam* (1962); A.F. KENDRICK, *Catalogue of Tapestries at the Victoria and Albert Museum* (1924); and H.C. MARILLIER, *The Tapestries at Hampton Court Palace, London*, 2nd ed. (1962).

(M.J.)

Tariffs

A tariff, or duty, is a tax levied on a commodity when it crosses the boundary of a customs area. The boundary may be that of a nation or a group of nations that have agreed to impose a common tax on goods entering their territory. Tariffs are often classified as either protective or revenue. Protective tariffs are designed to shield domestic production from foreign competition by raising the price of the imported commodity. Revenue tariffs are designed to obtain revenue rather than to restrict imports. The two sets of objectives are, of course, not mutually exclusive. Protective tariffs, unless they are so high as to keep out imports, yield revenue. Revenue tariffs give some protection to any domestic producer of the duty-bearing goods. A transit duty, or transit tax, is a tax levied on commodities passing through a customs area en route to another country. Similarly, an export duty, or export tax, is a tax imposed on commodities leaving a customs area. Finally, some countries provide export subsidies; import subsidies are rarely used.

Tariffs on imports may be applied in several ways. If they are imposed according to the physical quantity of an import (so much per ton, per yard, per item; etc.), they are called specific tariffs. If they are levied according to the value of the import, they are known as *ad valorem* tariffs.

Tariffs may differentiate among the countries from which the imports are obtained. They may, for instance, be lower between countries that have previously entered into special arrangements, such as the trade preferences accorded to each other by members of the British Commonwealth.

Tariffs may affect the economy of the country imposing them in a number of ways. By raising the prices of imported goods, tariffs may encourage domestic producers to increase production. As expenditures on domestic products rise, domestic employment tends to do likewise. This is why tariffs are favoured by industries that find themselves pressed by foreign competitors. The tariff may also encourage tendencies toward a monopolistic market structure to the extent it lessens foreign competition, with a resulting decrease in the incentive to modernize or innovate. But a tariff by increasing the price of a commodity may also reduce its consumption, depending on how consumers react to the higher price. If enough of them curtail purchases, the decrease in demand may be large enough in relation to the world market to force the price of the import down. This would register as an improvement in the tariff-imposing country's terms of trade. In other words, if export prices remained stable, the country would be paid the same amount for its exports but pay less for its imports.

Nontariff barriers. Other government regulations and practices may also act as barriers to trade. In some countries sales and turnover taxes are imposed on imports but are rebated on goods destined for export; if such taxes apply at identical rates to domestically produced commodities, foreign goods receive equal treatment. By contrast, excise taxes discriminate against foreign goods if they are levied on commodities that cannot be produced domestically.

Quotas or quantitative restrictions may prohibit the importation of certain commodities or limit the amounts imported. Such quotas are usually administered by requiring importers to have licenses to bring in particular commodities. Quotas raise prices just as tariffs do, but, being set in physical terms, their impact on imports is direct, with an absolute ceiling set on supply. Increased prices will not bring more goods in. There is also a difference between tariffs and quotas in their effect on revenues. With tariffs, the government receives the revenue: under

quotas, the import license holders obtain the revenue in the form of the difference between the high domestic price and the low international price of the import. Quotas are used to regulate oil imports in the United States; in other developed countries they serve to protect agriculture; and in a number of developing countries they are used in place of or in conjunction with tariffs, to prevent the outflow of money.

Still other barriers include state trading organizations and government procurement practices that may be used preferentially. In the United States, "buy American" legislation requires government procurement agencies to favour domestic goods. Finally, customs classification and valuation procedures, health regulations, and marking requirements may also have a restrictive effect on trade.

Measuring the effects of tariffs. It is difficult to gauge the effect of tariff barriers among countries. First of all, how import demand responds to changes in tariffs depends on a variety of factors—the reaction of producers and consumers to price changes, the share of imports in domestic production and consumption, the substitutability of imports for domestic products, and so on. And the responsiveness to tariff levels will differ from country to country as well as from commodity to commodity. Thus, the amount of a tariff does not necessarily determine its restrictive effect. For another thing, such comparisons also must be restricted to commodities for which tariffs are the major protective device. This is generally true for nonagricultural commodities in developed countries, but other devices are often employed to protect agricultural production. In the third place, a tariff levied upon imports of raw materials will protect domestic raw material producers but will increase the costs of manufacturers using those raw materials. It is necessary, therefore, to distinguish between nominal and effective rates of protection.

Figuring the effective rate of protection

One way of measuring effective protection is to take the value added in domestic production (*i.e.*, the value of output minus the costs of materials) after the imposition of tariffs and other protective measures and compare it with the value added before protection. To illustrate, consider a product (clothing) with an import price of 100, of which 60 represents the costs of material inputs (textile fabrics). The value added in manufacture is therefore 40. A 20 percent tariff on clothing would raise the domestic price by 20 to 120, while a 10 percent tariff on textile fabrics would increase material costs to the domestic producer by 6 to 66. Protection would thus enable the firm to operate with labour costs within a value-added margin of 54—the difference between the domestic price of 120 and the material cost of 66. The difference between the value added of 40 without tariff protection and that of 54 with it provides a margin of 14. That means that the effective rate of protection of the domestic processing activity—the ratio of 14 to 40—would be 35 percent.

The effective rate of protection also depends on the share of value added in the product price. Effective rates can be very high if value added to the imported commodity is a small proportion or very low if value added is a large proportion of the total price. Thus, effective protection in one country may be much higher than that in another even though its nominal tariffs are lower, if it tends to import commodities of a high level of fabrication with correspondingly low ratios of value added to product price.

The accompanying table shows the extent to which effective tariff levels differ from the nominal ones for principal industrial countries. The data pertain to manufactured products imported by the United States, the European Common Market, the United Kingdom, Sweden, and Japan—countries that between them account for four-fifths of world exports and about one-half of world imports of manufactured goods. The figures are estimated for the situation expected to obtain following the so-called Kennedy Round of tariff reductions negotiated in the 1960s. It is evident that effective rates of protection are generally much higher than nominal rates. This is explainable in part by the practice of setting higher tariffs on finished products in order to encourage the domestic processing of unfinished commodities, and in part by the absence of tariffs on nonmaterial inputs (services) that do not enter international trade. Among the five countries and country groups the overall average of nominal as well as that of effective tariffs is the highest in Japan, with the United Kingdom a close second, and Sweden at the opposite end of the scale. The United States and the European Economic Community occupy the middle ground.

Averages of Nominal and Effective Tariffs on Manufactured Products After the Kennedy Round

	percentages	
	nominal	effective
United States	6.8	11.6
United Kingdom	9.1	16.0
European Economic Community	6.6	11.1
Sweden	3.8	6.7
Japan	9.4	16.4

Source: B. Balassa in *The Kennedy Round: Estimated Effects of Tariff Barriers*, UNCTAD, 1968.

Efforts to liberalize trade. A great tide of protectionism rose during the depression of the 1930s, when industrial nations raised tariffs and other barriers to trade in an attempt to help domestic producers. Since most countries adopted these "beggar-my-neighbor" policies, the effect was to stifle trade. The gains to protected industries were offset by the harm to export industries. As the failure of these policies became apparent, proposals to stem the tide of protectionism and to reduce tariffs through multilateral agreement were put forward. Political and economic uncertainties prevented the realization of these goals in the years preceding World War II.

GATT, EEC, and EFTA

The post-World War II period witnessed several efforts to reduce tariff barriers. Prominent among these were the tariff reductions undertaken in the framework of the General Agreement on Tariffs and Trade (GATT) after that organization came into being in 1948. GATT members include virtually all major trading nations of the world, but only the more developed countries have participated in the successive rounds of tariff reductions. But such reductions are automatically applied to imports from developing member countries; under the most-favoured-nation clause, the contracting parties to GATT assume the obligation to extend to one another the tariff concessions accorded to any one member country.

Another central feature of GATT is the principle of reciprocity: mutual tariff concessions are designed to lead to equivalent increases in the exports and imports of each participant. The principles of nondiscrimination and reciprocity often conflict, since the agreed-upon tariff reductions are extended to countries not obliged to reciprocate. To lessen this effect of giving a "free ride" to third countries, much tariff bargaining was carried out on an item-by-item basis in the postwar period. This approach limited the scope of tariff reductions and, after initial successes, the progress of multilateral trade liberalization slowed down.

More comprehensive liberalizations occurred in the European Economic Community (EEC) and the European Free Trade Association (EFTA). The establishment of EEC and EFTA was made possible by Article 24 of the General Agreement on Tariffs and Trade, which provides for an exception to the principle of nondiscrimination; countries that mutually agree to abolish tariffs within a free-trade area or a customs union are not required to extend such concessions to other parties. The success of EEC and EFTA in reducing trade barriers was partly at the expense of outsiders, who were put in an unfavourable position in relation to member-country producers.

The Kennedy Round

This "trade diverting" effect of the EEC and EFTA, plus a desire to surmount the difficulties encountered with item-by-item tariff bargaining, underlay the U.S. initiative for trade liberalization under the Trade Expansion Act of 1962. Under that act, the Congress authorized the president to negotiate multilateral tariff reductions up to 50 percent on an across-the-board basis. The act brought about some important modifications in the traditional

method of tariff bargaining. The replacement of bilateral negotiations by across-the-board reductions in tariffs, with exceptions granted through an escape clause in the case of especially sensitive industries, made it possible to proceed with trade liberalization on a broad front. Firms injured by imports were to be given special assistance. The Trade Expansion Act led to the Kennedy Round of tariff reductions in 1967. In most cases tariffs were decreased by 50 percent, with smaller reductions for sensitive industries. On the average, tariff reductions on nonagricultural commodities amounted to about 35 percent. The Kennedy Round represented an important step toward international trade liberalization, but there was no indication that it would be followed by negotiations of a similar scope in years to come.

Protectionism in the less developed countries. The industrialization that has taken place in the postwar years in the developing countries has been characterized by the expansion of import-competing industries behind high tariff walls. In many of those countries tariffs and various quantitative restrictions on manufactured goods lie in the 100–200 percent range. The effective rates of protection are even greater, because the goods tend to be highly fabricated and the proportion of value added in production after importation is low. While some countries such as Hong Kong, Taiwan, and Korea have oriented their manufacturing industries mainly toward export trade, those are exceptional cases in the developing world. More commonly the new industries seek to compete with foreign-made goods for the domestic market. High protection in these countries has often contributed to a slowdown in the production and export of primary commodities and has discouraged expansion of exports of manufactured goods. Furthermore, while domestic production of nondurable consumer goods has permitted rapid economic growth at an early stage, developing countries have encountered considerable difficulties in producing more sophisticated commodities. They suffer the disadvantages of small domestic markets, along with a lack of incentives for technological improvement.

Arguments for and against tariffs. *The infant industry argument.* Advocates of protection often argue that new and growing industries, particularly in less developed countries, need to be shielded from foreign competition. They contend that costs decline with growth and that some industries must reach a minimum size before they are able to compete with well established industries abroad. Tariffs can protect the domestic market until the industry becomes internationally competitive and, it is often argued, the costs of protection can be recouped after the industry has reached maturity. This justification of tariffs is often abused. In many developing countries, industries have failed to attain international competitiveness even after 15 or 20 years of operation and might not survive if protective tariffs were removed. The infant industry is probably better aided by production subsidies than by tariffs. Production subsidies do not raise prices and therefore do not curtail domestic demand; and the cost of the protection is not concealed in higher prices to consumers. Production subsidies have the disadvantage of drawing upon government revenue rather than adding to it, which may be a serious consideration in countries at lower levels of development.

The employment argument. Tariffs raise import prices and tend to shift demand from imported to domestic products. Tariffs may therefore stimulate domestic employment. In developing countries, protection for the purpose of increasing employment may involve a misuse of scarce resources such as capital and entrepreneurship, but it is defensible if the employment objective is of paramount importance. In trade among developed countries, however, imposing tariffs for employment reasons means exporting unemployment to other countries, and it therefore invites retaliation.

Other arguments. One hypothesis in the theory of trade is that countries will tend to specialize in producing commodities that use relatively more of the factors of production that they have in abundant supply. For instance, a country with relatively more capital than labour will tend to specialize in goods that require more capital per unit of labour than other goods and to import labour-intensive goods. If, however, imports are restricted by tariffs, production will shift toward labour-intensive industries, increasing the demand for labour, and hence the wage rate.

Economists in developing countries have argued that, in order to attract labour, industry has to pay higher wages than agriculture and that tariffs are necessary to compensate the manufacturing sector for the higher cost of labour. The empirical evidence for such differential wage rates is not definitive, and therefore the worth of this argument must be evaluated case by case. Another frequently used argument for protection is based on the need to maintain so-called key industries that are vital for national defense. Domestic industries can be protected equally well, however, by subsidies. If the objective is to reduce dependence on foreign products, tariffs may be appropriate since they reduce both imports and the total consumption of the commodity in question. But the argument has lost much of its force in a world threatened by nuclear warfare.

Protection of key defense industries

As noted earlier, a country may be able to use tariffs to improve its terms of trade. But other countries may retaliate, and, furthermore, such power is not usually available to many countries over a long period. To the extent they restrict imports, tariffs also reduce the outflow of funds and thus may help remedy balance of payments deficits. This objective has been important in the less developed economies. A serious deficit in the balance of payments, however, calls for more fundamental treatment, such as devaluation of the currency.

In advanced industrial countries, tariffs result mainly from the political pressures of producers in particular industries. The costs of such protection are borne by consumers in general. The pressures of special interests on the parliaments and legislatures of industrialized countries are far more intense and sustained than the counterpressures from consumers, both in establishing tariffs and maintaining them once they exist.

BIBLIOGRAPHY. Elementary treatments of the theory and practice of tariffs may be found in standard textbooks on international economics, prominent among which are C.P. KINDLEBERGER, *International Economics*, 4th ed. (1968); and D.A. SNIDER, *Introduction to International Economics*, 4th ed. (1967). A more advanced treatment of the theory may be found in J. VANEK, *International Trade: Theory and Economic Policy* (1962).

Institutional and historical studies of tariffs are A. ISAACS, *International Trade: Tariff and Commercial Policies* (1948); and F.W. TAUSSIG, *The Tariff History of the United States*, 8th ed. (1967).

An analysis of the economic and political consequences of trade liberalization may be found in B.A. BALASSA, *Trade Liberalization Among Industrial Countries* (1967). Accounts of the General Agreement on Tariffs and Trade (GATT) may be found in G. CURZON, *Multilateral Commercial Diplomacy* (1965); and G. PATTERSON, *Discrimination in International Trade: The Policy Issues, 1945–65* (1966). On related issues, see also G. and V. CURZON, *After the Kennedy Round: What Trade Policies Now?* (1968).

Treatment of U.S. foreign economic policy may be found in P.B. KENEN, *Giant Among Nations* (1963); and W.B. KELLY (ed.), *Studies in United States Commercial Policy* (1963).

Studies of the effects of trade policy on the developing nations are H.G. JOHNSON, *Economic Policies Toward Less Developed Countries* (1967); and S. WEINTRAUB, *Trade Preferences for Less-Developed Countries: An Analysis of U.S. Policy.*

(B.B./T.J.B.)

Tarim River

The Tarim River (T'a-li-mu Ho) is the chief river of the Sinkiang Uighur Autonomous Region of China, an area situated immediately north of the Tibetan Plateau. It flows for most of its length through the Takla Makan Desert, although part of the watershed is located within mountainous territory of the Soviet Union.

Physical characteristics. The word *tarim* means a bank of a river that flows into a lake or that is undifferentiable from the sands of a desert. This is a characteristic hydrographic feature of many rivers that traverse the sands of

the Takla Makan Desert. Another characteristic of the rivers of the Tarim Basin, including the Tarim itself, is their active migration—*i.e.*, the shifting of their beds and banks.

The Tarim is formed by the confluence of the K'a-shih-ka-erh (Kyzylsu in the U.S.S.R.) and Yarkand rivers in the west, and by the A-k'o-su Ho and the Ho-t'ien Ho (Khotan) some 230 miles downstream. Only the A-k'o-su Ho flows for the entire year. It is the Tarim's most important tributary, supplying 70–80 percent of its water volume. The name Tarim is used after the Ho-t'ien Ho confluence. The total length of the Yarkand–Tarim river system is 1,260 miles (2,030 kilometres), although as the Tarim frequently changes its channel, the length tends to vary over the years. The area of the Tarim River Basin, including the basins of T'ai-t'e-ma Hu (T'ai-t'e-ma Lake), which is also known as Kara Buran Köl, and Lop Nor (Lop Lake), is 367,400 square miles (951,500 square kilometres). A considerable part of the Tarim is unformed, following no clearly defined riverbed. Here, as a result of extensive evaporation, the downstream volume of river water regularly decreases. This natural process of dispersion of the river's water is abetted by the diversion of water for irrigation. In the past, the waters of the Tarim merged with those of the Kunlun Mountains: the Ni-ya, the K'o-li-ya, and other smaller rivers. Presently, however, these rivers have been lost in the sands of the Takla Makan.

Irregularity of the Tarim's course

From the confluence of the Yarkand, Ho-t'ien, and A-k'o-su to the populated point of Sha-ya, the Tarim winds through a valley with cultivated bottomlands, the valley having widths to 3,300 feet and a clearly defined terraced relief with elevations of 13–20 feet. East of the settlement of Sha-ya, the river strays over the alluvial valley floor, generating many streams of varying size and nourishing numerous lakes and marshy areas. Below the village of Wei-li, the river valley narrows, and the Tarim approaches the K'ung-ch'iao Ho flowing toward the east and Lop Nor. In some years when the main volume of water in the Tarim is contributed by the K'ung-ch'iao, Lop Nor is filled; but if, as a result of the migration of the riverbed, the waters of the Tarim flow to the southeast along the ancient lake plain, then Lop Nor dries up, and the waters fill T'ai-t'e-ma Hu, located several hundred miles to the south of Lop Nor.

The Tarim's low-water period is from October through April. The spring and summer high waters begin in May, continuing through September.

The Lower Tarim Basin is a plain bordered by massive mountain ranges and composed of alluvium and lake sediments. A low, rocky ridge is situated in the western parts between the Yarkand and the Ho-t'ien rivers. The basin is the driest region of Eurasia and one of the most extraordinary regions in the whole world. The predominant part of the basin is occupied by the Takla Makan Desert, the sand area exceeding 105,000 square miles. In addition, there are several comparatively small sand massifs with areas of from 300 to 1,700 square miles. Sand dunes are the predominating desert relief. The large symmetrical ridges of friable sand, covered with sand dunes, have lengths up to 8 miles, widths of from 850 to 1,000 feet (sometimes up to 1,650 feet), and heights of from 100 to 500 feet. Outside the limits of the Takla Makan, there are wide stretches of salt flats, often having very thick salt crusts. The mineral content of the water increases from 200 milligrams per litre to 800 milligrams per litre from source to mouth.

Climate. Precipitation in the Tarim Basin is extremely scanty and, in some years, is nonexistent. In the Takla Makan Desert and in the Lop Nor plain, the average annual total of precipitation is 0.4–0.5 inch. In the foothills it increases to 1.6–2.4 inches. The western Kunlun mountains receive about 12 inches of precipitation a year, and in the remaining regions the total amounts to from two to four inches a year. In the Tien Shan (Tien Mountains) it is much wetter, precipitation often exceeding 20 inches. This decreases to from two to four inches a year in the ranges in the eastern end of the Tien Shan. The western part of the Tarim Basin receives westerly winds and the eastern part northeasterly winds. Maximum temperatures are around 104° F (40° C). The Tarim freezes over every year from December through March.

Vegetation and animal life. Vegetation in the Tarim Basin is mainly located along the river and its branches. Here, at the edge of the sands, shrublike vegetation and stunted trees, especially wormwood, is found. A thin forest of poplar is found in the Tarim Valley. Underbrush consists of willows, sea buckthorn, and dense growths of Indian hemp and Ural licorice. Poplar forests along the banks of the Tarim disappear downriver near the settlement of T'ieh-kan-li-k'o.

The Tarim is rich with fish, and animal life on the Tarim and the surrounding desert is varied. Such rare animals as the kulan (*Equus hemionus*), the wild two-humped camel (*Camelus bactrianus*), and the Siberian reindeer (*Cervus elaphus*) have survived. The wild boar (*Sus scrofa*), the jaran (a gazelle), and the hare (*Lepus tolai*) are frequently encountered. Predators include wolves and steppe foxes. Lynx and wildcat are occasionally found, and in the recent past tigers inhabited the region. The valley and lakes of the Tarim are used by migratory birds as a temporary stopover, those most frequently encountered including pintail duck, redbills, grey goose, cormorant, pochard, white tern, and long-legged heron.

Economic activity. The population of the area consists of various nationalities, basically Turkic. Uighurs and Uzbeks live along the Tarim and in the oases of the Tarim Basin, Tadzhiks in the western Kunlun Mountains and the Pamirs. Kirgizians and Mongols roam as far as the Nan Shan (Nan Mountains).

Irrigation works

The waters of the Tarim since ancient times have been used for irrigation. Such major oases as Kashgar, Khotan, Wei-li, and others are located in the basin. In recent years the area of irrigated land along the shores of the Ho-t'ien, Yarkand, A-k'o-su, and their main tributaries has rapidly increased. In addition to the irrigation system on the main rivers, dams have been created in the areas of the oases, protecting the surrounding land from flooding during the crest peaks of high water.

The lack of transport development has kept connections uncertain and movement both slow and costly. The Chinese have plans for marked development of several localities of the basin, but the comparatively small opportunities make them of relatively low priority, and, by the 1970s, little was achieved other than the political organization of the region and the settling of Chinese nuclei in most of the important oases. A few agricultural processing plants were established and resource exploration was begun.

Oasis agriculture remains the mainstay of the scattered settlements. Grains, cotton, silk, fruits, and wool are the chief agricultural products, and Khotan jades are the only other important item. Little is exported from the basin, but local surpluses find a market among travellers passing through when conditions permit. (N.T.K.)

Tasman, Abel Janszoon

Abel Janszoon Tasman, the greatest of Dutch navigators and explorers, was the discoverer of Tasmania, New Zealand, Tonga, and the Fiji Islands.

He was born at Lutjegast in Groningen in the Netherlands, probably in 1603. He entered the service of the Dutch East India company in 1632 or 1633 and made his first voyage of exploration to Ceram (modern Seram) Island (in modern Indonesia) as captain of the "Mocha" in 1634. He sailed in 1639 under Comdr. Mathijs Hendrickszoon Quast on an expedition in search of the "islands of gold and silver" in the seas east of Japan. After a series of trading voyages to Japan, Formosa, Cambodia, and Sumatra, he was chosen by the governor general of the Dutch East Indies, Anthony van Diemen, to command the most ambitious of all Dutch voyages for the exploration of the Southern Hemisphere.

By 1642, Dutch navigators had discovered discontinuous stretches of the western coast of Australia, but whether these coasts were continental and connected with

the hypothetical southern continent of the Pacific remained unknown. Tasman was to answer this problem, following instructions based on a memoir by Frans Jacobszoon Visscher, his chief pilot. He was to explore the Indian Ocean from west to east, south of the ordinary trade route, and, proceeding eastward into the Pacific (if this proved possible), to investigate the practicability of a sea passage eastward to Chile, to rediscover the Solomon Islands of the Spaniards, and to explore New Guinea.

Tasman's voyage of 1642 to 1643

Leaving Batavia (modern Djakarta) on August 14, 1642, with two ships, the "Heemskerk" and "Zeehaen," Tasman sailed to Mauritius (September 5–October 8), then southward and eastward, reaching his most southerly latitude of 49° S at about 94° E. Turning north he discovered land on November 24 at 42°20′ S, and he skirted its southern shores, naming it Van Diemen's Land (now Tasmania). A council of officers on December 5 decided against further investigation, so he missed the opportunity of discovering Bass Strait. Continuing eastward, he sighted on December 13, at 42°10′ S, the coast of South Island, New Zealand, and explored it northward, entering the strait between North Island and South Island supposing it to be a bay. He left New Zealand on January 4, 1643, at North Cape, under the impression that he had probably discovered the west coast of the southern continent, which might be connected with the "Staten Landt" (Staten Island), discovered by W.C. Schouten and Jacques LeMaire south of South America —hence the name of Staten Landt, which Tasman gave to his discovery in honour of the States General (the Dutch legislature).

Convinced by the swell that the passage to Chile existed, Tasman now turned northeast, and on January 21 he discovered Tonga and on February 6 the Fiji Islands. Turning northwest, the ships reached New Guinea waters on April 1 and Batavia on June 14, 1643, completing a ten-month voyage on which only ten men had died from illness. Tasman had circumnavigated Australia without seeing it, thus establishing that it was separated from the hypothetical southern continent.

The council of the company decided, however, that Tasman had been negligent in his investigation of the lands he discovered and of the passage to Chile. They sent him on a new expedition to the "South Land" in 1644 with instructions to establish the relationships of New Guinea, the "great known South Land" (western Australia), Van Diemen's Land, and the "unknown South Land." He sailed from Batavia on February 29, steering southeast along the south coast of New Guinea, coasting Australia's Gulf of Carpentaria, sailing northeast into Torres Strait, which he mistook for a shallow bay, and then following the east coast of Australia to 22° S.

The second voyage of exploration

Although he was rewarded with the rank of commander and was made a member of the Council of Justice of Batavia, his second voyage was also a disappointment to the company because it had failed to reveal lands of potential wealth. In 1647 Tasman commanded a trading fleet to Siam (now Thailand), and in the following year he commanded a war fleet against the Spaniards in the Philippines. He had left the service of the Dutch East India company by 1653 and died probably before October 22, 1659. (H.M.W.)

Tasmania

The island state of the Commonwealth of Australia, Tasmania, lies about 150 miles (240 kilometres) south of the State of Victoria, from which it is separated by the relatively shallow Bass Strait. Physically, Tasmania forms part of the Great Dividing Range; culturally, it is part of Melbourne's metropolitan sphere of influence. The state comprises a main island called Tasmania; Bruny Island, nestling close to the southeast coast near Hobart, the capital; King and Flinders islands in Bass Strait; numerous smaller islands off the coast of the main island; and sub-Antarctic Macquarie Island, some 1,000 miles to the southeast. The main island is roughly heart-shaped, and its latitude and climate are broadly comparable to those of northern California and northwestern Spain. Its 26,383 square miles (68,332 square kilometres), although slightly larger than the area of Sri Lanka, comprise less than 1 percent of Australia.

Location and general character of the state

The state owes its name to the Dutch navigator-explorer Abel Tasman, the first European to discover the island, in 1642, though until 1856 it was known as Van Diemen's Land, after the governor of the East Indies who had sent Tasman on his voyage of exploration. The island of Tasmania contains some of the most spectacular mountain, lake, and coastal scenery in the continent, has much of the country's hydroelectric-power potentiality, and displays a great diversity of natural resources. The Tasmanian population of 390,000 at the 1971 census was only 3 percent of the Australian total, however, and its growth rate lags behind that of the mainland. Throughout much of its history, Tasmania has experienced a net out-migration that has deprived the state of a wealth of talent but has contributed, perhaps disproportionately, to the leadership of the nation. Although insularity renders much of the political, economic, and social life distinctive, proximity to Melbourne and modern air travel make the island less isolated and more progressive than is often assumed in other states. (For related information, see AUSTRALIA; AUSTRALIA, COMMONWEALTH OF; AUSTRALIA, HISTORY OF.)

THE NATURAL AND HUMAN LANDSCAPE

The physical environment. *Topography.* Tasmania is essentially mountainous. In the west, where Mt. Ossa reaches 5,305 feet (1,617 metres), the landscape comprises parallel northwest–southeast ridges and valleys; eastward lies a series of plateaus at various altitudes, the highest being Ben Lomond at 5,160 feet (1,573 metres) in the northeast. But the dominant feature is the glaciated, lake-studded Central Plateau, bounded on the north and east by a 2,000-foot fault scarp and sloping gently southeastward from 3,500 to 2,000 feet (1,100 to 600 metres). Much of the east is made up of a low, dissected plateau averaging about 1,200 feet (365 metres). Extensive plains are confined to the far northwest, the lower South Esk River Valley, and the northeast. The Bass Strait islands represent outliers of the northern coastal platforms. In the southeast, postglacial submergence has produced one of the finest drowned coastlines in the world.

Drainage. There are two major river systems—the Derwent in the southeast and the South Esk in the northeast—many small systems, especially westward flowing to the west coast, and innumerable lakes. The Central Plateau is studded with more than 4,000 lakes in a landscape similar to northern Canada and Finland; almost all, including Great Lake, are shallow. Lake St. Clair, the deepest lake in Australia (over 700 feet [215 metres]), is a piedmont lake similar to the lakes of northern Italy. Several lakes, notably Lake King William, have been created by hydroelectric-power development.

Major river systems

Soils. Most Tasmanian soils are leached, acidic, poorly drained, high in humus, and low in fertility. Least fertile and most extensive are the soils, especially the moor peats, of the west and northeast. Fertile areas occur extensively in the northwest and locally elsewhere, notably in the northeast. Brown earths occupy the drier areas east of the Central Plateau; black earths, the southeast; and alluvial soils, the narrow valley floors to the east. Other fertile soils are those of former swamps in the far northwest and the Bass Strait islands.

Climate. Tasmania, located in the mid-latitude westerly wind belt and dominated by southern maritime air masses, enjoys a moist, equable climate, with mild to warm summers, mild winters in most settled areas, and rain during all seasons. Yet the occasional incursion, in summer, of tropical continental air masses and, in spring and autumn, of tropical Tasman air masses along the east coast, together with the mountainous surface, results in greater annual rainfall variability, seasonal moisture deficiencies, and temperature changes than are the norm in similar climates elsewhere. Climatic averages thus mask considerable variations. The average annual precipitation exceeds 100 inches (2,500 millimetres) on the western

Rainfall and temperature ranges

ranges and declines eastward to under 20 inches in some places; along the north coast it exceeds 30 inches in all locations. The seasonal incidence in the north and west is greatest in winter, and in the south and east it is greatest in spring. Summer rainfall may vary markedly from year to year, especially in the drier east. Mean January temperatures are higher in the north and east than elsewhere, reaching 64° F (18° C) at Launceston; mean July temperatures are 46° to 49° F (8° to 9° C) in all coastal stations, declining sharply with altitude. Almost everywhere, the mean daily range of temperature exceeds the mean annual range.

Vegetation and animal life. In general, the wettest areas have temperate rain forest, especially the beech or myrtle; areas having 30 to 60 inches of rain annually carry good-quality eucalypt forest; and the drier areas carry poor-quality eucalypt forest or savanna woodland. In certain areas, particularly in the forest of the south and southwest, an almost impenetrable thicket known as horizontal scrub develops. This is caused by the growth of a remarkable small tree called the horizontal (known scientifically as *Anodopetalum biglandulosum*). The slender trunk of the tree falls over under its own weight, and from it branches arise that behave in the same way. On the mountain plateaus are found many plants having subantarctic affinities. These include Tasmania's only deciduous tree or shrub, myrtle beech, and certain cushion plants. Rain forest would be more widespread in the absence of fires. Other vegetation includes the sedgeland along the west coast, the high moorlands, and the coastal heaths of the far northwest, the far northeast, and the Bass Strait islands.

Mammals

Animal life is virtually absent from the true rain forest but abounds in the extensive eucalypt forest. The avian fauna includes the honey eaters, the black jay, black magpie, the black cockatoo, and various parrots. Among the mammals are the wallaby, brush and ringtail possum, and the marsupial carnivores—the native cat, tiger cat, Tasmanian devil, and the rare thylacine. The sedgeland and moorland are distinctive for the wombat, and the coastal heath for the green rosella, the platypus, and the echidna.

The human imprint. The inhospitable terrain of much of Tasmania naturally has had much influence on settlement patterns. The nomadic original Tasmanians have left a few archaeological traces, including geometric designs on exposed rock surfaces and evidence of cremations and corroborees, or ceremonial gatherings. The European settler has left his imprint according to economic activity, whether it be in the mining settlements of the west; the intensive cropping or dairying of the northern coastal belt, or of the southeast lowlands; the intensive beef fattening of the northwestern tip; or the dryland sheep farming of most of the eastern sector. The southwestern segment of the island is scarcely utilized by man.

Traditional regions. A widely recognized regionalism bears witness to the diversity of the Tasmanian landscape. In the north the distinctive regions are the far northwest around Smithton; the northwest coast extending along the north coast west of Port Sorell; the Tamar Valley; the northeast around Scottsdale and Lilydale; and the far northeast from Ringarooma eastward. In central Tasmania the regions are the west coast, usually comprising only the Queenstown and Rosebery area; the Central Plateau and the Derwent Valley; the midlands, comprising the area east of the Central Plateau between Hobart and Launceston; and the east coast. Southern regions are the southwest, the Huon Valley, and the southeast. Numerous subdivisions, such as the north midlands comprising the South Esk plains around Launceston and the lower midlands around Tunnack, are also recognized.

Rural building and town patterns

Rural and urban settlement. While the pattern of rural settlement differs strikingly by region, the basic contrasts stem from farm size and the length of settlement. In general, the older settled areas, including the midlands, the central north, the east coast, and the southeast, have larger properties, dispersed homesteads, buildings often of stone or brick, architecture that typically is early Georgian, and nucleated villages laid out on a grid. Higher rainfall areas settled since 1850, chiefly the northwest, the northeast, and the Huon Valley, have generally small farms, buildings mainly of weatherboard, and houses and villages mainly aligned along roads. In all areas, villages normally contain a post office and store, but most villages also have a primary school, public hall, church, service station, and transport services.

Tasmania is less urbanized than other Australian states and exhibits a dispersed pattern of peripheral growth that falls into three urban regions. Hobart (population 130,000 in 1971), located at the foot of Mt. Wellington on the Derwent estuary, is not only the capital city, the premier port, and leading industrial centre but the metropolitan focus for the southeast, the upper Derwent, the Central Plateau, the midlands south of Woodbury, and the east coast south of Swansea. Launceston (62,000), at the head of the Tamar Valley, is a secondary administrative centre and hub of the state's transport network, with important textile and engineering industries. Its sphere of influence extends westward to Elizabeth Town and incorporates the entire north and northeast. The third region centres on both Burnie (20,000) and Devonport (18,000); it includes the northwest and the west coast. Significantly, the Hobart region contains only one full-fledged town; the Launceston region, three; and the Burnie–Devonport region, five. Each, however, has six or seven minor towns.

THE PEOPLE OF TASMANIA

The Aboriginal Tasmanians

Ethnic and religious groups. The original Tasmanians were an anthropologically interesting Negritoid people, with the widest nasal index ever recorded and shorter and broader heads than the Aboriginal peoples of the continental mainland. They may have originally drifted across from the mainland or arrived from as far as the New Hebrides, several thousand miles to the northeast. Their five tribal groups numbered only some 2,000 people at the onset of European settlement in the early 19th century, and the last full-blooded Tasmanian died in 1876. As a consequence of the "black war" and attacks from white outlaws, they were removed in the interim to Flinders Island, off the northeast tip of Tasmania, where the survivors languished and died. Within Australia as a whole, the population of Tasmania has a distinctive composition both by birthplace and nationality. Of all the Australian states Tasmania has the highest proportion born in Australia (over 90 percent in 1966), the lowest proportion born in the British Isles (5 percent), the lowest proportion born in continental Europe (3.5 percent), and the lowest proportion born elsewhere in the world. Nationality is thus overwhelmingly British by descent, the proportion (98.5 percent) being higher than that of any other state.

The origins and nationality of the population are reflected in its religious affiliations. Of the four major denominations, Tasmania has, relative to its population, more Anglicans and Methodists but fewer Roman Catholics and Presbyterians than Australia as a whole. In the postwar years, the only major denomination to show a relative increase has been Roman Catholicism, partly because of immigration from Catholic countries and partly because of stringent religious beliefs regarding the use of modern birth-control methods.

Birth and death rates and migratory patterns

Contemporary demography. Tasmania continues to display a higher birth rate than any other state, but, as in Australia generally, crude birth rates have tended to fall. In the 1950s, they averaged 25 births per 1,000 persons and in the 1960s, 23; in the early 1970s the rate had dropped to just over 20. Birth rates are lower in the cities than in the smaller towns and rural areas. Death rates remain fairly constant (8.1 in the early 1970s); partly because the ratio of children to adults in the population is high, death rates tend to be among the lowest in Australia. Infant mortality rates are also among the lowest.

After World War II Tasmania experienced a net in-migration of population from other states and overseas, but since 1959 there has been, in most years, a resumption of net out-migration to the mainland, particularly of young persons, chiefly men, entering the work force. This out-migration has been accompanied by a pronounced in-

ternal rural-to-urban migration, largely through the increasing scale of farming and the substitution of capital for labour in the farming sector. As a result, Tasmania differs from the mainland in having the smallest proportion of population in the labour force and the lowest growth rate of any state.

Since more than two-fifths of the island in the west and the south is too rugged and too wet for agriculture, the population is largely confined to the north and the southeast, with the sparsely settled midlands connecting them with an isolated cluster on the west coast. Hobart has one-third of the state's population in its metropolitan area and nearly one-half within its sphere of influence; the rest of the people are distributed more or less equally between the respective spheres of influence of Launceston and the Burnie–Devonport area. No other state has so equitable a balance of population among the capital city, other urban centres, and the rural areas, the nearest equivalent being Queensland. One-quarter of the population live in centres of less than 1,000 inhabitants and on farms.

THE STATE'S ECONOMY

The Tasmanian economy enjoys many comparative advantages, notably in its mineral, water, and tourist resources, in the diversity of the economic activity, and in fairly stable labour relations. It suffers markedly, however, from the small scale of much of its resource base, from restricted local markets, and from problems of transport to external markets. At the 1966 census, 14 percent of the work force was in the primary sector (the production of raw agricultural and mineral resources), 23 percent in the secondary (manufacturing and processing), and 63 percent in the tertiary (services).

Resources and primary industries. Resources are highly diversified. Important mineral deposits include iron at Savage River, zinc, lead, and copper at Rosebery, copper at Queenstown, and tin and tungsten in the northeast; poor-quality coal abounds in the Fingal Valley but is of declining importance. The heavy, well-distributed rainfall and the rugged terrain in the centre and the west facilitate hydroelectric-power development. The western forests furnish not only excellent hardwoods but also the raw material for pulp-and-paper industries, while the drier eastern forests yield wood chips. Dairying and mixed farming characterize the wetter north; extensive sheep grazing, the drier midlands and the east coast; and specialized horticulture, notably for apples and hops, the southeast. Apples and pears are also locally important in the Tamar and Mersey valleys, potatoes and canning peas in the northwest, and vegetables in the northeast. Fisheries occur in all coastal waters, the catch mainly comprising crayfish, abalone, shark, barracouta, and scallops. A tourist potential lies in variety of landscape.

Agriculture and fishing

The diversity of resource development is reflected in the relative contribution of each primary industry to the net value of primary production: mining generally contributes about one-quarter, dairying one-fifth, wool growing one-sixth, fruit growing one-seventh, and forestry one-eighth. Iron ore mined by the opencut method at Savage River is converted to slurry, piped to Port Latta near Stanley, pelletized, and exported to Japan. Copper ore mined underground at Queenstown and Rosebery is shipped directly to Japan, Tasmania accounting for nearly one-sixth of the Australian copper output. Tin production fluctuates with market conditions. Butter, wool, and fruit are exported mainly to Britain. Tasmanian hops meet almost all of Australia's requirements. In timber production, Tasmania furnishes about one-sixth of the Australian total; wood chips are exported to Japan. The fishing industry depends largely on exports, fairly evenly distributed between mainland and overseas markets.

Manufacturing. Secondary industries are twice as important in net value of production as are the primary industries. Foremost among them are the electrometallurgical and electrochemical specialties dependent on cheap, bulk power provided by the Hydro-Electric Commission. Thus Tasmania, with only 3 percent of Australia's population, produces more than 10 percent of its electricity. The resultant industries include the electrolytic refinery in Risdon, which treats zinc concentrates from Broken Hill (New South Wales) and Rosebery, the aluminum and ferro-manganese plants at Bell Bay, near George Town, pulp-and-paper mills in Burnie and Boyer near New Norfolk, a pulp mill in Port Huon, and a cement plant in Railton. Other important industries include the manufacture of chocolate near Hobart, titanium pigments (Burnie), textiles (Launceston and Devonport), carpets (Devonport), and alginate from sea kelp (Orford).

Promotion and management. Government intervention is largely at the federal level in the primary industries and at the state level in the secondary industries. The Directorate of Industrial Development seeks to foster manufacturing growth by the provision of financial and other assistance. The state government also is active in promoting tourism and trade. Tourist development is favoured by Tasmania's highly diversified and spectacular scenery, its equable climate, its coastal, lake, and mountain recreation areas, and its colonial heritage of Georgian architecture. As a result, Tasmania is visited by more than 100,000 tourists each year. The private sector has well-established chambers of commerce, a chamber of manufacturers, a tourist council, and more than 100 trade unions.

Transportation. Given its island setting and dispersed pattern of development, transportation is especially important to Tasmania. The Transport Commission regulates and, where necessary, provides internal transport. Tasmania has the highest density of roads per square mile of any Australian states: The government railways comprise about 500 miles (805 kilometres) of line. Of the intrastate freight movement, 60 percent is by road, 30 percent by rail, and 10 percent by sea; virtually all passenger movement is by road. Almost all interstate and overseas freight moves by sea and almost all passengers by air. Of the major seaports, Hobart has about one-third of the trade and Launceston, Burnie, and Devonport each about one-fifth; port administration is decentralized. Minor ports are Stanley, Currie, and Strahan. Of the airports, Hobart handles the most passengers (with Launceston close behind) and Launceston the most freight. Regularly scheduled air services also operate from Devonport, Wynyard, King Island, and Flinders Island.

ADMINISTRATION AND SOCIAL CONDITIONS

Government and politics. The state parliament comprises a House of Assembly, with 30 members, and a Legislative Council, by tradition a largely nonparty house, with 19 single-member constituencies. The system of elections for the House of Assembly is proportional representation by the single transferable vote; the system for the Legislative Council is the preferential system, with an obligation to record preferences, used for the federal House of Representatives. The main political parties are the Australian Labor Party and The Liberal Party of Australia; the Australian Centre Party affiliated with The Australian Country Party dates from 1966. The Labor Party was in power from 1934 to 1969, when it was replaced by a Liberal–Centre party coalition; in May 1972 the Labor Party once more regained office. Executive government is by the Cabinet system, the governor representing the British monarch and the Executive Council of Ministers of State giving Cabinet decisions where necessary to legal form. Local government is exercised through cities and municipalities, which comprise the 49 divisions of the state.

Legislative, executive, and judicial institutions

Although the state now administers some of the lower courts, most are still administered by municipal authorities, a feature unique to Tasmania; in other states the administration is vested in a state department. Courts of petty session have jurisdiction over all summary offenses and certain indictable offenses at the option of the defendant. Minor civil proceedings are dealt with by courts of request in the cities and some municipalities or by courts of general sessions. The Supreme Court of Tasmania sits regularly in Hobart, Launceston, and Burnie; it has jurisdiction over all causes save those reserved to the high court of Australia under the commonwealth constitu-

tion. The high court normally sits in Hobart each year, but Tasmanian cases are also heard in Melbourne and Sydney. Children's courts have jurisdiction over persons under the age of 17.

The social milieu. Wages and working conditions of Tasmanian employees are regulated either by awards of the Commonwealth Conciliation and Arbitration Commission or by state wages boards. Tasmania tends to have the lowest average weekly earnings per employed male of all the states except Queensland and, usually, South Australia. That weekly earnings are not the lowest is in line with the cost of living, which is relatively high, partly because of freight charges on imported goods. Although the standard of public health is good, there is a high incidence of hydatids in some rural areas, and goiter is endemic locally. Social status, as in Australia generally, derives primarily from occupation, but Tasmanians accord a lower ranking to professionals and a higher ranking to businessmen than do Australians as a whole. The standing and numerical importance of small and medium property owners is reflected in the relatively conservative tendency of the two main political parties. There are no great extremes of wealth or poverty.

The general standards of living

Education. School attendance is compulsory between the ages of six and 16. Government schools include infant, primary (some with preschool facilities), area, district, and high schools (nonselective, comprehensive, mostly coeducational), together with matriculation colleges (a Tasmanian innovation) and special schools. Independent (nongovernment) schools with one-sixth of the school population are mostly controlled by religious denominations, the great majority being Roman Catholic; since 1967 they have received state aid. Institutions of higher learning include technical colleges, teachers colleges at Hobart and Launceston, the University of Tasmania (founded 1890) at Hobart, and the College of Advanced Education (opened 1972), which provides technical, education, liberal arts, and various courses.

Health. The state government controls directly or through hospital boards the general hospitals at Hobart, Launceston, Latrobe, and Burnie–Wynyard; numerous district and district nursing hospitals; a mental hospital at New Norfolk; maternity hospitals at Hobart and Launceston; a chest hospital at New Town; and hospitals for the aged at Hobart, Launceston, and Wynyard. It also provides district medical officers for the more remote areas, a district nursing service, a school medical service, a school dental-health service for which it trains dental nurses, and a child-health service. Private institutions include seven hospitals and numerous nursing homes. Annual X-ray examinations for adults are compulsory by law. Public water supplies are fluoridated at Hobart, Launceston, and Beaconsfield.

Welfare. Commonwealth social-service benefits include maternity allowances; child-care payments; unemployment, sickness, and special benefits; age, invalid, and widows' pensions; and funeral benefits. The commonwealth also provides paid employment for the disabled, a rehabilitation service, and subsidies for aged persons' homes. The state Department of Social Welfare assists deserted wives or husbands with children, wives with husbands in prison, and neglected or deserted children. It also maintains homes for maladjusted and delinquent children and wards of the state.

Housing. Three-quarters of all Tasmanian houses are owner occupied, and two-thirds are built of weatherboard; apartments make up less than one-tenth of all dwellings. The Commonwealth Housing Commission administers the homes-savings-grant plan, which makes a gift to young persons with certain minimum savings for a first home, a housing-loans-insurance plan, and assistance to certain former members of the armed forces. The State Housing Commission caters to the lower income groups, providing mainly three-bedroom timber units for families to rent or buy and apartments for elderly persons to rent. Rental rebates are allowed as necessary. Advances for home building are made available by the Agricultural Bank of Tasmania and cooperative building societies.

Public aid for housing

CULTURAL LIFE AND INSTITUTIONS

The arts. For its population, Tasmania has a remarkably vigorous cultural life. At the amateur level, there are numerous musical groups, ranging from the full orchestra to the ensemble, as well as several choral societies and repertory companies. The College of Advanced Education incorporates a conservatorium of music and a school of art. The Australian Broadcasting Commission, with financial support from the Hobart and Launceston city councils, maintains the Tasmanian Symphony Orchestra, which gives regular concerts in the main centres, often with visiting artists from the mainland or overseas. The National Theatre and Fine Arts Society of Tasmania, in association with the Elizabethan Theatre Trust, is responsible for bringing drama and opera from the mainland to the Theatre Royal in Hobart, Australia's oldest theatre (1834), and the National Theatre in Launceston. Both cities have museums and art galleries that exhibit paintings, pottery, and sculpture. Tasmania has well-established film festivals and eisteddfods (Welsh poetry festivals); in 1972 it also held its first arts festival. At Hobart's Battery Point, Narryna, a Georgian town house dating from 1839, has been preserved as the Van Diemen's Land Folk Museum, with furniture and furnishings of the early 19th century. Other Georgian houses that have been restored for public benefit include Franklin House and Entally House, both near Launceston, and Runnymede, at New Town, Hobart.

Library services. The state provides lending-library services to adults and children, including a central service with film and recorded music in Hobart, regional services in Launceston and Burnie, and country services in almost all municipalities. Bookmobiles operating from Hobart, Launceston, and Burnie serve schools and other institutions as well as rural areas that lack library facilities. A state reference library and the state archives are located in Hobart.

The communications media. Tasmania has three daily newspapers, in Hobart, Launceston, and Burnie, and dailies from Sydney and Melbourne are available. Each of the smaller centres also publishes a weekend paper, and weeklies of regional interest are published in Queenstown, Smithton, George Town, Scottsdale, New Norfolk, and Huonville.

In Tasmania, as throughout Australia, broadcasting and television services are produced by both the Australian Broadcasting Commission and commercial transmitters. National radio stations operate from Hobart (two), Launceston, and Queenstown, while commercial radio stations relay from Hobart (two), Launceston (two), Devonport, Burnie, Queenstown, and Scottsdale. There are four television stations, two national and two commercial, Hobart and Launceston both having one of each. Tasmania's rugged relief necessitates the provision of an elaborate network of translator stations to ensure reception in all districts. The Australian Broadcasting Control Board exercises control in certain fields over the commercial private-enterprise services. The island is connected with the mainland by a broad-band radio link that enables television programs to be relayed directly from other states.

PROBLEMS AND PROSPECTS

Compared with the major growth areas of southeastern Australia, the island state of Tasmania is underdeveloped, limited in potential, distinctively more rural, and industrially less mature. Manufacturing continues to be dominated by the first-stage processing of primary products; economic expansion will almost certainly focus increasingly on tertiary activity—especially in the great tourist potential of the extensive areas of spectacular wilderness, in which case the island's face may become its fortune. The greatest potential for economic growth lies in the western half of the island; the eastern half, particularly in the north, displays symptoms of relative economic debility.

Given Tasmania's population dynamics, the most important single export is likely to remain human talent. With a population decentralized yet homogeneous, small

yet culturally vigorous, conservative yet independently innovative, remote yet close to both wilderness and metropolis, Tasmania will long remain the quietest state in a quiet continent.

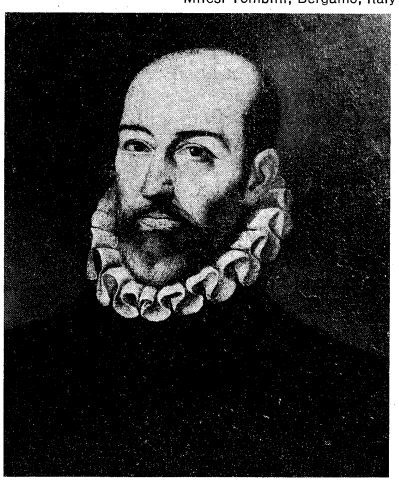
BIBLIOGRAPHY. J.L. DAVIES (ed.), *Atlas of Tasmania* (1965), comprehensive topical and topographical map coverage with detailed commentary; HOBART, COMMONWEALTH BUREAU OF CENSUS AND STATISTICS, *Tasmanian Year Book* (published annually since 1967), the most comprehensive and important single reference on Tasmania; WINIFRED M. CURTIS, *The Endemic Flora of Tasmania*, 3 pt. (1967–71), a definitive study, beautifully illustrated; MICHAEL SHARLAND, *Tasmanian Birds*, 3rd ed. rev. (1958), a field guide; E.R. GUILER, *Marsupials of Tasmania* (1960), an illustrated handbook of diagnostic features; HUNTER VALLEY RESEARCH FOUNDATION, *Tasmania in the Seventies* (1970), an analysis of economic growth and growth potential; P.G. PAK POY AND ASSOCIATES, *Study of the Transport of Goods for Tasmania*, 5 reports (1971), an economic evaluation of Tasmanian transportation commissioned by the state government, including (no. 4) an economic growth study; R.L. WETTENHALL, *A Guide to Tasmanian Government Administration* (1968), a descriptive classification.

(P.Sc.)

Tasso, Torquato

The greatest Italian poet of the late Renaissance, Torquato Tasso wrote the epic *Gerusalemme liberata* (*Jerusalem Liberated*). This poem, through the imitation of classical models, marks stylistically the culmination of the Renaissance poetic manner and anticipates the Baroque trend of the 17th century. It also reveals the "modern" sensibility of its author and the half-morbid feelings of his melancholic disposition, which made such a disconcerting impression on his contemporaries.

By courtesy of L. Locatelli-Milesi-Tombini, Bergamo, Italy



Tasso, oil painting by F. Zuccheri, 1594. In a private collection.

Early life and works

Tasso was born in Sorrento, near Naples, on March 11, 1544, the son of Bernardo Tasso, a poet and courtier, and of Porzia de'Rossi. His childhood was overshadowed by family misfortunes: his father followed the Prince of Salerno into exile in 1552; the family estates were confiscated; his mother died in 1556; and there was subsequent litigation about her dowry. Tasso joined his father in Rome in 1554 and two years later at the court of Urbino, where he was educated with the Duke's son. His imagination had already been fired by stories of the Crusades, and he was struck in 1558 by news of an attack by the Turks on Sorrento, where his sister Cornelia narrowly escaped the accompanying massacre. While in Venice the following year, Tasso began to write an epic in ottava rima (an Italian stanza of eight 11-syllabled lines), *Gierusalemme*, about the First Crusade (which recovered Jerusalem from the Turks in 1099). He soon interrupted its composition, probably realizing that he was too inexperienced to write a historical epic, and turned to themes of chivalry. The resulting *Rinaldo* exhibited his technical ability but not as yet his poetic genius. In 1560 he was sent to study law in Padua and there met the Humanist and critic Sperone Speroni, under whose guidance he studied Aristotle's *Poetics*. It was probably then that he started writing his *Discorsi dell'arte poetica* explaining therein his qualified acceptance of the rules supposedly laid down by Aristotle in 4th-century-BC Greece. (For instance, Tasso maintained that unity of action should not exclude a variety of episodes.) In 1562 he also wrote love poems to Lucrezia Bendidio, a maid-in-waiting to Princess Leonora of the powerful Este family.

In 1565 Tasso entered the service of Luigi, cardinal d'Este, and frequented the court of Duke Alfonso II d'Este at Ferrara, where he enjoyed the patronage of the Duke's sisters, Lucrezia and Leonora, for whom he wrote some of his finest lyrical poems. In 1569 his father died; the following year Lucrezia left Ferrara, and Tasso followed the Cardinal to Paris, where he met a fellow poet, the Frenchman Pierre Ronsard. Back in Ferrara in 1571, he became one of the Duke's courtiers and devoted himself to intense poetic activity. In 1573 he wrote the pastoral drama *Aminta*, which transcends the convention of artificial rusticity with the sensuous, lyrical inspiration of its picture of Arcadia. The play reflects in its idealization of court life the ephemeral period of happiness he enjoyed at Ferrara. In the same year he also began a tragedy, "Galealto re di Norvegia."

Composition of the *Gerusalemme liberata*

In 1575 Tasso completed his masterpiece, the *Gerusalemme liberata*, on which he had been working since his stay in Padua. In this epic poem Tasso narrates the actions of a Christian army led by Godfrey of Bouillon (*Goffredo* was an early alternative title for the poem) during the last months of the First Crusade. In its composition he tried to establish a balance between the moral aspirations of his times and his own sensuous inspiration, and between the requirements of the formal rules laid down for the epic by Renaissance scholars and the impulse of his own lyrical fantasy. He succeeded in reconciling invention with historical truth by adding a number of romantic and even supernatural episodes to the firm groundwork of the principal historical action. Aware of his epic's poetic novelty, Tasso went to Rome in order to arrange its revision by a group of critics. Back in Ferrara in 1576, he started revising his work in a contradictory mood, in which he felt the urge both to accept the criticism he himself had sought and yet to rebel against this kind of authority. He developed a persecution mania, accompanied by unwarranted scruples about his own religious orthodoxy, and the following years were characterized by sudden departures from Ferrara and by violent crises, the latter culminating in his incarceration in the hospital of S. Anna (1579–86) by order of the Duke of Ferrara. During his confinement Tasso wrote a number of philosophical and moral dialogues that, together with his numerous letters, are among the best examples of 16th-century Italian prose. In 1581 the first editions of the *Gerusalemme liberata* and portions of the *Rime e prose* were published. A long controversy started among Italian critics on the respective merits of his epic and of his immediate predecessor Ariosto's chivalric poem *Orlando furioso*, Tasso himself taking part in the controversy with an *Apologia*.

In July 1586 Tasso was released from S. Anna, thanks to the intervention of Vincenzo Gonzaga, prince of Mantua, who received him at his court. After a revival of creative inspiration—at Mantua he completed his tragedy *Galealto*, re-entitled *Re Torrismondo*—he relapsed into his usual inquietude and fled from Mantua, wandering mainly between Rome and Naples, where he composed his religious poems *Monte Oliveto* and *Le sette giornate del mondo creato*. In May 1592 he was given hospitality in Rome by Cardinal Cinzio Aldobrandini, a nephew of Pope Clement VIII. To this patron he dedicated a new version of his epic (*Gerusalemme conquistata*, published 1593), a poetic failure that reveals the extent of Tasso's final submission to the moral and literary prejudices of the times. After writing two more religious poems (*Lagrime di Maria Vergine* and *Lagrime di Gesù Cristo*), in

June 1594 he went to Naples again, where his *Discorsi del poema eroico* was published. In the *Discorsi* he tried to justify the new version of his epic according to his modified conception of art. On Tasso's return to Rome in November 1594, the Pope granted him an annual pension and promised to make him poet laureate. But Tasso fell ill in the following March, was moved to the Convent of S. Onofrio, and died on April 25, 1595.

Reputation and influence

It was not long before *Gerusalemme liberata* was translated and imitated in many European languages, but its mixed reception following publication aggravated Tasso's melancholic disposition. This melancholy was partly caused by a contrast between his hedonistic ideals—as displayed in his pastoral drama *Aminta*—and his religious aspirations in accordance with the rigid morality of the Counter-Reformation, the Roman Catholic Church's movement of revival in answer to Protestantism. The restlessness of the poet's life, his mental illness, his supposed romantic loves, and his alleged persecutions all gave rise to a legend about the man: Tasso became a literary subject first in 17th-century Italy and then in 18th- and 19th-century Europe, being finally represented as the man of genius who is misunderstood and persecuted. Modern criticism refers the peculiarities of his life and character to the moral incertitude of the times.

MAJOR WORKS

POEMS: *Rinaldo* (1562; trans. by J. Hoole, 1792); *Rime* (1581–87; a few poems translated into English, including those by C. Chorley, *From the Italian of Tasso's Sonnets*, 1867); *Gerusalemme liberata* (pirated ed. pub. as *Il Goffredo*, 1580; 1st authorized ed., 1581; trans. by R. Carew, *Godfrey of Bulloigne; or, the Recoverie of Hierusalem*, with Italian text, 1594; by E. Fairfax as *Godfrey of Bulloigne; or, the Recoverie of Ierusalem*, 1600; J.K. James, 2 vol., 1884); *Gerusalemme conquistata* (1593); *Monte Oliveto* (1605); *Mondo creato* (*I due primi giorni del mondo creato*, 1600; *Le sette giornate del mondo creato*, 1607).

DRAMA: *L'Aminta* (performed 1573; pub. 1581; trans. by A. Fraunce, 1591; by L.E. Lord, 1931); *Re Torrismondo* (1587).

PROSE: *Discorsi del'arte poetica* (pub. 1587; rewritten as *Discorsi del poema eroico*, pub. 1594); *Apologia*, 2 pt. (1585); *Dialoghi* (written 1578–94; pub. 1581 and afterward); *Lettere* (1587 and afterward).

BIBLIOGRAPHY

Standard editions: *Le Rime de Torquato Tasso* by ANGELO SOLERTI, 3 vol. (1898–1902), is still fundamental while a new critical edition is in preparation; *Gerusalemme liberata*, by LANFRANCO CARETTI (1957), is the nearest to a critical edition; *Aminta*, by BARTOLO T. SOZZI (1957); *Gerusalemme conquistata*, by LUIGI BONFIGLI (1934); *Dialoghi*, by EZIO RAIMONDI, 3 vol. (1958), a critical edition; *Le Lettere di Torquato Tasso*, 5 vol. (1852–55) and *Le Prose diverse di Torquato Tasso*, 2 vol. (1875), by CESARE GUASTI; *Appendice alle opere in prosa*, by ANGELO SOLERTI (1892); *Opere minori in versi*, by ANGELO SOLERTI, 3 vol. (1891–95); *Mondo creato*, by G. PETROCCHI (1951), a critical edition; *Il re Torrismondo* in *Opere*, by BARTOLO T. SOZZI, vol. 2 (1956); *Discorsi dell'arte poetica e del poema eroico*, by LUIGI POMA (1964), a critical edition.

Biographical and critical studies: ANGELO SOLERTI, *Vita di Torquato Tasso* (1895), a detailed biography with an appendix of documents; LANFRANCO CARETTI, *Ariosto e Tasso* (1961), a valuable introduction to the study of Tasso's poetry considered in relation to the contemporary historical climate; C.P. BRAND, *Torquato Tasso: A Study of the Poet and of His Contribution to English Literature* (1965), includes a revaluation of the *Gerusalemme conquistata* and an original study of the poet's fortunes in England.

(G.A.)

Taxation

This article is concerned with taxation in general, its principles, its purposes, and its effects; specifically, the article discusses the nature and purposes of taxation, classes of taxes, the history of taxation, canons and criteria of taxation, shifting and incidence, and the structure of national tax systems in various countries. A more extended discussion of particular kinds of taxes will be found in the articles INCOME TAX, PERSONAL; INCOME TAX, CORPORATION; PROPERTY TAX; SALES AND EXCISE TAXES; TARIFFS; and DEATH AND GIFT TAXES.

THE NATURE AND PURPOSES OF TAXATION

In modern economies taxes are the most important source of governmental revenue. They are compulsory levies that are regularly imposed and, as a rule, not designated for a special purpose; they are regarded as a contribution to the general revenue pool from which most government expenditures are financed. Taxes differ from other sources of revenue in that they are unrequited—*i.e.*, they are not paid in exchange for some specific thing, such as the sale of public property or the issue of public debt. While taxes are presumably collected for the sake of the welfare of taxpayers as a whole, the liability of the individual taxpayer is independent of any benefit received.

Tax legislation customarily distinguishes between the tax *object* and the tax *base*. The tax object may consist of goods, transactions (*e.g.*, sales, purchases of real estate, imports, etc.), or sums of money (*e.g.*, income, net wealth, inheritances). The tax base is the physical unit or monetary amount to which the tax rate is applied. For example, a levy on automobiles (the tax object) may use as the tax base the weight of the automobile, its horsepower, its age, its value, etc. Similarly, the property tax may be based on gross value or rental; an excise duty on sugar may be levied as a percentage of the retail price or as a fixed sum per ton of the finished product, etc.

The various functions of taxation

During the 19th century the prevalent idea was that taxes should serve mainly to finance the government. In earlier times, and again today, governments have utilized taxation for other than merely fiscal purposes. Current theories suggest that governments should not use the tax instrument as a revenue-raising device exclusively. Taxes are considered to have three functions: (1) fiscal or budgetary, to cover government expenditures insofar as they are not financed from other sources (fees, profits from public enterprises, the issue of public debt, the creation of money); (2) economic, to promote such general goals as full employment, monetary stability, and a satisfactory rate of economic growth within the framework of a market economy; and (3) social or redistributive, to lessen inequalities in the distribution of income and wealth to the extent they are considered excessive and unjust.

Since these three functions are interrelated, there are likely to be conflicts among them. Thus the level or composition (or both) of taxes considered necessary for budgetary reasons may tend to hold back the rate of economic growth. Or taxes that are highly redistributive may also conflict with the desired goal of economic growth. On the other hand, a relatively high and steady rate of economic growth will bring with it higher tax revenues, which in turn will enable the government to pursue other aims, fiscal and redistributive.

Aside from its main functions, taxation has many lesser purposes. Certain consumption goods considered undesirable, such as alcoholic beverages and cigarettes, may be taxed heavily on the grounds of national health (though more often than not this justification has been put forward to conceal a purely fiscal desire for more revenue). Income taxes and succession duties have been used since ancient times to affect population growth; the most conspicuous examples are bachelor taxes and income taxes on childless couples graduated according to the length of time they have been married. It is doubtful whether such tax inducements actually achieve the objectives sought; even if they do, there are probably more efficient ways of influencing human behaviour.

CLASSES OF TAXES

In the literature of public finance, taxes have been classified in various ways according to who pays them, who bears the ultimate burden of them, the extent to which the burden can be shifted, and various other criteria. The most common distinction is between direct and indirect taxes, an example of the former being the income tax and of the latter the sales tax. There is much disagreement among economists as to the criterion for distinguishing between direct and indirect; it is usually said that a direct tax is one that cannot be shifted by the taxpayer to someone else. Some theorists have attempted to show that in principle any tax may be shifted, in whole or in part,

depending upon the conditions under which the tax is levied. Nevertheless, it is conceded that some taxes are more likely to be shifted than others.

Direct taxes. Direct taxes are primarily taxes on persons; they are aimed at the individual's ability to pay as measured by his income or his net wealth. The main types of direct taxes are the following:

Individual income taxes. These are taxes levied on the total personal net income in excess of some stipulated minimum. They are adjusted to take into account the circumstances influencing the ability to pay of the individual, such as family status, number and age of children, financial burdens resulting from illness, etc.

Taxes on net worth. These are levied on the total net worth of a person—that is, the value of his assets minus his liabilities. As with the income tax, the personal circumstances of the individual are taken into consideration. Very often there is double taxation when a net worth tax falls upon both corporations and individuals.

Death duties. Taxes at death take two forms: the inheritance tax, where the tax object is the bequest received by the person inheriting, and the estate tax, where the object is the total estate left by the deceased. Inheritance taxes usually allow for the personal circumstances of the taxpayer, including his net worth before receiving the bequest and his relation to the donor. Estate taxes are generally graduated according to the size of the estate, although in some countries they make allowance for the number of children involved. In order to prevent the death duties from being circumvented, any rational and efficient system has to include a tax on gifts between living persons, particularly those made in anticipation of death.

General expenditure or spending taxes. These are taxes on all income that is not channelled into savings. In contrast to indirect taxes on spending, such as the sales tax, an expenditure tax can be adjusted to an individual's ability to pay by allowing for marital status, age, number of dependents, etc.

Indirect taxes. Indirect taxes include taxes on sales of consumer goods, on goods in the process of production, on legal transactions, etc., where the assumption (sometimes erroneous) is that the tax will be shifted in whole or in part to the ultimate consumer of the goods, who presumably has the ability to pay.

Consumption taxes and customs duties. The tax objects in these instances are commodities or services. They are levied on almost everything, from necessities such as bread, meat, and salt, to nonessentials such as cigarettes, wine, liquor, coffee, and tea, to luxuries such as jewels and furs. Taxes on specific commodities are called excises, as distinguished from sales taxes and other general consumption levies.

Taxes on consumer durables. In former times these taxes were applied to luxury commodities such as pianos, saddle horses, carriages, and billiard tables. Today the main tax object is the automobile.

Taxes on intermediate goods and production factors. These are levied on raw materials, intermediate goods (*e.g.*, mineral oil, alcohol), machines, or labour; in the last case, the tax base may be the number of persons employed, as in the British "Selective Employment Tax," or the employer's payroll, as in the German *Gewerbesteuer*.

Taxes on legal transactions. Examples of these are taxes on the issue of shares, on the sale of houses and land, and on stock exchange transactions. They are frequently levied in the form of stamp duties.

Taxes that may be direct or indirect

Some taxes cannot be said with certainty to be either direct or indirect. This is because they may, under certain assumptions, be shifted by the taxpayer to someone else. The current prevailing opinion is that corporation income taxes are partially or wholly shifted, although the results of some empirical studies are not altogether convincing on this point. (When the corporation income tax is accompanied by a personal income tax, it also involves double taxation of the dividend element in profits.)

Other taxes that are ambiguous in their incidence include the so-called real taxes on elements of income or net worth, including real estate, business enterprises, and earnings on capital investment. In the 19th century they were common in western Europe. They are found now mainly in the less developed countries (on the national level) and at the local community level of advanced economies. Property taxes are the mainstay of local tax revenue in the United States.

There are other ways of classifying taxes. Some taxes are specific—*i.e.*, they are levied on the basis of weight, length, volume, and other specific characteristics of the tax object. Other taxes are ad valorem—levied on the value of the goods as measured by the price.

One can also distinguish between taxes on the basis of the effect they have on the distribution of income and wealth. A proportional tax is one that imposes the same relative burden on all taxpayers—*i.e.*, where the tax liability and the tax base grow in equal proportion. A progressive tax is characterized by a more than proportional rise in the tax liability relative to the increase in the tax base, and a regressive tax is characterized by a less than proportional rise in the relative burden. Examples of taxes that are generally considered progressive are income taxes, expenditure taxes, and death duties. Taxes that are nominally proportional include, as a rule, the corporation income tax, the various taxes on elements of income or net worth, and ad valorem excises and stamp duties. But there is more to be said: an ad valorem tax on cigarettes or sugar that is nominally proportional will turn out to be decidedly regressive in practice, because the share of personal income spent on cigarettes and sugar declines as the level of personal income rises. Poll taxes, levied as a fixed amount per capita, obviously are regressive. Even income taxes that are nominally progressive may become less so in the upper income categories, when it is possible to reduce the tax base by means of deductions or by the exclusion of certain income components from taxable income; on the other hand, in the lower income categories to which a proportional tax rate applies, some progression is introduced through the personal exemptions.

Progressive and regressive taxes

HISTORY OF TAXATION

Taxes played a relatively minor role in the ancient world. Taxes on consumption were levied in Greece and Rome. Tariffs—taxes on imported goods—were often of considerably more importance than internal excises so far as the production of revenue went. As a means of raising additional funds in time of war, taxes on net worth would be temporarily imposed. For a long time these net worth taxes were confined to real property, but later they were extended to other elements of net worth. Real estate transactions also were taxed. In Greece free citizens had different tax obligations from slaves, and in Rome the tax laws distinguished between nationals and residents of conquered territories.

In Rome, along with consumption taxes and customs duties, there were certain "direct" taxes. The principal of these was the *tributum* paid by citizens and usually levied as a head tax; later, when additional revenue was required, the base of this tax was extended to real estate holdings. In the time of Julius Caesar a 1 percent general sales tax was introduced (*centesima rerum venalium*). The provinces relied for their revenues on head taxes and land taxes; the latter consisted initially of fixed liabilities regardless of the return from the land, as in Persia and Egypt, but later the land tax was modified to achieve a certain correspondence with the fertility of the land, or, alternatively, a tenth of the produce was collected as a tax in kind (the "tithe"). It is noteworthy that at a relatively early time Rome had an inheritance tax of 5 percent, later 10 percent; however, close relatives of the deceased were exempted. For a long time tax collection was left to middlemen or "tax farmers," who contracted to collect the taxes for a share of the proceeds; under Caesar collection was delegated to civil servants.

In the Middle Ages many of these ancient taxes, especially among the direct levies, vanished and gave way to a variety of obligatory services and a system of "aids" (most of which amounted to gifts). The main indirect taxes were transit duties and market fees. In the cities, the concept developed of a tax obligation encompassing all residents: the burden of taxes on certain foods and beverages was

intended to be borne partly by consumers and partly by producers and tradesmen. During the latter part of the Middle Ages, some German and Italian cities introduced several direct taxes: head taxes for the poor and net worth or occasionally crude income taxes for the rich. (The income tax was administered through self-assessment and an oath taken before a civic commission.) The use of land taxes and taxes on houses gradually spread.

Modern trends

The development of taxation in recent times can be summarized by the following general statements, although allowance must be made for considerable national differences: The level of most taxes has risen substantially and so has the ratio of tax revenues to the national income. Taxes today are collected in money, not in goods. Tax farming has been abolished; taxes are assessed and collected by civil servants. There has been a reduction in special excises and a shift from the taxation of necessities to taxes on nonessential commodities of general consumption and on luxuries. Many countries increasingly rely on sales taxes and other general consumption taxes (including turnover taxes paid by manufacturers, wholesalers, and retailers on their gross sales and value-added taxes levied on the difference between gross returns and the costs of materials, supplies, packaging, and overhead). While objective taxes on selected elements of income or net worth, especially those levied on business and real property, played a considerable part in the tax systems of the 19th and early 20th centuries, they have since lost ground and are important today mainly as revenue sources for local communities. The general tendency seems to be toward a closer correspondence between individual tax liability and ability to pay; consequently, the absolute and relative weight of direct personal taxation has been increasing in most of the developed countries. Income taxation, both of individuals and of corporations, occupies a leading position in modern tax systems. The income tax has ceased to be a rich man's tax; it is now paid by the general populace, and in several countries it is complemented by a tax on net worth. The growing emphasis on the ability-to-pay principle and on the redistribution of wealth has led to graduated rates, especially in the case of income and inheritance or estate taxes and sometimes in the case of the net worth tax. A good deal of fiscal centralization has occurred, as reflected in the kinds of taxes levied by central governments. The latter now control the most important taxes (from an economic and social point of view, as well as budgetary): income and corporation taxes, inheritance or estate duties, turnover and value-added taxes. The authority of the sovereign to levy taxes in a more or less arbitrary fashion, however, has been lost, and the power to tax now generally resides in parliamentary bodies.

Taxes as an issue in politics, revolutions, and wars

Taxes have been a major subject of political controversy throughout history, even before they constituted a sizable share of the national income. A famous instance is the rebellion of the American colonies against Great Britain when the colonials refused to pay taxes imposed by a Parliament in which they had no voice; hence the slogan, "No taxation without representation." Another instance is the French Revolution of 1789, in which the inequitable distribution of the tax burden was a major factor. It is worth noting that revolutions often bring with them radical tax reforms (usually including high capital levies) but that these mostly fail to achieve their aims. In the long run, however, such reforms often become reality; for example, in 19th-century France some of the tax proposals of the Revolution—elimination of tax exemption for clergy and nobility, lower taxes for farmers, craftsmen, and other small businessmen—were finally put into effect.

Wars have influenced taxes much more than revolutions. Many taxes, notably the income tax (first introduced in England in 1799) and the turnover or purchase tax (Germany, 1918; Great Britain, 1940) began as "temporary" war measures. Similarly, the withholding method of income tax collection began as a wartime innovation in France, the U.S., and Britain.

It is hardly necessary to mention the role that tax questions play in peacetime politics, where the influence of powerful, well-organized pressure groups is great. Arguments for tax reform, particularly in the area of income and inheritance taxes, are at issue in the domestic politics of many countries today.

PRINCIPLES OF TAXATION

Adam Smith first systematized the rules that should govern a rational system of taxation. In *The Wealth of Nations* (Book V, Chapter II) he set down four general canons: that taxes should be based on the individual's ability to pay, and that they should be certain, convenient, and economical.

> I. The subjects of every state ought to contribute towards the support of the government, as nearly as possible, in proportion to their respective abilities; that is, in proportion to the revenue which they respectively enjoy under the protection of the state. . . .
>
> II. The tax which each individual is bound to pay ought to be certain, and not arbitrary. The time of payment, the manner of payment, the quantity to be paid, ought all to be clear and plain to the contributor, and to every other person. . . .
>
> III. Every tax ought to be levied at the time, or in the manner, in which it is most likely to be convenient for the contributor to pay it. . . .
>
> IV. Every tax ought to be so contrived as both to take out and keep out of the pockets of the people as little as possible over and above what it brings into the public treasury of the state. . . .

Today only the first of these canons is considered of prime importance. The other three have been overshadowed by more recent developments in the theory of taxation. A history of tax theory would include a number of other criteria such as "moderation," "distributive neutrality," "preservation of a purely fiscal orientation," "imperceptibility," and "insensitivity to business cycles"—all of them swept away in the fiscal revolution of the mid-20th century. A similar fate befell the "cost" or "benefit" principles of taxation, based on the idea that there should be some equivalence between what the individual pays and the benefits he derives from governmental activities.

Nowadays the criteria of a good tax system are wide ranging and complex. Some of them fall under the heading of *revenue productivity*. Others involve considerations of *social justice*. There is a further category of considerations having to do with *economic goals*. And finally there are the Smithian criteria of *ease of administration and compliance*. With regard to the first category, revenue productivity, a tax system should fulfill two requirements.

Revenue productivity

The principle of adequacy. A national tax system should guarantee revenues adequate to cover the expenditures of government at all levels. Since public expenditures tend to grow at least as fast as the national product, taxes as the main vehicle of government finance should produce revenues that grow correspondingly. In developed economies this criterion would give first place to income, sales, and turnover or net-value-added taxes; other productive taxes are levies on net worth and, to a limited extent, property taxes and taxes on nonessential articles of mass consumption such as tobacco products, alcoholic beverages and—to a lesser degree—sugar, coffee, etc. (Automobile and gasoline taxes are omitted in this connection because of their special nature and because the revenue derived from them tends to be designated for particular purposes.)

The principle of adaptability. A tax system should be flexible enough to produce additional revenue at short notice (when this is necessary to cover unforeseen expenditures) without causing economic disruption. The need may be to finance a war or to pay for an important social reform (*e.g.*, the introduction of old-age pensions). The most important taxes in time of war are income tax surcharges, excess profits or turnover, or net-value-added taxes. While an increase in net worth, consumption, and inheritance taxes is of much less budgetary importance, it may be a useful psychological accompaniment of the other tax increases.

Social justice

The modern conception of social justice requires in addition that taxes be universal, equal, and responsive to the individual's ability to pay. In the modern welfare state, furthermore, taxes are expected to be more or less redistributive of income or wealth or both.

The principle of universality. This postulates that all persons able to pay should be taxed uniformly, and that exemptions from any tax must serve overall economic, social, and other goals. The universality principle is not violated by the basic exemption from the income tax, the purpose of which is to leave some minimal level of income untouched, nor by the exemption of necessities from a general consumption tax. But differentiations based on such factors as nationality, social class, religion, or race do violate the universality principle.

The equality principle. This requires that persons in the same or similar positions so far as tax purposes are concerned be subject to the same tax liability. In practice the equality principle is often disregarded, both intentionally and unintentionally. Intentional violations of it are usually motivated by considerations of policy (*e.g.*, the tax advantages granted to farmers, artisans, or members of the middle class in general; the exclusion of interest on government securities; the preferred treatment of capital gains). It is debatable whether income splitting for tax purposes, as practiced in the U.S., France, and Germany, is compatible with the equality principle: on the one hand, it seems inequitable for a married couple to pay higher taxes than two unattached individuals with the same total income, as happens in the absence of income splitting; on the other hand, a married couple may have certain economic advantages over two single persons, so that the optimal solution may be some modified form of splitting (like that introduced in the U.S. by the tax reform of 1969).

The ability-to-pay principle. This requires that the total tax burden be distributed among individuals according to their capacity to bear it, taking into account all of the relevant personal characteristics in such a way that the relative loss in economic capacity resulting from the tax is equal. The primary indicator of ability to pay is income, complemented by the subsidiary consideration of net wealth. To include the wealth criterion is desirable not only because wealth produces income but also because its ownership implies some degree of tax capacity even if it generates no tangible income (*e.g.*, collections of art or jewelry). The most suitable taxes from this standpoint are personal levies (income, net worth, expenditure, and inheritance taxes). Indirect taxes such as excise, sales, turnover or net-value-added taxes can be adapted to the ability-to-pay criterion only to a limited extent—for example, by the exemption of necessities or the differentiation of tax rates according to "urgency of need." The ability-to-pay principle also requires that direct personal taxes have a progressive rate structure, although there is no way of demonstrating that any particular degree of progression is the right one.

The principle of redistribution. Currently it is felt that the distribution of the tax burden among individuals ought to have the effect of reducing the large disparities of income and net wealth that result from the functioning of the market economy. This is a complete volte-face from the 19th-century liberal view that the distribution of income ought to be left alone. The modern argument is that because of various restrictions and monopoly elements, the market mechanism does not necessarily (even in the long run) distribute income and wealth according to effort and merit; moreover, poverty may result from some work disability or from an unfavourable social or family background. The most suitable means of carrying out this redistribution is through progressive income and inheritance (or estate) taxes. But since a considerable part of the population does not pay income or inheritance taxes at all, such taxes ordinarily need to be supplemented by direct income transfers such as subsidies or "negative income taxes" if a satisfactory redistribution is to be achieved.

Consistency with economic goals

As governments have come to play a larger part in controlling their national economies, they have had to consider the ways in which taxes help or hinder them in achieving their economic goals. To the extent that tax policy helps to bring about a higher level of economic activity or to moderate downswings, it also helps to produce higher revenues.

Taxes as countercyclical measures. Modern fiscal policy advocates tax flexibility as a way of promoting overall economic stability. One may distinguish two kinds of flexibility: automatic flexibility that is tied to changes in the level of economic activity, without intervention by the authorities, and "discretionary" flexibility involving decisions by the authorities to change tax rates or the tax base. *Automatic flexibility* implies a tax system that cushions the cyclical fluctuations in prices, employment, and production through variations in tax liabilities (and, consequently, revenues) without changes in the tax laws and without detrimental effects on long-term economic growth. Taxes differ in their cyclical characteristics: individual and corporation income taxes show the strongest sensitivity to cyclical movements, their returns varying more than the variations in the national product; moreover, since they account for a huge portion of total tax revenues, their cushioning effect is quite large. Of course, such built-in flexibility requires that the extra revenues received by the government in boom periods be taken out of circulation through a budgetary surplus and not spent, while the decreased revenues received during recessions must register as a budgetary deficit—*i.e.*, the government must not cut back its expenditures during a recession but instead must cover its deficit through additional short-term borrowing. *Discretionary flexibility* calls for more active measures, primarily changes in the tax rates themselves, to increase the anticyclical effects of tax receipts. The effectiveness of discretionary changes depends on the speed at which they can be carried out and the time required for the changes to be felt in the economy. Since the normal legislative process is very time consuming, this policy would be facilitated by giving to the executive limited authority to change tax rates. The West German federal government was given the power in 1967 to change individual and corporation income tax rates by decree (to a maximum of 10 percent up or down in a given year) with the consent of the legislature. The British chancellor of the exchequer can vary the rates of the purchase and excise taxes without requiring approval by Parliament. In the U.S., Presidents John F. Kennedy and Lyndon B. Johnson asked Congress in the 1960s for "standby authority" to make a limited reduction of federal income tax rates but failed to obtain it.

Taxation and economic growth. Policies that promote cyclical stability are generally considered to be favourable to long-term economic growth as well. But since numerous factors influence growth, some economists have proposed separate criteria for shaping tax policies so as to specifically promote growth. These would require that tax policies eschew measures hindering growth and emphasize measures stimulating it. The precise measures would depend on the situation. In general, this approach implies a qualitative restructuring of the tax system (for example, the substitution of indirect taxes for direct taxes to some extent) or special tax advantages to stimulate saving, labour mobility, research and development, etc.

Other economic ends. Additional requirements of an efficient tax system arise from the nature of a market economy. Tax policy should refrain from pursuing particular ends to the point where it leads to the misallocation of economic resources. It should not discriminate in favour of, or against, particular means of transportation, particular forms of retail organization, or particular industries. It should not seek to manipulate the balance of payments, as does the German interest tax on bonds held by foreigners, the U.S. "Interest Equalization Tax," and the British "Selective Employment Tax." In general, so far as the efficient working of a market economy is concerned, taxation should imply a minimum of interference with individual privacy and freedom of activity. This does not mean, of course, that major social and economic goals may not take precedence over these considerations.

Ease of administration and compliance

In discussing these general principles one must not lose sight of the fact that taxes have to be administered. This imposes certain limitations on the fiscal process. There are four general requirements for the efficient administration of tax laws.

The requirement of clarity. Tax laws and regulations must be comprehensible to the taxpayer; they must be unambiguous and certain, both to the taxpayer and to the tax administrator. While the principle of certainty is better adhered to today than in the time of Adam Smith, and arbitrary administration of taxes has been reduced, every country has tax laws that are far from being generally understood by the public. Not only does this result in a considerable amount of error; it also tends to discriminate against the ignorant and the poor, who cannot take advantage of the various legal tax-saving opportunities that are available to the educated and the affluent.

The requirement of continuity. Tax laws should be changed seldom, and when changes are made they should be carried out in the context of a general and systematic tax reform.

The requirement of economy. The costs of assessing, collecting, and controlling taxes should be kept to the lowest level consistent with other goals of taxation. This principle is of secondary importance; obviously, equity and economic rationality should not be sacrificed to cost considerations. The costs to be minimized include not only government expenses but also those of the taxpayer and of private fiscal agents such as employers who collect taxes for the government through the withholding procedure.

The requirement of convenience. The payment of taxes should cause the taxpayer as little inconvenience as possible, subject to the limitations of higher ranking tax principles. Governments usually allow the payment of large tax liabilities in installments and set generous time limits for the completion of returns. (F.N.)

SHIFTING AND INCIDENCE

It is fundamental that the real burden of taxation does not necessarily rest upon the person who pays the tax. General sales taxes are paid by retailers, but most of the burden is assumed to rest upon those who buy the goods that are taxed. In other words, the tax is shifted from the retailer to the consumer. Previously in this article "direct" and "indirect" taxes were distinguished according to whether they were borne entirely by the taxpayer or were shifted to someone else. Taxes may be shifted in several directions. *Forward shifting* takes place if the burden falls entirely on the user, rather than the supplier, of the commodity or service in question. (Thus an excise tax or tariff on luxuries increases their price to the purchaser, a tax on bonds leads to a rise in the before-tax price of the bond, etc.) *Backward shifting* occurs when the price of the article taxed remains the same and the cost of the tax is borne by those engaged in producing it—*e.g.*, in lower wages and salaries, less employment, lower prices for raw materials, or a lower return on capital. *Tax capitalization* occurs if the burden of the tax is incorporated in the value of long-term assets—*e.g.*, if the price of land declines by an amount that offsets an increase in property taxes. At first glance, capitalization does not seem to imply any shifting, since the present owner of the asset takes a capital loss. But when he sells the asset, its price will be lower by the capitalized value of the increased tax payments.

Who ultimately pays

The difficulty with this classification is that a tax may be partly borne by the taxpayer and partly shifted—in any or all of three directions. What actually happens will depend on all of the forces at work. In a market economy the introduction of any tax triggers a whole series of adjustments in consumption, production, the supply of productive factors, and the pattern of foreign trade. These adjustments in turn will have repercussions on the prices of various commodities, productive factors, and assets that may be far removed from the area of the initial impact. In other words, a tax levied on a certain object may affect the prices of nontaxed goods and services that are not even used in the production of the tax object. Thus the initial impact of a tax does not indicate where the ultimate burden will rest unless one knows what repercussions the tax will have throughout the system of interrelated economic variables—*i.e.*, unless recourse is made to general economic theory. In what follows, the attempt will be made to isolate some of the factors involved.

The basic principle of tax shifting

The direction and extent of tax shifting is determined basically by one principle: The user of a tax object can avoid the tax burden to a greater (lesser) extent the easier (the more difficult) it is for him to find nontaxed or less-taxed alternatives or substitutes for the tax object; the supplier of a production factor that is taxed itself or used in the production of a taxed good can avoid the tax burden to a greater (lesser) extent the easier (the more difficult) it is for him to find equivalent nontaxed or less taxed alternative employment opportunities for this factor. Since the demand for substitute goods will increase, their prices may rise, thus placing part of the tax burden on those individuals who used them before the tax was imposed. Likewise, the productive factors that seek alternative employments to avoid the tax will tend to receive lower returns in those employments, thus placing part of the burden on individuals who supplied the factors in those sectors before the tax was imposed. For example, if wine is taxed while beer is not, and these two beverages are regarded as perfect substitutes and the price of beer does not rise with increased demand, the tax burden will fall on the owners of land used for viticulture and on the workers engaged in it. It will fall mainly on the landowners if the soil is specific to winegrowing and if labour has alternative employment possibilities. If, on the other hand, wine drinkers are determined to drink only wine, they will bear most of the tax burden. If some substitution of beer for wine takes place and the price of beer rises somewhat, both wine and beer drinkers will bear the burden.

If a tax causes a significant redistribution of real income, then, in addition to the substitution effect discussed above, one must take into account an income effect: if different income classes have different propensities to save and different patterns of consumption, then income redistribution will influence the demand for various goods and for the labour and materials used in making them.

There are other considerations that affect tax shifting, but they are derived from the basic principle of substitution. The following are most common.

The time period. The direction and extent of shifting may vary over time, depending on how long it takes to adjust consumption patterns, reallocate land and capital, retrain labour, etc. Those users and suppliers who have the most difficulty in adjusting will bear the largest burden.

The kind of tax. The broader (narrower) the tax base (*i.e.*, the more [less] inclusive the scope of the tax from the user's point of view), the higher (lower) his tax burden, since the range of nontaxed or less taxed substitutes at his command is narrower (wider). Thus an excise tax allows partial shifting through a change in the consumption pattern, while a general consumption tax does not. In a similar fashion, the returns on capital will be affected less by the taxation of corporation profits alone than by the taxation of both corporation and noncorporation profits.

The geographical unit affected. The smaller the jurisdictional unit imposing the tax, the easier it tends to be for a user to obtain nontaxed or less-taxed substitutes from outside the jurisdiction and for a supplier to find nontaxed or less-taxed outside employment opportunities for his goods and services. Unless there are border tax adjustments for commodity, labour, and capital movements between jurisdictions (as there are today at national borders), tax differences among regions usually will be shifted to the users or suppliers of commodities and productive factors that are immobile. In small communities the only really immobile factors are likely to be real estate, certain local services, and poor families.

Market imperfections. The rigidities of imperfect markets are likely to increase the uncertainty of the shifting response. Thus a monopolist may absorb part of a tax in lower profits rather than shift all of the burden to the user of the product. In industries where there are few firms (oligopoly), the price behaviour of a firm is mainly determined by what it expects its competitors to do. There is also the phenomenon of "money illusion," which leads individuals to adjust their saving and consumption decisions to changes in their money income rather than their real income; this makes them less responsive to price

increases than they are to decreases in their pay, and thus a tax increase may have very little effect on their consumption. Rigid prices are likely to increase the incidence of taxes on employment, unless monetary policy allows the tax-induced changes in relative prices to take place in the setting of a generally rising price level.

Empirical studies

All of these considerations are analytical and theoretical. Efforts have been made to measure the impact of taxation by studying the actual effects of a particular tax on income and employment. These studies reflect the obvious difficulty that the tax impact cannot be isolated from the economic consequences of other events. For example, studies of corporation income tax shifting vary in their results from the conclusion that the tax is not shifted at all to the conclusion that it is shifted by more than 100 percent, depending mainly on the methods used to isolate the tax impact. Table 1 shows an estimate of the incidence

Table 1: Taxes as Percent of Income, Canada, 1961
(family income class in dollars)

kind of tax	under 2,000	2,000 to 2,999	3,000 to 3,999	4,000 to 4,999	5,000 to 6,999	7,000 to 9,999	over 10,000
Individual income tax	1.2	2.2	3.8	5.2	8.3	10.2	12.0
Corporation income tax	8.5	4.5	3.7	3.0	3.1	3.5	8.0
Sales, excises, and tariffs	25.2	13.6	13.6	12.2	13.2	13.0	8.5
Property tax	16.3	6.8	5.4	4.8	4.3	4.0	3.8
Estate and succession duties	—	—	—	—	—	—	2.9
Other taxes	2.7	1.6	1.4	1.3	1.4	1.5	2.2
Social security and hospital insurance	6.1	4.1	4.4	3.9	2.5	1.9	1.1
Total tax incidence (federal, state, and local)	60.0	32.8	32.3	30.4	32.8	34.1	38.5

Source: W.I. Gillespie, "The Incidence of Taxes and Public Expenditures in the Canadian Economy," *Studies of the Royal Commission on Taxation*, 1966.

of taxes in Canada on various income classes, based on certain assumptions as to the degree and direction of shifting. It suggests that individuals in the lower income classes were unable to shift their tax burden while property owners and rich stockholders were. The greatest impact on the lower income classes came from consumption taxes, as one would expect, but the table also suggests that the corporation income tax and property tax were largely shifted to these classes. The extreme tax regression at the lower end of the income scale may overstate the inequity, since the underlying assumptions are biassed toward the hypothesis of "forward" shifting—*i.e.*, taxes shifting toward consumers and away from property owners.

In the United States the tax burden is also highly regressive, falling much more heavily on the lower income levels than on the middle and upper income classes, but this is offset by so-called transfer payments from the government such as social security, veterans' benefits, pensions, and welfare benefits. Table 2 shows how these compare, according to one estimate. (M.Sc.)

Table 2: Taxes and Transfer Payments as Percent of Income, United States, 1965

income classes	federal taxes	state and local taxes	total taxes*	transfer payments	taxes less transfers*
Under $2,000	19	25	44	126	−83†
$2,000–$4,000	16	11	27	11	16
$4,000–$6,000	17	10	27	5	21
$6,000–$8,000	17	9	26	3	23
$8,000–$10,000	18	9	27	2	25
$10,000–$15,000	19	9	27	2	25
$15,000 and over	32	7	38	1	37
Total	22	9	31	14	24

*Detail will not always add to totals because of rounding. †The minus sign indicates that families and individuals in this class received more as a group from federal, state, and local governments than they paid to these governments in taxes.
Source: Joseph A. Pechman, *The Rich, the Poor, and the Taxes They Pay*, Reprint 168, p. 33, 1969. (*Cf. Economic Report of the President*, pp. 160–161, January 1969).

THE TAX STRUCTURE IN VARIOUS COUNTRIES

The share of the Gross National Product taken by taxes varies considerably at different times and in different countries. This is partly explained by differences in economic development: the most productive taxes, such as general income and turnover or net-value-added taxes, can be applied effectively only in the more developed countries. Ideological differences also play a part; in countries where conservative beliefs about the role of government prevail, the level of taxation is likely to be lower than in countries where public opinion favours a relatively high degree of government intervention. There may also be some correlation between taxation and per capita income: lower levels of taxation in respect to national income tend to be found in countries where the levels of per capita income are low. One might suppose that a country with high per capita incomes would have less need for high government expenditures, especially in the form of social welfare, but the fact is that the demand for government services tends to increase with the level of national income. (Some countries in the middle range do, however, collect relatively more tax revenues than the very rich countries.) Wars are another factor that are an influence on the level of taxation, since they create the need for greatly increased government expenditures not only while they are being fought but sometimes for many years afterward.

International comparisons such as the foregoing make sense only if they take into account the taxes imposed by all levels of government. One frequently finds comparisons limited to the taxes levied by central governments, but these are misleading, since the share of central governments in total national tax revenues varies greatly. It is also important to include compulsory social security contributions as well. Finally, it should be emphasized that the comparability and reliability of statistics are not to be taken for granted; so far as tax revenues are concerned the figures may be taken as accurate, but estimates of Gross National Product involve complex procedures that are not the same in every country.

Table 3 shows total taxes, social security contributions, and Gross National Product for nine industrial countries. Keeping in mind the qualifications made above, it can be said in general that differences among these countries with respect to the per capita burden of taxes are relatively minor except for Italy, where per capita tax revenues are rather low ($312), and Sweden and the U.S., where they are very high ($1,083 and $888, respectively). If taxes and social security contributions are considered together, there seems to be a strong (but not complete) correlation between the per capita burden (column 11) and per capita income (column 9). Scrutinizing the figures in column 8, it is not surprising that Sweden as the prototype of a modern welfare state shows the highest ratio of taxes and social security contributions to Gross National Product (40.4 percent). The corresponding ratio for Switzerland, one of the richest and at the same time most conservative countries, which has not been involved in a war for more than a century, is only 21.8 percent. One also may note that national differences in the ratio of social security contributions to GNP (column 7) are in most cases much more substantial than those in the ratio of taxes to GNP (column 6).

National tax systems also vary in the emphasis they give to different types of taxes. Comparisons are difficult because of differences in terminology and practice, but Table 4 presents an estimate for seven Western countries of the relative weight of different taxes in their national revenues. In France and Italy, indirect taxes bulk larger than direct taxes as a source of governmental revenues. Property taxes figure heavily in state and local revenues in the United States and in the United Kingdom, but they are at a minimum or wholly nonexistent in the other countries listed.

Finally, it is of some interest to compare the distribution

Table 3: Share of Taxes and Social Security Contributions in Gross National Product and Their Per Capita Amounts

country and year	gross national product	taxes	social security contributions	taxes + social security contributions	col. 3 as % of GNP	col. 4 as % of GNP	col. 5 as % of GNP	col. 2 in $U.S. per capita	col. 3 in $U.S. per capita	col. 5 in $U.S. per capita
	billions of national currency									
1	2	3	4	5	6	7	8	9	10	11
U.S. (1967)	789.7	177.20	41.89	219.08	22.4	5.3	27.7	3,955	888	1,097
U.K. (1967)	39.7	12.52	1.96	14.48	31.5	4.9	36.5	1,735	546	633
Fed. Rep. of Germany (1968)	529.8	121.81	55.36	177.17	23.0	10.4	33.4	2,284	525	764
France (1968)	595.4	140.28	78.42	218.70	23.6	13.2	36.7	2,311	568	886
Netherlands (1968)	89.6	22.32	11.50	33.82	24.9	12.8	37.7	1,967	486	736
Italy (1968)	46,740.0	10,270.00	5,006.00	15,276.00	22.0	10.7	32.7	1,418	312	463
Sweden (1968)	131.2	44.33	8.68	53.01	33.8	6.6	40.4	3,455	1,083	1,295
Switzerland (1967)	68.0	11.36	3.48	14.84	16.7	5.1	21.8	2,593	433	566
Japan (1967)	41,563.1	6,441.30	1,507.70	7,979.00	15.5	3.6	19.1	1,198	178	219

Source: *Finanzbericht 1970*, Bundesministerium der Finanzen, pp. 358–359 (1970).

of tax revenues between different levels of government. Table 5 makes the comparison between central (or federal) government revenues and local (or state and local)

Table 4: Composition of National Tax Revenues
(percent of total revenue)

	U.S.	U.K.	W. Ger.	France	Sweden	Italy	Japan
Direct							
Individual income tax	32.3	35.5	19.7	14.7	52.8	11.3	22.3
Corporation income tax	16.3	4.0	6.9	6.0	—	2.1	32.1
Employees' soc. sec. contrib.	6.7	7.3	13.5	6.9	7.4	6.3	—
Other	2.0	3.2	10.0	2.2	1.0	5.5	0.8
Total	57.3	50.0	50.1	29.8	61.2	25.2	55.1
Indirect							
Excises, sales, customs	21.4	34.6	32.7	44.8	33.8	46.6	37.4
Property tax	13.2	9.5	1.6	1.0	—	—	—
Employers' soc. sec. contrib.	7.5	5.9	13.7	19.3	4.9	27.1	—
Other	0.6	—	1.9	5.1	0.2	1.2	7.6
Total	42.7	50.0	49.9	70.2	38.9	74.9	44.9

Source: *Excise Tax Compendium*, Part 1 (1964). *Japan Statistical Yearbook 1968*. Data for 1961.

revenues. The chief differences among these countries obviously relate to the national constitutional structures. The share of revenues going to the central government tends to be lower in federal states (ranging from 56 percent to 65.8 percent), and the share of revenues going to local government tends to be lower in unitary states (ranging from 10.9 percent to 25.6 percent).

Table 5: Division of Tax Revenues by Level of Government
(percent of total revenue)

	central (federal) government	local (local and state) government
U.S.	65.8	34.2
Canada	62.6	37.4
U.K.	89.1	10.9
West Germany	56.0	44.0
France	84.8	15.2
Sweden	74.4	25.6
Italy	85.3	14.7
Japan	71.1	28.9

Sources: *Excise Tax Compendium*, Part 1 (1964), and *Japan Statistical Yearbook 1968*. Data for 1961.

BIBLIOGRAPHY. Problems of taxation are treated in every textbook on public finance and fiscal policy. Outstanding examples are: U.K. HICKS, *Public Finance*, 2nd ed. (1955); R.A. MUSGRAVE, *The Theory of Public Finance* (1959), probably the most influential contemporary work in this field; G. COLM, *Essays in Public Finance and Fiscal Policy* (1955); H. BROCHIER and P. TABATONI, *Économie Financière* (1959); C. COSCIANI, *Principii di Scienza delle Finanze* (1953); J.M. BUCHANAN, *Fiscal Theory and Political Economy* (1960); E.E. ROLPH and G.F. BREAK, *Public Finance* (1961); H.M. GROVES, *Financing Government*, 6th ed. (1964); A.R. PREST, *Public Finance in Theory and Practice*, 3rd ed. (1967); J.F. DUE, *Government Finance: Economics of the Public Sector*, 4th ed. (1968); and C.S. SHOUP, *Public Finance* (1969).

Among older studies in tax problems and principles, the following are still of great importance: A. SCHAEFFLE, *Die Steuern*, 2 vol. (1895–97), the most comprehensive monograph on both theoretical and empirical tax problems; J. STAMP, *The Fundamental Principles of Taxation*, new ed. (1936); and A.C. PIGOU, *A Study in Public Finance*, 3rd ed. (1952).

A modern theory of taxation is developed in H. HALLER, *Die Steuern* (1964); and a more eclectic view in G. SCHMOELDERS, *Allgemeine Steuerlehre*, 4th ed. (1965).

Tax theory, as well as the tax policies of many countries, is treated by European and American authors in W. GERLOFF and F. NEUMARK, *Handbuch der Finanzwissenschaft*, 2nd ed., 4 vol. (1962–65). Contemporary issues of U.S. fiscal policy are examined in the various publications of the Joint Economic Committee of the U.S. Congress; one of the most important is *Federal Tax Policy for Economic Growth and Stability* (*Papers*, 1955 and *Hearings*, 1956). For Canadian problems, see *Report of the Royal Commission on Taxation* (known as the Carter Report), 5 vol. (1966).

Important works on specific tax problems include the following: O. VON MERING, *The Shifting and Incidence of Taxation* (1942); W.S. VICKREY, *Agenda for Progressive Taxation* (1947); N. KALDOR, *An Expenditure Tax* (1955); C.A. HALL, JR., *Fiscal Policy for Stable Growth* (1960); and C. GREEN, *Negative Taxes and the Poverty Problem* (1967).

Useful works on the historical development of fiscal thought include: F.K. MANN, *Steuerpolitische Ideale* (1937), a comprehensive history of the ideals, ideologies, and theories of taxation from both the economic and the political-sociological standpoints; R.A. MUSGRAVE and A.T. PEACOCK (eds.), *Classics in the Theory of Public Finance* (1958); and AMERICAN ECONOMIC ASSOCIATION, *Readings in the Economics of Taxation*, selected by R.A. MUSGRAVE and C.S. SHOUP (1959). A contemporary monograph on principles of taxation is F. NEUMARK, *Grundsätze gerechter und ökonomisch rationaler Steuerpolitik* (1970).

(F.N.)

Taxila

Taxila (Takṣaśilā), an ancient city of northwestern India, was from the 5th century BC a meeting place of Indian, Iranian, and Greek cultural influences. Its site is between the Indus and the Jhelum rivers, about 22 miles northwest of Rawalpindi in the country of Pakistan. Its prosperity in ancient times resulted from its position at the junction of three great trade routes: one from eastern India described by a Greek writer, Megasthenes, as the "Royal Highway," the second from western Asia, and the third from Kashmir and Central Asia. When these routes ceased to be important the city sank into insignificance and was finally destroyed by the Huns in the 5th century AD.

History. Taxila is known from references to it in Indian and Greco-Roman literary sources and from the accounts of two Chinese pilgrims, Fa-hsien and Hsüan-tsang.

Legendary origins

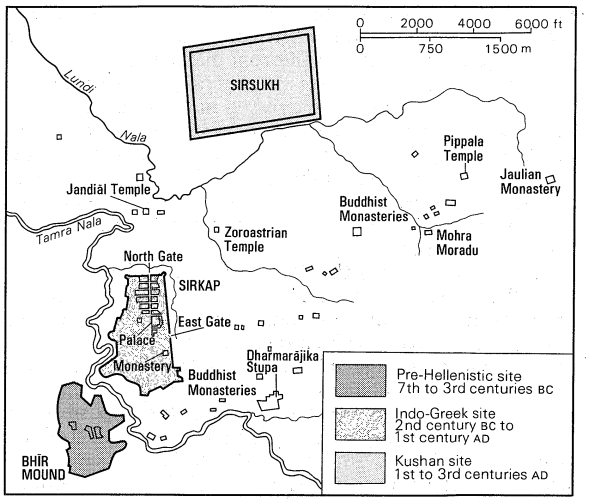
Literally meaning "the city of cut-stone" or "the rock of Takṣa," Takṣaśilā (rendered by Greek writers as Taxila) was founded, according to the Indian epic *Rāmāyaṇa*, by Bharata, younger brother of Rāma, an incarnation of the Hindu god Viṣṇu. The city was named for Bharata's son, Takṣa, its first ruler. The great Indian epic *Mahābhārata* was, according to tradition, first recited at Taxila at the great snake sacrifice of King Janamejaya, one of the heroes of the story. Buddhist literature, especially the Jātaka, mentions it as the capital of the kingdom of Gandhāra and as a great centre of learning. Gandhāra (Gandarii) is also mentioned as a satrapy or province in the inscriptions of the Achaemenian (Persian) king Darius I in the 5th century BC. Taxila, as the capital of Gandhāra, was evidently under Achaemenian rule for more than a century. When Alexander the Great invaded India in 326 BC Āmbhi (Omphis), the ruler of Taxila, surrendered the city and placed his resources at Alexander's disposal. Greek historians accompanying the Macedonian conqueror described Taxila as "wealthy, prosperous, and well governed."

Within a decade after Alexander's death, Taxila was absorbed into the Mauryan empire founded by Candragupta Maurya, under whom it became a provincial capital. But this was only an interlude in the history of Taxila's subjection to conquerors from the west. After three generations of Mauryan rule, the city was annexed by the Indo-Greek kingdom of Bactria. It remained under the Indo-Greeks till the early 1st century BC. They were followed by the Śakas, or Scythians, from Central Asia, and by the Parthians, whose rule lasted till the latter half of the 1st century AD.

According to early Christian legend, Taxila was visited by the Apostle Thomas during the Parthian period. Another distinguished visitor was the Neo-Pythagorean sage Apollonius of Tyana (1st century AD), whose biographer Philostratus described Taxila as a fortified city that was laid out on a symmetrical plan and compared it in size to Nineveh.

Taxila was taken from the Parthians by the Kuṣāṇas under Kujūla Kadphises. The great Kuṣāṇa ruler Kaniṣka founded Sirsukh, the third city on the site. (The second, Sirkap, dates from the Indo-Greek period.) In the 4th century AD the Sāsānian king Shapur II (310–379) seems to have conquered Taxila, as evidenced by the numerous Sāsānian copper coins found there. There is little information about the Sāsānian occupation, but when Fa-hsien visited the city at about the beginning of the 5th century AD, he found it a flourishing centre of Buddhist sanctuaries and monasteries. Shortly thereafter it was sacked by the Huns. Taxila never recovered from this calamity. Hsüan-tsang, visiting the site in the 7th century AD, found the city ruined and desolate, and subsequent records do not mention it. Excavations begun by Sir Alexander Cunningham, the father of Indian archaeology, in 1863–64 and 1872–73, identified the local site known as Saraikhala with ancient Taxila. This work was continued by Sir John Marshall, who over a 20-year period completely exposed the ancient site and its monuments.

Archaeology. The structural remains at Taxila include the Bhīr mound area, the palace area at Sirkap, the Jandiāl and Pippala temples, the Giri fortress, the Mohra Moradu and Jaulian monasteries, and the Dharmarājika, Bhallar, and Kuṇāla *stūpas* (burial mounds). Different types of masonry used in the monuments indicate their period of origin. The earliest remains are those of the Bhīr mound. The palace area, modelled on the same lines as its Assyrian counterpart, has several entrances and outer fortifications. It reveals traces of successive settlements with the oldest parts of the buildings consisting of rubble masonry. Uncovered there were a spacious Buddhist temple, several small shrines, and blocks of dwelling houses. The shrine of the double-headed eagle is interesting for its pilasters of the Corinthian order on the front and niches in the interspaces. Other antiquities of the palace area include terra-cottas and potteries; small bronze, copper, and iron objects; beads, gems, and coins of Indo-Greek, Parthian, and early Kuṣāṇa rulers. The stratigraphic relation between the dated structure inside the palace area and the stone city wall in the southern sector was revealed in the 1944–45 excavations directed by the British archaeologist Sir R.E. Mortimer Wheeler.



Taxila.
Adapted from *Westermann Grosser Atlas zur Weltgeschichte*, Georg Westermann Verlag, Braunschweig

The Dharmarājika *stūpa*, popularly known as the *Chir tope*, is a circular structure with a raised terrace around its base. A circle of small chapels surround the great *stūpa*. Three distinctive types of masonry in the buildings around the main *stūpa* suggest the contributions of different periods to the building activity. A silver scroll inscription in Kharoṣṭī, and a small gold casket containing some bone relics of Buddha were found in one of the chapels. The inscription refers to the enshrinement of the relics by a Bactrian named Ursaka from the town of Noacha in the year 136 BC for the bestowal of health on "the great King, Supreme King of Kings, the Son of Heaven, the Kuṣāṇa" (probably Vīma Kadphises, son of the Kuṣāṇa conqueror Kujāla). This site also contained several statues of the Buddha and *bodhisattvas*.

The Jandiāl temple, set up on an artificial mound, closely resembles the classical temples of Greece. Its Ionic columns and pilasters are composed of massive blocks of sandstone. Built in the Scythio-Parthian period, it is probably the temple described by Philostratus in his *Life of Apollonius of Tyana*. Though the Jandiāl temple is not Buddhist, the Jaulian remains are. These include a monastery and two *stūpa* courts.

Taxila, besides being a provincial seat, was also a centre of learning. It was not a university town with lecture halls and residential quarters, such as have been found at Nālandā in Bihar. At Taxila, the preceptor housed his own pupils, who paid for their board and lodging in cash or in the form of service to the teacher and his family. The Buddhist monasteries also catered to the needs of the students and monks.

BIBLIOGRAPHY. JOHN MARSHALL, *Taxila*, 3 vol. (1951), provides the most exhaustive material for the history and archaeological excavations of Taxila. The report of Sir Mortimer Wheeler's excavations at Sirkap in Taxila is published in *Ancient India*, no. 4 (1947–48). RADHA KUMUD MOOKERJI, *Ancient Indian Education*, 2nd ed. (1951), includes a comprehensive account of Taxila as a centre of learning. For a general study of Taxila as an ancient city see S. PIGGOT, *Some Cities of Ancient India* (1945); and also B.N. PURI, *Cities of Ancient India* (1966).

(B.N.P.)

Tax Law

Tax law is the body of rules under which the public authority has a claim on taxpayers requiring them to transfer to the public authority part of their income or their property. Tax law is concerned only with the legal aspects of taxation, not with its financial, economic, or other aspects. The making of decisions as to the merits of various kinds of taxes, the general level of taxation, and the rates of specific taxes, is not a legal but a political

The nature of tax law

process; and a question such as whether it is advisable to grant privileged tax treatment to capital gains is not a legal problem but a matter of economic policy.

The power to impose taxes is generally recognized as a right of governments, not only of national governments but also, in most countries, of local communities. It also has been recognized recently as a right of an international or supranational community (as in the case of the levy introduced in December 1953 by the European Coal and Steel Community on steel-producing enterprises in the member countries).

Tax law falls within the domain of public law—*i.e.*, the rules that determine and limit the activities and reciprocal interests of the political community and the members comprising it—as distinguished from relationships between individuals (the sphere of private law). The specific purpose of the relationship between the individuals and the community regulated by tax law is the imposition on individuals (and corporations) of a contribution to cover public expenditure. International tax law is concerned with the problems arising when an individual or corporation is taxed in several countries. Tax law can also be divided into material tax law, which is the analysis of the legal provisions giving rise to the charging of a tax; and formal tax law, which concerns the rules laid down in the law as to assessment, enforcement, procedure, coercive measures, administrative and judicial appeal, and other such matters.

The development of tax law as a comprehensive, general system is a recent phenomenon. One reason for this is that no general system of taxation existed in any country before the middle of the 19th century. In traditional, essentially agrarian, societies, government revenues were drawn either from nontax sources (such as tribute, income from the royal domains, and land rent) or, to a lesser extent, from taxes on various objects (land taxes, tolls, customs, and excises). Levies on income or capital were not considered an ordinary means for financing government. They appeared first as emergency measures. The British system of income taxation, for example, one of the oldest in the world, originated in the act of 1799, introduced by the younger Pitt as a temporary means for meeting the increasing financial burden of the Napoleonic Wars. Another reason for the relatively recent development of tax law is that the burden of taxation—and the problem of definite limits to the taxing power of public authority—became substantial only with the broadening in the concept of the proper sphere of government that has accompanied the growing intervention of modern states in economic, social, cultural, and other matters.

THE TAXING POWER

Predominance of the legislature

The limits to the right of the public authority to impose taxes are set by the power that is qualified to do so under constitutional law. In a democratic system this power is the legislature, not the executive or the judiciary. The principle of the legality of taxation has been asserted by constitutional texts in many countries, including the United States, France, Brazil, and Sweden. In Great Britain, which has no written constitution, taxation is also a prerogative of the legislature.

The historical origins of this principle are identical with those of political liberty and representative government —the right of the citizens

> to take cognizance, either personally or through their representatives, of the need for the public contributions, to agree to it freely, to follow its use and to determine its proportion, basis, collection and duration

(in the words of the Declaration of the Rights of Man and the Citizen proclaimed in the first days of the French Revolution, August 1789). Other precedents may be found in the English Bill of Rights of 1689 and the rule "no taxation without consent" laid down in the Declaration of Independence of the United States.

Under this principle all that is necessary is that the rights of the tax administration and the corresponding obligations of the taxpayer be specified in the law; that is, in the text adopted by the people's representatives. The implementation of the tax laws is generally regulated by the executive power (the government or the tax bureau). There have been many encroachments on the principle of the legality of taxation: sometimes the base or the rate of taxation is determined by government decree rather than by law. The encroachment of the executive power on the territory reserved to the legislature in matters of taxation is generally explained by the need to make tax policy more flexible; urgent amendments may be required by sudden changes in the economic situation, changes so sudden that recourse to relatively slow parliamentary procedure would take too long. A compromise may be reached between the orthodox doctrine of the legality of taxes and the need, under special circumstances, to amend texts on taxation almost immediately, by modifying the text through a decree or an order of the executive (treasury) and ratifying it by the legislative power as soon as possible thereafter.

Limitations on the taxing power. While the legislature is omnipotent in its exercise of the taxing prerogative, certain restraints on the legislative power are imposed by tradition, custom, and political considerations. In many countries there are also constitutional limitations. But tax (or other) legislation once enacted cannot be nullified by the courts on the ground that it is in opposition to constitutional principles—except in countries such as the United States, where no tax may be imposed by an act of Congress if it has been found unconstitutional by the judiciary power.

Jurisdictional limitations

Certain limitations on the taxing power of the legislature are self-evident. Since the power to raise taxes is a prerogative of the public authority, a government only has the right to impose a levy insofar as it is competent to do so. There must be a minimum connection between the subject of taxation and the taxing power. The extent of income-tax jurisdiction, for example, is essentially determined by two main criteria: the residence (or nationality) of the taxpayer, and his source of income. As for the first criterion, the test in almost all countries is residence; *i.e.*, the generally permanent home place of the individual, or the place in which control of a corporation is situated. Some countries (including the United States) exercise the right to tax the whole income of their nationals even if they reside abroad. The second basic criterion is the source of income: almost all countries consider it their right to tax income arising within their borders, whether or not the income is earned by individuals or corporations having their residence or exercising their management and control in the country. The application of both criteria together often results in burdensome double taxation, which can be avoided (or restricted) by international treaties.

Taxes other than income taxes—such as retail-sales taxes, turnover taxes (sales taxes levied on production and distribution), death taxes, registration fees, and stamp duties—are imposed by the authority (national or local) on whose territory the goods are delivered or the taxable assets are located.

Another self-evident limitation on the taxing power of the public authority is that the same authority cannot impose the same tax twice on the same person on the same ground.

Taxes are generally not levied retroactively, except in special circumstances. One example of retroactive taxation was the taxation of wartime benefits in some European countries by legislation enacted in 1945 when the war and enemy occupation were over.

The requirement of uniformity

A common limitation on the taxing power is the requirement that all citizens be treated alike. This requirement is specified in the United States Constitution. A similar provision in other constitutions is that all citizens are equal and that no privileges can be granted in tax matters. Unfortunately, the rule is often violated through the influence of pressure groups; it is also difficult to enforce. In countries in which local governments are under the control of the national government, a local tax can be nullified by the central authority on the ground that it violates the rule of uniformity and equality of taxpayers.

Aside from the foregoing constitutional, traditional, or political limitations, there is no restraint on the power of

the legislative body to determine the object or the rate of a tax and the circumstances under which such special adjustments as exemptions or deductions will be granted. The amount of taxation may be excessive: only the legislature can decide that point, by rejecting or amending a tax bill before it is passed. Once enacted by the legislature, a tax cannot be judicially restrained.

Taxation need not be intended for revenue purposes. In modern societies in which government revenues comprise up to 35 percent or more of the gross national product, taxation is shaped by ideology and by economic policy. Taxes may be adjusted to provide an incentive to save (for example, by reducing the registration duties on the purchase of small properties), to encourage business concentration (through tax exemptions for mergers of corporations), business investment (through higher depreciation allowances in corporation income taxation), and to achieve other purposes. Taxes may be punitive (as in the above-mentioned taxation of wartime benefits or intended to discourage unwanted social phenomena such as gambling or alcoholism.

There is no way of mounting a legal attack upon a tax law on the ground that it is arbitrary or unjust. But the application of the law must be correct; it is a matter of public interest that taxpayers be protected against unlawful decisions and actions by the authorities administering the tax laws. The constitutional and legal guarantees protecting the citizen are outlined in the last part of this article.

Double taxation. The problem of double and concurrent taxation by overlapping governmental authorities has become increasingly important, particularly in international law. The growth of international contacts has multiplied the possibility of an individual or corporation being taxed in several countries. Moreover, the expanding financial needs of states have led them to extend their powers of taxation as far as possible, with the result that cases of double taxation are becoming increasingly frequent and serious.

International tax law

International tax law has two parts. One consists of the provisions of internal tax law whereby national taxes are made applicable to nonresidents and to facts or situations located outside the frontiers. The other part has its source in the growing number of international agreements designed to prevent double taxation, either by defining the field of application of the tax laws of each of the contracting states or, without limiting the field of application, by providing for the granting of credits in each of the contracting states for taxes paid under the legislation of the other.

Nearly all the agreements aimed at preventing international double taxation are bilateral; that is, between two countries. Among the few multilateral treaties are the Geneva Convention of 1931, which concerns the tax system applicable to motor vehicles registered abroad, and the conventions concerning the immunities of certain international or supranational organizations. Many bilateral conventions are intended not only to prevent double taxation but also to enable cooperation between the fiscal administrations of the contracting states in combatting tax evasion.

Problems of internal double taxation exist in federal countries (including the United States, Switzerland, and West Germany). A state legislature may, for example, tax all income arising in the state, whether received by residents or nonresidents, or all income received by residents, even when the source of income is located outside the state borders. Therefore, interstate arrangements must be made, similar to international conventions, delimiting the amount of state and federal taxes or granting a credit against the state tax in calculating the federal tax paid on the same object.

Special tax problems arise when countries are involved in economic integration with each other. When two or more countries form a customs union (free-trade zone), each member state keeps its own system of taxation. The aims of an economic union are more ambitious, entailing far-reaching limitations on the sovereignty of the member states; when countries decide to form an economically integrated area, as have the member countries of the European Economic Community (Common Market), they agree to establish a unified economic and financial market. In tax terms, this means the abolition of all tax (and other) discriminations and distortions, since they are likely to give rise to abnormal movements of goods and capital. To this end the sales and turnover taxes of the European Economic Community were "harmonized," as provided in the Rome Treaty of March 1957; all member countries have had to change their sales and turnover taxes to conform to a model prescribed by the Community.

ADMINISTRATION OF TAX LAWS

Whereas the right to impose taxes and to determine the circumstances under which they will be due is a privilege of the legislative power, administration of the tax law is the responsibility of the executive power. The head of tax administration in a central government is the minister of finance, secretary of the treasury, or chancellor of the exchequer. The actual administration is generally separated into departments because taxes differ so greatly in their bases and methods of collection. In most countries the ministry of finance has three branches charged with the levying of taxes. One collects income taxes; another levies taxes on the transfer of goods and on such legal transactions as stamp fees, inheritance taxes, registration dues, and turnover taxes; a third is responsible for customs and excise duties.

The levying of taxes can be divided into three successive phases: (1) assessment, or the definition of the exact amount subject to taxation under the statute; (2) computation or calculation; and (3) enforcement.

Assessment. The definition of the amount subject to taxation under a particular statute requires an analysis of the taxpayer's situation and of the legal provisions that apply to him. With the income tax (and also some taxes on the transfer of property, such as the inheritance tax), the taxpayer submits a tax return providing information as to his occupation, his real and personal property, his professional expenditures, and other pertinent matters; a corporation supplies, additionally, copies of the balance sheet, profit and loss statement, and minutes of the general meeting that approved these financial reports. The return, with the attached reports and statements, is meant to provide such complete information that the assessing tax official can rely on it to compute the correct tax. In the United States, the income taxpayer's liability is computed by himself subject to review by the taxing authority. Most tax systems also collect information in other ways, in order to inform the authorities as to potential tax liabilities. Records are kept of such matters as the allocation of income by partnerships, trusts, or estates, or the payment of fees, interest, dividends, and other sums exceeding a certain minimum amount.

Tax returns

In the case of an annual levy such as the income tax, a return must be filed every year. In many countries, however, individuals who, on the basis of the return previously filed, appear to earn an income below the taxable limit do not have to file a new return annually (this facility is subject to revision at any time). Because it is not easy for some categories of taxpayers to determine the precise amount of their occupational net income, the tax administration frequently reaches an agreement with professional associations, fixing an estimated basis on which the net taxable income of their members will be determined for a period usually exceeding one year; members are then allowed to provide the tax administration with simplified factual information (*e.g.*, for farmers the area of land cultivated, for butchers or bakers the amount of goods sold), instead of filing the standard return.

In many countries a separate assessment procedure has been organized for income from real property; such is the case in the various European countries in which the French system of land register (cadastre) was introduced at the end of the 18th century. The theoretical income of each piece of real property is then determined by the administration of the land register and remains fixed for a relatively long period, except when important changes

are made in the property. In examining tax returns, the basic principle is that a return is assumed to be correct until the assessing official determines otherwise. In countries such as the United States, where the self-assessment method prevails, a minority of returns is selected for audit; most, however, are only checked as to timely arrival, inclusion of all required forms and attachments, and arithmetical accuracy. Except in special circumstances—when, for example, the statute introduces a suspicion of fraud (*e.g.*, if no return has been filed) or creates certain presumptions (as when personal living expenses exceed the reported income)—the administration has no right to shift onto the taxpayer the burden of the proof that he has complied with his liabilities. The golden rule of the tax administrator consists not only in getting as much money as possible for the treasury, but in displaying fairness. The rules of taxation naturally have an authoritarian character, but tax law does not grant the taxing authority a privileged position nor deprive the individual of means of defense against arbitrary taxation.

Tax auditing

In all countries, the assessing officials have extensive powers in determining the amount subject to taxation. In addition to the routine check, there are numerous sources of information. The return of one taxpayer can be checked against that of another: whenever an individual or a corporation includes within deductible expenses the interest paid on borrowed money or the fee paid to a professional expert, the return must show the name and address of the payee, and when this information is placed before the appropriate assessing official he can readily determine whether the payee has included the payment in his declared income.

The procedure varies from one country to another, and depends largely on the circumstances of the case. An audit may be performed either in the office of the tax agent, by correspondence, or in the taxpayer's office. The tax agent is entitled to examine the books and records kept by the taxpayer, within reasonable limits. He is, within the same limits, entitled to question not only the taxpayer but other persons acquainted with the case. There are, however, legal guarantees, protecting confidential communications. In the United States, for example, federal and common law protect communications between husband and wife, or between a client and his attorney acting as such. Under Belgian income-tax law, certain taxpayers, in the course of the assessing official's interrogation, may assert that they are bound by professional secrecy and unable to communicate what they claim to be privileged information; the assessing officer may then consult a special advisory board, composed of the president and two members of the taxpayer's professional or occupational group (lawyers, doctors, notaries, etc.), which will give its opinion as to the taxpayer's probable income.

The problem of bank secrets is a very important one. In some countries, such as Switzerland, banks are legally entitled to refuse information to tax agents concerning their clients. But banks in most countries are required to make reports of cash deposits or similar transactions. In most countries, a safe-deposit box in a bank cannot be opened after the death of the client unless a tax official is present.

Tax authorities do a great deal of intelligence work, using tips from informers such as employees, competitors, and neighbours of the taxpayer. In the United States, informers are encouraged by the payment of fees of up to 10 percent of the additional tax collected. But it is a fundamental principle of tax law that information cannot be used against the taxpayer if it has been obtained by unlawful means, and that no evidence or testimony is a valid proof of tax liability unless the taxpayer has had the opportunity to discuss it.

Assessment consists primarily of an analysis of the facts giving rise to taxation and of the legal provisions applicable to a particular case. The assessor may find himself in disagreement with the taxpayer, either as to the facts (the amount of income, of deductible expenses, etc.) or as to the manner in which the taxpayer has resolved a question of law or a mixed question of law and fact. The tax agent may use his discretion as to questions of fact, and frequently a compromise is reached on those questions between the taxpayer and the tax agent.

Whenever the tax agent decides questions of law, he is bound by the treasury's position on the particular problem. On unresolved issues, lower taxation officers (field offices) usually request the advice of a higher echelon. In some countries, including the United States, written rulings are issued by the administration in advance, thus avoiding disputes at the level of the assessment official. In countries in which this is not done, officials at all levels are free to give informal advice concerning the tax effects of proposed transactions. But the taxpayer can file a petition with the competent administrative or judicial authority whenever he believes that the interpretation of the law by the assessing official is wrong.

Computation of the tax. The second phase in levying taxes is the calculation of the amount to be paid. In the American self-assessment method, the liability for income tax is primarily established by the taxpayer himself. Under this method, as a rule, the tax liability reported on the return forms the basis of the assessment record. If the tax administration discovers that additional tax is due, a deficiency statement is issued. Many countries provide for prepayment of the withholding tax on dividends and other income from personal property, and have set up a "pay-as-you-go" system for professional income. Such provisional payments are calculated by the taxpayer. Advance payment of all or part of the income tax (on a voluntary or compulsory basis) before the return is filed, on the basis of expected income or of the taxable income of the previous year, is also provided for in some countries. In general, however, the final computation of taxes levied on income, on inherited property, or on the transfer of property is made by the tax administration. Turnover taxes, sales taxes, and value-added taxes (*e.g.*, taxes on a manufacturer's sales minus taxes included in his production costs) are calculated by the taxpayers.

Enforcement. If the taxpayer fails to pay within the legally prescribed period, or within a very short time afterward, the competent tax office undertakes to collect the amount due. In proceedings against the taxpayer, the tax administration is not in the position of an ordinary creditor suing an ordinary debtor. In political communities where powers are divided among the legislative, the executive, and the judiciary arms of government, the application of the tax law is the privilege of the executive power—the tax administration. The collecting agent will not embark upon civil litigation to obtain payment of the delinquent taxes or to secure adequate guarantees from the taxpayer.

Civil and criminal sanctions

In addition to interest charges on the amount due, various kinds of coercive measures are available to ensure payment. Civil penalties consist generally of a fine added by the collecting agent when the violation is the result of negligence rather than of willful neglect or bad faith. Examples of negligence are the failure to file a required return on time, and understatement or underpayment of the tax liability without intent to mislead. Civil penalties are fixed by assessment, so that the procedural remedies of the taxpayer are identical with those provided for the assessment of the tax itself. The law also confers a privileged position on the tax administration among the creditors of the taxpayer.

Criminal tax fraud is severely punished in all tax legislation. Certain tax crimes are classed as misdemeanours (such as willful failure to pay certain taxes, to file certain returns, to keep proper records, and to supply proper information); these are punishable by fines or imprisonment or both. Heavier punishment is provided for crimes classed as felonies (such as the making of false statements and, in the United States, tax evasion). In most countries the criminal penalties can be combined with the civil penalties.

Criminal penalties cannot be imposed by the tax administration. Offenses against tax law, whether misdemeanours or felonies, must be tried by courts. The procedure in criminal tax cases is almost identical with that in other criminal cases. The accused is deemed to be innocent

until proved guilty; the burden of the proof inevitably rests upon the prosecutor and not upon the taxpayer-defendant.

THE JUDICIARY AND TAX LAW

The taxpayer has a guarantee against unfairness or error in the application of taxes in the right to appeal to impartial authorities when he disagrees with the determination of the assessing officer.

In some countries disputes between taxpayers and the authorities are settled by special commissions consisting of high-ranking civil servants (and also of members of various occupational organizations). In others, the decision is the privilege of the judiciary power. But in the vast majority of countries a combination of both systems prevails. "Out-of-court" jurisdictions—commissions composed of tax officials and laymen—frequently act as preliminary settlement committees that decide factual questions, leaving the interpretation of the tax law to the courts. In general, when a taxpayer disagrees with the amount of the tax as calculated by the administration, or thinks he has paid too much, he files a petition with a tribunal, which is either a specialized court or the ordinary court competent for civil litigation. Even if he must exhaust the administrative processes before he may take a dispute with the tax authorities to court, he can still invoke the jurisdiction of a judicial court to re-examine the case, in respect both of the facts and of the legal arguments.

In almost all countries, the judiciary is headed by a supreme court whose jurisdiction is limited to questions of law. The supreme court is generally competent in tax matters, but an appeal to the court must be based solely on an alleged misapplication of law; should it appear that questions of fact or mixed questions of law and fact are involved, the claim, under most judicial systems, is dismissed. (The United States Supreme Court can also decide on the constitutionality of an act of the legislative power.) The judiciary thus has final authority to interpret the tax statute. This interpretation is binding only in the matter submitted to the tribunal. But, in any legal system, reference to interpretative decisions of the courts in comparable cases, especially of the supreme court, is obviously the best argument in a litigation about a disputed point of tax law.

Interpretation of tax laws

When a legal text—in tax law or in other law—requires interpretation because there is a reasonable doubt as to its meaning or scope, the first step is to determine the meaning of the words used, according to the rules of grammar and syntax, not in isolation but in their context and taking into account the subject discussed. As stated by R.E. Paul, "the meaning of a sentence may be more than that of the separate words, as a melody is more than the notes" (*Taxation in the United States,* 1954). This is the necessary "literal" interpretation of the law: when the current sense of a term is its wide sense, then it must be accepted in its wide sense in a tax law as in any other, unless it can be shown that the legislature used the term in a narrower sense. If the meaning of the text cannot be determined with certainty by the literal method, then the interpreter, in seeking the legislature's intention, will resort to the "historical" method (the study both of the preparatory work and of the place occupied by the text in successive laws on a specific matter) and to the "systematic" or "teleological" interpretation (the position occupied by a legal provision in a legal system as a whole, and the object pursued by the legislature in producing that system).

In addition to these general principles common to the interpretation of all legal texts, some special rules apply to the interpretation of tax laws. One rule is the autonomy of tax law, meaning that tax laws pursue aims that are different from those of other bodies of law. The tax claim is a claim under public law. Its cause lies not in a contractual obligation but in an expression of unilateral will, a decision by the public authority. The function of taxes in the organization of the budget is incompatible with the principles of the law of contracts; such principles apply only to relationships under private law and, therefore, cannot be invoked to interpret provisions of tax law.

Other special rules of interpretation of tax laws derive from the nature of the tax obligation. In a democratic system, a tax can only be imposed by law. Thus the courts or the administration do not have a "creative power" to make things or operations taxable through an analogic interpretation of the text, in cases where it is not proved that the legislature wished them to be taxable. On the other hand, the rule of legality of taxation does not always operate in favour of the taxpayer: the person or body entrusted with the task of applying or interpreting the law cannot introduce any attenuation or relaxation of its effect, even though this might be more than amply justified by circumstances, except in cases where the legislature has authorized the judge or the administration to apply the rules of equity within certain legally prescribed limits.

BIBLIOGRAPHY. Comparative studies of the tax systems of various countries include: CONFEDERATION OF BRITISH INDUSTRY, *Taxation in Western Europe* (annual); W.J. GIBBONS, *Tax Factors in Basing International Business Abroad: A Study of the Law of the United States and of Selected Foreign Countries* (1957); L. HART WRIGHT *et al., Comparative Conflict Resolution Procedures in Taxation* (1968); and J. VAN HOUTTE, *Principles of Interpretation in Internal and International Tax Law* (1968). Studies of the tax systems of particular countries may be found in the "World Tax Series," published under the auspices of Harvard University.

(J.M.J.v.H.)